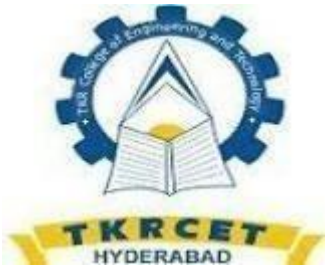
MACHINE LEARNING IN FINANCE



BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE & ENGINEERING

By

R.Gayathri	**20K91A05F2**
Neha	**20K91A05C6**
MD.Abdul Gafoor	**20K91A05B2**
M.Ajay Kumar	**20K91A05B7**

Under the guidance of
Ms.Y.Latha
Assistant professor

DEPARTMENTMENT OF COMPUTER SCIENCE & ENGINEERING
TKR COLLEGE OF ENGINEERING & TECHOLOGY (AUTONOMOUS)
(Accredited by NBA and NAAC with 'A' Grade)
Medbowli, Balapur(M), Hyderabad-500097

CERTIFICATE

This is to certify that the narrowed topic entitled "Machine Learning in Finance", being submitted by **R.Gayathri**, bearing Roll.No:**20K91A05F2**, **Neha**, bearing Roll.No:**2OK91A0C6**, **MD.Abdul Gafoor**, bearing Roll.No:**20K91A05B2, M.Ajay Kumar** bearing Roll.No:**20K91A05B7**,completed literature survey with best of their knowledge and submitting this report to the **Department of Computer Science & Engineering**, **TKR College of Engineering & Technology** under my guidance and supervision.

Signature of the Guide
Ms.Y.LATHA
Assistant Professor

CONTENTS

1. Machine Learning Models for Predicting Bank Loan Eligibility
2. Customer Loan Eligibility Prediction using Machine Learning
3.A Study on Machine Learning Algorithm for Enhancement of Loan Prediction
4. Loan Prediction using Machine Learning Algorithms
5. Analysis of Loan Availability using Machine
Learning Techniques
6.Machine Learning Techniques for Recognizing the Loan Eligibility
7. Prediction for Loan Approval using Machine Learning Algorithm
8. Prediction for Loan Approval using Machine Learning Approach
9. Customer Loan Prediction Using Supervised Learning
Technique
10.A comparative Study on Loan Eligibility
11. Loan Prediction using Decision Tree and Random Forest
12. Machine Learning Techniques for Recognizing the Loan Eligibility
13. Predict Loan Approval in Banking
System Machine Learning Approach
for Cooperative Banks Loan Approval
14.Logistic Regression Based Loan Approval Prediction
15. ML Based Loan Approval Prediction System
A Novel Approach
16.Prediction of Customer Loan Elibility Using Random Forest Algorithm
17. Loan Default Identification and its Effect
18. Comparative Analysis of Bank Loan Defaulter
Prediction Using Machine Learning Techniques
19.Machine Learning Algorithm to Predict Fraudulent Loan Requests
20.Loan Analysis Predicting Defaulters
21.Accuracy Prediction for Loan Risk Using Machine Learning Models
22.An Improved light gradient boosting machine algorithm based on swaen
algorithms for predicting loan default of peer-to-peer lending
23.Loan Prediction using Machine Learning Algorithms
24.Predicting Bank Loan Risks Using Machine Learning Algorithms
25.Fraud Detection in Bank Loan administration using decision tree

2022 IEEE NIGERCON

2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON) | 978-1-6654-7978-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/NIGERCON54645.2022.9803172

Machine Learning Models for Predicting Bank Loan Eligibility

Ugochukwu .E. Orji
Dept. of Computer Science
University of Nigeria, Nsukka
Enugu, Nigeria
ugochukwu.orji.pg00609@unn.edu.ng

Chikodili .H. Ugwuishiwu
Dept. of Computer Science
University of Nigeria, Nsukka
Enugu, Nigeria
chikodili.ugwuishiwu@unn.edu.ng

Joseph. C. N. Nguemaleu
Dept. of Computer Science
University of Nigeria, Nsukka
Enugu, Nigeria
nguemaleu.ngako.dp000058@unn.edu.ng

Peace. N. Ugwuanyi
Dept. of Computer Science
University of Nigeria, Nsukka
Enugu, Nigeria.
Peace.ugwuanyi@unn.edu.ng

***Abstract* — Machine learning algorithms are revolutionizing processes in all fields including; real-estate, security, bioinformatics, and the financial industry. The loan approval process is one of the most tedious task in the banking industry. Modern technology such as machine learning models can improve the speed, efficacy, and accuracy of loan approval processes. This paper presents six (6) machine learning algorithms (Random Forest, Gradient Boost, Decision Tree, Support Vector Machine, K-Nearest Neighbor, and Logistic Regression) for predicting loan eligibility. The models were trained on the historical dataset 'Loan Eligible Dataset,' available on Kaggle and licensed under Database Contents License (DbCL) v1.0. The dataset was processed and analyzed using Python programming libraries on Kaggle's Jupyter Notebook cloud environment. Our research result showed high-performance accuracy, with the Random forest algorithm having the highest score of 95.55% and Logistic regression with the lowest score of 80%. Our Models outperformed two of the three loan prediction models found in the literature in terms of precision-recall and accuracy.**

Keywords— KNN, SVM, Bagging and Boosting techniques, Efficient ML Algorithms, Loan approval prediction.

I. INTRODUCTION

Like many other business ventures, the banking sector is increasingly looking to take advantage of the opportunities presented by modern technologies to improve their processes, productivity and reduce costs. According to [1], the predictive analytics feature of Machine learning was the most utilized feature for applications in the banking sector worldwide in 2020. The success or failure of most lending platforms largely depends on their ability to evaluate credit risk [2]. The loan approval process is a challenging task for any financial institution. Before giving credit loans to borrowers, the bank decides whether the borrower is bad (defaulter) or good (non-defaulter). This paper focused on developing Machine Learning (ML) models to predict loan eligibility, which is vital in accelerating the decision-making process and determining if an applicant gets a loan or not. Our objectives in this study include; (1) Clean and Preprocess the data for modeling, (2) Perform Exploratory Data Analysis (EDA) on the dataset, (3) Build various ML models to predict loan eligibility, and (4) Evaluate and Compare the different Models built.

II. REVIEW OF RELATED LITERATURE

The authors in [3] carried out a systematic literature review to identify and compare the best fit ML-based models for credit risk assessment. The authors aimed to show the various ML algorithms utilized by researchers for credit assessment of rural borrowers, especially those with inadequate loan history. Their finding showed that the ML algorithms we utilized in this research were widely used and showed great results.

The adverse impact of low loan repayment rates on banks is a major issue globally, and banks are looking for more effective ways to handle loan approval processes. The authors in [4] evaluated the loan default prediction of the Chinese peer-to-peer (P2P) market using R.F, XGBoost, GBM, and Neural Network machine learning models. Their four models exceeded 90% accuracy, with RF being the superior model. This research is closely related to our study in terms of methods used and algorithms deployed; however, they aimed to predict P2P loan default while we aimed to predict customers' eligibility for loans.

In their research, [5] deployed various ensemble ML techniques such as AdaBoost, LogitBoost, Bagging, and Random Forest model to predict loan approval of bank direct marketing data. Their research result showed that AdaBoost had the highest accuracy of 83.97%. When compared to our study, the SMOTE technique we utilized to balance our dataset proved to be the key difference as our models achieved better performance.

The research by [6] studied actual bank credit data to predict customers' creditworthiness and help the banks formulate an automated risk assessment system. They deployed different ML algorithms, including; neural network, naive Bayes, KNN, decision tree, and ensemble learning algorithms. Their model accuracy ranged from 80% to 76% respectively, which is also below the accuracy of our models.

III. METHODOLOGY

This research was done using Python on Kaggle's Jupyter Notebook cloud environment. The proposed model predicts customers' loan eligibility based on the available data. The input to the model includes attributes from the dataset, as shown in table 1. The output from the model is a decision on whether the customer is eligible to get the loan. The following section discusses the dataset and explains the methods used to cleanse and preprocess the dataset for modeling.

978-1-6654-7978-3/22/$31.00 ©2022 IEEE

A. Dataset

The dataset used in this study is the historical dataset 'Loan Eligible Dataset,' available on Kaggle [7] and licensed under Database Contents License (DbCL) v1.0. Table 1 below gives a brief description of the dataset attributes.

Table 1: Dataset description

Variable Name	Description	Data Type
Loan_ID	Loan reference number (Unique I.D.)	Numeric
Gender	Applicant gender	Categorical
Married	Applicant marital status	Categorical
Dependents	Number of family members	Numeric
Education	Applicant educational qualification (graduate or not graduate)	Categorical
Self_Employed	Applicant employment status (yes for self-employed, no for employed/others)	Categorical
Applicant_Income	Applicant's monthly salary/income	Numeric
Coapplicant_Income	Additional applicant's monthly salary/income	Numeric
Loan_Amount	Loan amount	Numeric
Loan_Amount_Term	The loan's repayment period (in days)	Numeric
Credit_History	Records of applicant's credit history (0: bad credit history, 1: good credit history)	Numeric
Property_Area	The location of the applicant's home (Rural/Semi-urban/Urban)	Categorical
Loan_Status	Status of loan (Y: accepted, N: not accepted)	Categorical

B. Data preprocessing and Analysis

To ensure optimal performance of the model, the following techniques were deployed to analyze and preprocess the data for modeling;

1. Synthetic Minority Oversampling Technique (SMOTE): This technique is highly effective for handling imbalanced classification problems, a significant source of error in ML models. The imbalance occurs when there is a limited amount of the minority class in the dataset, which makes it difficult for a model to effectively learn the decision boundary [8]. In this research, we used the SMOTE technique to overcome this challenge by oversampling the examples in the minority class. We achieved this by producing duplicates of the minority class in the training dataset before fitting the model.
2. One-hot encoding technique helps convert categorical variables in a dataset into binary form so that the ML model will understand the data.
3. Normalization: The goal of normalizing data for ML models is to transform features and ensure that they are all on a similar scale. Normalization helps improve the training stability and performance of the model.
4. Exploratory Data Analysis (EDA) is a method of exploring the dataset to discover patterns, trends, and spot anomalies. Also, the dataset was cleaned at this stage to remove/handle missing or incomplete data by performing data imputation (substituting missing values with close estimations).
   On exploring the dataset, we found that;
   - The dataset contains more male applicants than female applicants.
   - The dataset contains more married applicants.
   - The dataset contains more applicants with good credit (1) than those with bad credit (0).

Furthermore, fig 1 below shows the correlation of key variables in the dataset and that Applicant_Income is the most positively correlated attribute to Loan_Amount.

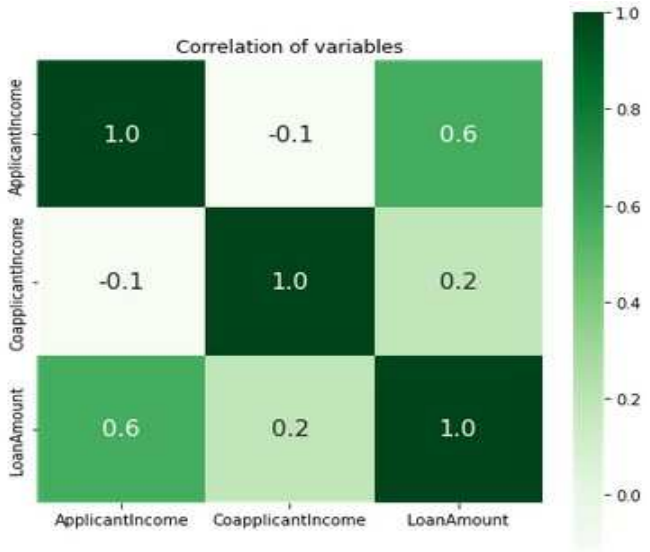


Fig 1: correlation of key variables in the dataset

IV. MODEL DEVELOPMENT AND RESULT

Evaluation metrics explain the performance of an ML model; the performance metrics used for this research include the Confusion Matrix and F1 Score.

The confusion matrix summarizes the number of correct and incorrect predictions by an ML model and breaks it down into classes:

i. "True positive" = Actual positive cases that the model correctly predicts.
ii. "False positive" = Actual positive cases that the model incorrectly predicts.
iii. "True negative" = Actual negative cases that the model correctly predicts.
iv. "False negative" = Actual negative cases that the model incorrectly predicts.

v. "Accuracy" = Overall correct predictions.

F1 score represents the mean of precision and recall values. The formula is given as follows:

$$2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad [9]$$

Where;

Precision = $\frac{\text{True positive}}{\text{True positive} + \text{False positive}}$ and Recall = $\frac{\text{True positive}}{\text{True positive} + \text{False negative}}$ [9]

A. Logistic Regression (LR) Algorithm

LR is a simple classification algorithm used to model a binary (0,1) variable. LR predicts the outcome of a response/dependent variable based on one or more other variables, called predictor/independent variable [10].

The logistic function is given as follows:

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \quad [11]$$

Where;

$1 + e$ denotes the exponential function.

β_0 is the intercept

$\beta_1 x$ is the regression coefficient

Fig 2 below shows the evaluation result of our LR model.

```
In [243]:
LRclassifier = LogisticRegression(solver='saga', max_iter=500, random_state=1)
LRclassifier.fit(X_train, y_train)

y_pred = LRclassifier.predict(X_test)

print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))

from sklearn.metrics import accuracy_score
LRAcc = accuracy_score(y_pred,y_test)
print('LR accuracy: {:.2f}%'.format(LRAcc*100))
```

```
              precision    recall  f1-score   support

           0       0.79      0.75      0.77        20
           1       0.81      0.84      0.82        25

    accuracy                           0.80        45
   macro avg       0.80      0.79      0.80        45
weighted avg       0.80      0.80      0.80        45

[[15  5]
 [ 4 21]]
LR accuracy: 80.00%
```

Fig 2: LR model evaluation

B. K-Nearest Neighbor (KNN) Algorithm

KNN is a supervised ML algorithm that uses the Euclidean distance to calculate the distance between attributes and then matches the data points using the 'feature similarity' in the dataset. The formula is given as follows:

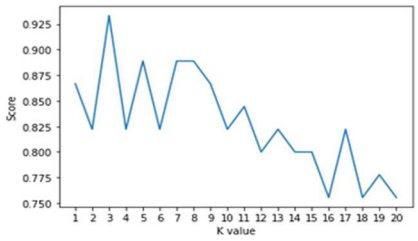
$$\text{dist}((x,y),(a,b)) = \sqrt{(x-a)^2 + (y-b)^2} \quad [12]$$

Where: (x, y) and (a, b) are the coordinates of two points in the plane.

Fig 3 below shows the evaluation result of our KNN model.

```
In [244]:
scoreListknn = []
for i in range(1,21):
    KNclassifier = KNeighborsClassifier(n_neighbors = i)
    KNclassifier.fit(X_train, y_train)
    scoreListknn.append(KNclassifier.score(X_test, y_test))

plt.plot(range(1,21), scoreListknn)
plt.xticks(np.arange(1,21,1))
plt.xlabel("K value")
plt.ylabel("Score")
plt.show()
KNAcc = max(scoreListknn)
print("KNN best accuracy: {:.2f}%".format(KNAcc*100))
```




```
KNN best accuracy: 93.33%
```

Fig 3: KNN model evaluation

C. Support Vector Machine (SVM)

The SVM algorithm explicitly seeks to find a hyperplane in an N-number of features that uniquely classify the data points (vectors) in a dataset [13]. Furthermore, the SVM algorithm is known for higher speed and better performance with a limited number of data samples.

It is given as follows:

$$w^* = \arg_w max \frac{1}{\|w\|_2} [min_n y_n \mid w^T(\phi(x) + b \mid] \quad [14]$$

Where: $[min_n y_n \mid w^T(\phi(x) + b \mid]$ represents the minimum distance of a point to the decision boundary, and $\arg_w max$ represents the maximum points of a function domain.

Fig 4 below shows the evaluation result of our SVM model.

```
In [249]:
SVCclassifier = SVC(kernel='rbf', max_iter=500)
SVCclassifier.fit(X_train, y_train)

y_pred = SVCclassifier.predict(X_test)

print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))

from sklearn.metrics import accuracy_score
SVCAcc = accuracy_score(y_pred,y_test)
print('SVC accuracy: {:.2f}%'.format(SVCAcc*100))
```

```
              precision    recall  f1-score   support

           0       0.88      0.75      0.81        20
           1       0.82      0.92      0.87        25

    accuracy                           0.84        45
   macro avg       0.85      0.83      0.84        45
weighted avg       0.85      0.84      0.84        45

[[15  5]
 [ 2 23]]
SVC accuracy: 84.44%
```

Fig 4: SVM model evaluation

D. Decision Tree (DT) Algorithm

The decision tree algorithm uses the features/attributes present in a dataset to make informed decisions. The objective of the DT algorithm is to maximize the value of information gain [15]. It achieves this by splitting the features (nodes) starting with the highest information gain. It can be calculated using the below formula:

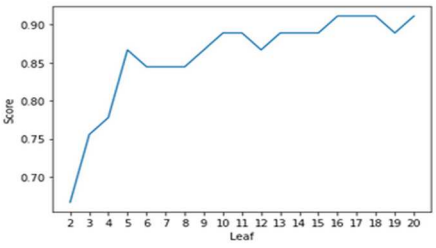
$$IG(T, a) = \mathrm{H}(T) - \mathrm{H}(T \mid a) \qquad [16]$$

Where: $\mathrm{H}(T \mid a)$ is the conditional entropy of T, and a is the value of the attribute.

Fig 5 below shows the evaluation result of our DT model.

```
In [248]:
scoreListDT = []
for i in range(2,21):
    DTclassifier = DecisionTreeClassifier(max_leaf_nodes=i)
    DTclassifier.fit(X_train, y_train)
    scoreListDT.append(DTclassifier.score(X_test, y_test))

plt.plot(range(2,21), scoreListDT)
plt.xticks(np.arange(2,21,1))
plt.xlabel("Leaf")
plt.ylabel("Score")
plt.show()
DTAcc = max(scoreListDT)
print("Decision Tree Accuracy: {:.2f}%".format(DTAcc*100))
```




```
Decision Tree Accuracy: 91.11%
```

Fig 5: DT model evaluation

E. Bagging and Boosting Algorithm

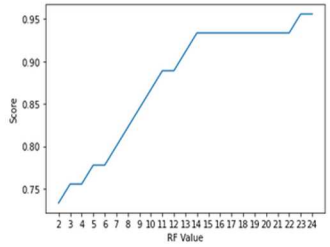
Bootstrap Aggregation (Bagging) is a very efficient ensemble technique that reduces ML models' variance. The Bagging technique achieves this by merging results from multiple classifiers modeled on the various sub-samples of a given dataset [17]. An example of the Bagging technique deployed for this research is the random forest (RF) Algorithm.

On the other hand, the boosting technique cuts across a special class of algorithms tasked with merging weak learners into strong learners. This is achieved by weighing the weak classifiers according to their accuracy and then iteratively learning and merging them into a final robust classifier [18]. The boosting techniques reweight the training set after every iteration and assign weights to any misclassified instances identified in the sequence [19]. In this research, the boosting technique deployed is the Gradient Boost (GBM) Algorithm.

Fig 6 below shows the evaluation result of our RF model.

```
In [250]:
scoreListRF = []
for i in range(2,25):
    RFclassifier = RandomForestClassifier(n_estimators = 1000, random_state = 1, max_leaf_no
des=i)
    RFclassifier.fit(X_train, y_train)
    scoreListRF.append(RFclassifier.score(X_test, y_test))

plt.plot(range(2,25), scoreListRF)
plt.xticks(np.arange(2,25,1))
plt.xlabel("RF Value")
plt.ylabel("Score")
plt.show()
RFAcc = max(scoreListRF)
print("Random Forest Accuracy:  {:.2f}%".format(RFAcc*100))
```




```
Random Forest Accuracy:  95.56%
```

Fig 6: RF model evaluation

Fig 7 below shows the evaluation result of our GBM model.

```
In [251]:
paramsGB={'n_estimators':[100,200,300,400,500],
      'max_depth':[1,2,3,4,5],
      'subsample':[0.5,1],
      'max_leaf_nodes':[2,5,10,20,30,40,50]}
GB = RandomizedSearchCV(GradientBoostingClassifier(), paramsGB, cv=20)
GB.fit(X_train, y_train)
GBclassifier = GradientBoostingClassifier(subsample=0.5, n_estimators=400, max_depth=4, max_
leaf_nodes=10)
GBclassifier.fit(X_train, y_train)

y_pred = GBclassifier.predict(X_test)

print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))

from sklearn.metrics import accuracy_score
GBAcc = accuracy_score(y_pred,y_test)
print('Gradient Boosting accuracy: {:.2f}%'.format(GBAcc*100))
```

```
              precision    recall  f1-score   support

           0       0.90      0.95      0.93        20
           1       0.96      0.92      0.94        25

    accuracy                           0.93        45
   macro avg       0.93      0.94      0.93        45
weighted avg       0.93      0.93      0.93        45

[[19  1]
 [ 2 23]]
Gradient Boosting accuracy: 93.33%
```

Fig 6: GBM model evaluation

V. Discussion

Loan defaulting is a significant financial risk for the banking industry as it damages the interests of lenders and breaks the social trust. The academic field has put extensive effort into developing efficient machine learning techniques to help regulators carry out an accurate loan approval process in real-time. This research utilized state-of-the-art machine learning methods to build credible and accurate prediction models. Fig 7 below shows the comparison of all the machine learning algorithms deployed in this research. Our models achieved high-performance accuracy based on the precision and recall metrics, with the R.F. model achieving a 95% score.

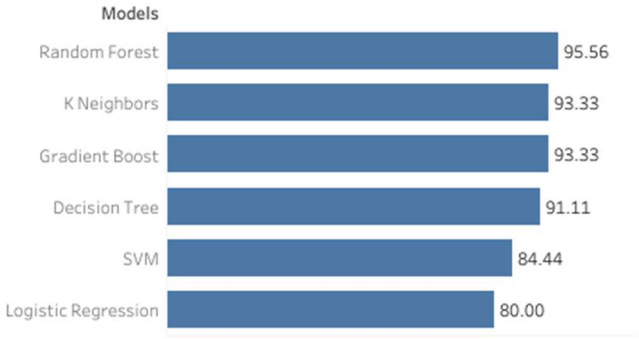


Fig 7: Model result comparison

VI. Conclusion

As more decision-makers in the financial industry seek to understand ways to improve their processes and maintain a balance between the security and reliability of their financial lending system, machine learning techniques can play a vital role in helping achieve this goal. Our ML models achieved high-performance accuracy in predicting loan eligibility in this research. We used the ensemble ML methods (bagging and boosting) and other techniques like SMOTE to ensure optimal predictive models. In general, the methods and algorithms deployed in this research could be instrumental to the successes of financial regulators, corporate, and individual borrowers in their effort to improve their overall loan approval process.

Additional Information

The datasets analyzed and complete documentation of the data analysis and model development process are available at: https://www.kaggle.com/orjiugochukwu/using-ml-algorithms-for-loan-approval-prediction.

References

[1] "Most commonly used A.I. application in investment banking worldwide 2020, by types." Statista, 15-Sept-2021 [Online]. Available: https://www.statista.com/statistics/1246874/ai-used-in-investment-banking-worldwide-2020/ [Accessed: 29-Jan-2022]

[2] G. Dorfleitner, E.M. Oswald, & R. Zhang. "From Credit Risk to Social Impact: On the Funding Determinants in Interest-Free Peer-to-Peer Lending." J Bus Ethics. 2021 Vol.170, pp. 375–400. https://doi.org/10.1007/s10551-019-04311-8

[3] A. Kumar, S. Sharma, & M. Mahdavi, "Machine Learning (ML) Technologies for Digital Credit Scoring in Rural Finance: A Literature Review." Risks 9.11 (2021): 192

[4] J. Xu, Z. Lu, and Y. Xie, "Loan default prediction of Chinese P2P market: a machine learning methodology." Scientific Reports, 2021, Vol. 11(1), pp. 1-19.

[5] H. Meshref, "Predicting Loan Approval of Bank Direct Marketing Data Using Ensemble Machine Learning Algorithms."International Journal of circuits, systems, and signal processing. 2020, Vol. 14, pp. 914-922 DOI: 10.46300/9106.2020.14.117

[6] A.S. Aphale, and S.R. Shinde, "Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval." International Journal of Engineering Research & Technology (IJERT). 2020, Vol. 9 pp. 991-995

[7] "Loan Eligibility Dataset." Kaggle, 15-Aug-2020. [Online] Available:https://www.kaggle.com/datasets/vikasukani/loan-eligible-dataset

[8] A.S. Hussein, T. Li, C.W. Yohannese, & K. Bashir. "A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE." International Journal of Computational Intelligence Systems. 2019, Vol. 12(2), PP.1412.of the 2003 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Stroudsburg, PA. 2003, pp. 129-136.

[9] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." arXiv preprint arXiv:2010.16061 (2020).

[10] J.M. Hilbe. "Logistic Regression." International encyclopedia of statistical science. 2011, Vol 1: pp. 15-32.

[11] A. Saini. "Logistic Regression | What is Logistic Regression and Why do we need it?" 26-Aug-2021[Online] Available: https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/#h2_5 [Accessed: 28-Jan-2022]

[12] M. Shouman, T. Turner, and R. Stocker. "Applying k-nearest neighbour in diagnosing heart disease patients." International Journal of Information and Education Technology. 2012 Vol. 2(3), pp. 220-223.

[13] L.K. Ramasamy, S. Kadry, Y. Nam, & M.N. Meqdad. "Performance analysis of sentiments in Twitter dataset using SVM models. International Journal of Electrical and Computer Engineering (IJECE). 2021 Vol. 11, No. 3, pp.2275-2284 https://doi.org/10.11591/ijece.v11i3.

[14] R. Kunchhal. "Mathematics Behind SVM | Math Behind Support Vector Machine." 28-Dec-2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/the-mathematics-behind-svm/ [Accessed: 27-Jan-2022]

[15] K. Yadav, and R. Thareja. "Comparing the performance of naive bayes and decision tree classification using R." International Journal of Intelligent Systems and Applications. 2019, Vol.11(12), p.11.

[16] K. Ramya, Y. Teekaraman, & K.R. Kumar. "Fuzzy-based energy management system with decision tree algorithm for power security system." International Journal of Computational Intelligence Systems. 2019, Vol.12(2), pp.1173.

[17] L.G. Kabari, & U.C. Onwuka. "Comparison of bagging and voting ensemble machine learning algorithm as a classifier." International Journals of Advanced Research in Computer Science and Software Engineering. 2019, Vol. 9(3), pp.19-23.

[18] Z. Tian, J. Xiao, H. Feng, & Y. Wei. "Credit risk assessment based on gradient boosting decision tree." Procedia Computer Science. 2020, Vol.174, pp.150-160.

[19] "Bagging vs Boosting in Machine Learning." GeeksforGeeks. 07-Jul-2021[Online]. Available: https://www.geeksforgeeks.org/bagging-vs-boosting-in-machine-learning/ [Accessed 28-Jan-2022]

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2053-2062
https://publishoa.com
ISSN: 1309-3452

Customer Loan Eligibility Prediction using Machine Learning

1. **Amjan Shaik**, Professor and HoD-CSE,St.Peter's Engineering College, Telangana, India, amjansrs@gmail.com
2. **Kunduru Sai Asritha**, St. Peter's Engineering College, Telangana, India, saiasritha1813@gmail.com
3. **NeelamLahre**,St. Peter's Engineering College, Telangana, India, neelamlahre05@gmail.com
4. **BolluJoshua**, St. Peter's Engineering College, Telangana, India, joshjohn456@gmail.com
5. **Velagapudi Sri Harsha**, St.Peter's Engineering College, Telangana, India, harshavelagapudi1002@gmail.com

Corresponding Author: Amjan Shaik, St.Peter's Engineering College, Telangana, India, amjansrs@gmail.com

Received 2022 April 02**; Revised** 2022 May 20**; Accepted** 2022 June 18

Abstract:
The revenue from the loans directly accounts for the majority of the bank's profit. Additionally, one of the key characteristics of the banking industry is the credit danger.The vaticination of credit defaulters is one of the delicate tasks for any bank, but by vaticinating the loan defaulters, the banks surely may reduce their loss by reducing its non-profit means so that recovery of approved loans can take place without any loss and it can play as the contributing parameter of the bank statement. This makes the study of this loan eligibility vaticination important. Machine learning ways are veritably pivotal and useful in the vaticination of these types of data. In order to save a lot of money and bank sweats, we attempt to lessen the threat element that drives people to choose the secure person in this paper. This is accomplished by mining the Big Data belonging to the previously loan issued individuals, and based on this data, machine learning models were taught to produce the most accurate results. This paper's main goal is to determine which machine algorithm performs best at predicting whether or not a person is qualified for a loan.

Keywords: Prediction, Training, Testing,Loan, Machine Learning.

I. Introduction:

A loan is a third-party gift of money, property, or other tangible items in exchange for the subsequent early repayment of the loan amount including interest [4].Until the loan is returned, the grantee incurs a debt for which they are often liable to pay interest. Numerous people are seeking for bank loans as a result of the banking industry's progress, but because banks only have limited resources to distribute among their customers, it is usually best for them to take a chance on someone they know. Credit threat evaluation is, as we all know, critically important. Threat position computation can be done in a number of different methods [12].Even though the bank authorises the loan following a lengthy verification andvalidationprocess, there is still no guarantee that the grantee chosen is +secure. When done manually, this operation takes a long time. Every business that engages in lending encounters the difficulty of conformation

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2053-2062
https://publishoa.com
ISSN: 1309-3452

of loans on a regular basis. However, it can reduce the number of man-hours required and speed up client service. The improvement in customer satisfaction and cost savings by automating the loan conformation procedure are substantial [13].However, in order to remove the implicit threat, the bank must have a reliable model in place that allows it to directly read which customer loans it should accept and which it should reject. Only then will the benefits be realised. The credit system controlled by banks is one of the most crucial elements that determine our nation's thriftiness and financialsituation [14]. We can prognosticate whether or not a specific seeker is secure, and the entire validation process is done automatically using machine learning techniques. However, the appropriate characteristics include things like gender,education,dependents, income,loan amount, conjugal status, history of credit and others [9].If the company want to partially automate the loan qualifying procedure based on information furnished by the customer during the online operation form submission. The customer as well as a bank's retainer both benefit greatly from loan prediction. Even if many people are seeking for loans, it can be challenging to choose the sincere candidate who would pay back the loan [11]. Many misconceptions may arise when choosing candidates manually, and since accurate predictions are crucial for maximising returns, it's crucial to understand the differences between the various approaches and compare them [5].As a result, we are creating a method for automatically determining if a loan seeker is qualified by reviewing various machine learning models. Both clients and bank personnel will benefit from this. There will be a significant reduction in the loan approval process's timeframe.

The organization of this article is in following manner i.e. Section-II describes the research background, where we reviewed and analysed about all the literatures, Section-III denotes about the existing regime,Proposed methodology explained in Section-IV, results and discussions are demonstrated in Section-V and finally conclusions are noted in Section-VI.

II. Research Background:

Observation 1:

In their research paper titled "Prediction of Loan Risk using NB and Support Vector Machine," S. Vimala and K.C.Sharmili suggested a loan valuation model that utilised both Support Vector Machine and Naïve Bayesmethods [1]. A self-reliant presumption method called Naive Bayes includes probability propositions related to data stratification. SVM, on the other hand, stratifies predictions using a statistical learning model. In order to estimate the suggested system, a data set from the UCI depository containing 21 attributes was employed. The results of the trials showed that the merging of the Support vector Classifier and Naïve Bayes resulted in an efficient stratification of loan eligibility rather than the independent efficiencies of the classifiers.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2053-2062
https://publishoa.com
ISSN: 1309-3452

Observation 2:

Ranpreet Kaur and Anchal Goyal's research work, "Loan prediction using ensemble technique" offered a useful vaticination method that aids bankers in predicting the credit threat for consumers who have filed for loans [2]. The paper describes a prototype that organisations can use to decide correctly whether to accept or reject the applicants' loan request. In addition to three separate models, the paper employs the ensemble model, which combines the three models and analyses the credit threat for the best results(Support Vector Machine Model, Tree Model for inheritable Algorithmand Random Forest Network).

Observation 3:

The study "Overdue Prediction of Bank Loans Based on LSTM-SVM" by authors Xianzhong Long,Xin Li, Guozi Sun, Huakang Liand Geng Yang elaborates the current successful background, traditional threat soothsaying system and substantially introduces the main operation of the Long Short TermMemory-Support Vector Machine model in customer loan threat vaticination [6,15]. On this foundation, the LSTM framework and SVM system-based vaticination methodology is suggested, the vaticination outcomes are compared to those of the conventional approach, and the model's viability is validated. However, the LSTM-SVM system suggested in this paper has a number of shortcomings and has to be improved in subsequent research.

Observation 4:

The study paper by Sandip Pandit,MrunalSurve, Priya Shinde,Swati Sonawaneand Pooja Thitme, titled "Data mining techniques to analyse risk giving loan (bank)" primarily focuses on identifying and analysing the threat of providing a loan to commercial banks [7]. They have employed data mining methodology to determine risk when disbursing loans. It entails analysing and recycling data from various agencies in order to recapitulate it as priceless data. For forecasting the threat chance for a person to grant loans, they employed the C4.5 stratification algorithm.

Observation 5:

The authors of the study "A machine learning strategy for forecasting bank credit worthiness" Regina Esi,Turkson, Edward YeallakuorBaagyere, and Gideon Evans Wenya have delved into the plethora of machine learning models that may be used to predict a loan applicant's eligibility [8]. In order to discover which algorithms are the best fits for analysing bank credit data sets, they applied 15 different algorithms to the data set. Among the algorithms utilised are Linear Regression,Logistic Retrogression, Discriminant Analysis, Naive Bayes, K-Nearest Neighbor, Neural Networks, Decision Trees and Ensemble expert systems [15,16]. Excluding the Gaussian Naive Bayes andNearest Centroid, the trial showed that the remaining algorithms perform cogently well in accordance of accuracy and sundry performance indicators.

III. Existing Regime:

The existing regime, CIBIL, is a credit scoring system that gathers and keeps

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2053-2062
https://publishoa.com
ISSN: 1309-3452

track of information about individual and marketable reality payments related to loans and credit cards. Every month, banks and other lenders send these records to TransUnion, a credit reporting firm. With the use of this data, a CIBIL Score and profile for each individual is created. A three-number numerical representation of your credit history is known as a CIBIL Score. The CIBIL Report's credit history is used to calculate the score. A CIR is a record of a person's credit payment history over a period of time across several loan kinds and credit agencies. Your funds, financing, or fixed income securities are not mentioned in a CIR [3].This allows lenders to organise and authorise loan operations. While the CIBIL Score is vital in the loan application process, this doesn't always provide a full view. The CIBIL Score is the issuer's first conviction; the higher the score, the more likely the loan will be evaluated and sanctioned.The lender alone must decide whether to advance, and they must also take into account a number of other variables. The loan or credit card should not be sanctioned, according to CIBIL. Based on the information supplied by the customer, such as their gender, conjugal status, earnings,quantity of dependents, loan amount, degree,history of credit, etc., the lender must manually analyse each application and determine whether it is creditworthy.

IV. Proposed Methodology:

Here, under this approach, we determine a client's eligibility for a loan from a bank and contrast various models to determine which one is the best fit for this use. In order to do this, we created an automatic loan predication platform using machine learning approach. We trained the machine using the prior dataset so that it could analyse and comprehend the process. It will also determine whether the loan applicant is eligible. For the purpose of determining the best functional model for determining loan eligibility, five learning algorithms are assessed. Thirteen aspects make up the dataset we're using: Loan ID, Gender, Married, Dependents, Education, Self-Employed, Applicant Income, Co-applicant Income, Loan Amount, Loan Amount Term, Credit History, Property Area, and Loan Status.

The following inflow illustration demonstrates how the proposed system contains various stages. In order to facilitate communication between the user and the system, a website must be developed.

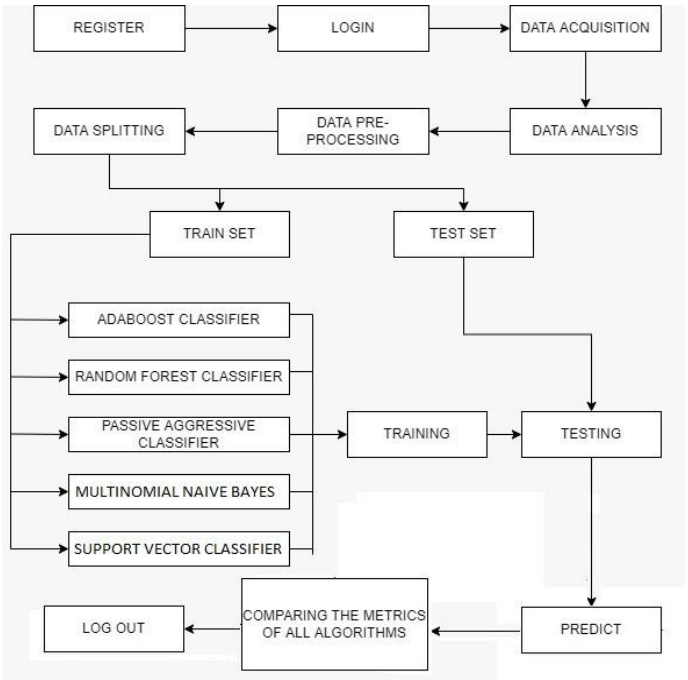


Figure 4.1. Flow diagram for the proposed methodology

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2053-2062
https://publishoa.com
ISSN: 1309-3452

The structure of the process described here is seen in figure 4.1 above.

1. **Register:**
   The user registers by providing information such their first name, password, email, and other details that are recorded in the SQLite database.
2. **Login:**
   The user may also log in using their username and password, which, if they are accurate, will take them to the website's homepage.
3. **Data collection:**
   Next, the dataset required for training and test sets is acquired from the internet.
4. **Data Analysis:**
   Using descriptive statistics and visualizations, we carry out original examinations on data to look for trends, identify deviations, put theories to the test, and confirm suppositions
5. **Data Pre-processing:**
   The raw data which has been acquired during the data accession process is converted into a form that can be understood by the machine.
6. **Data Splitting:**
   The dataset is partitioned into two non-overlapping sets, one with a stratification column and one without it.
7. **Test Set:**
   A reasonable sampling of the dataset's data that is used to give an unprejudiced evaluation of a final model fit on the training dataset.
8. **Train Set:**
   It's a portion of a data set used to fit a model for the vaticination or stratification of values that are familiar in the training set, but unknown in other (hereafter) data.

a) Instantiating the algorithms:

Starting the training process by feeding the train data to the various algorithms, each of which can receive input and produce results. The various algorithms this system employs include:

i. Adaboost Classifier:

During the data training phase, the Adaboost Classifier creates a specific amount of decision trees known as stumps. The incorrectly classified record in the first model is given precedence as the first decision tree / model is constructed. For the alternative model, just these records are sent as input. Until we define the quantity of base learners we would like to produce, the procedure continues. The following formula is used to update sample weights:

Current Weight of Sample = Weight of Sample * e^ (Performance)-------Equation- 1

ii. PassiveAggressive Classifier:

Belonging to the order of online learning algorithms in machine learning, responds passively in cases of proper stratification and aggressively in cases of misinterpretation.

iii. Random Forest Classifier:

In order to increase the predictive accuracy of the supplied dataset, this classifier employs several decision trees across numerous sections of the data. The following information gain calculation serves as the foundation for the data splitting:

Gain(V, Y) = Entropy(V)-Entropy(V, Y)----------Equation- 2

Where,

- V is the target variable
- Y is a point to be resolved on

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2053-2062
https://publishoa.com
ISSN: 1309-3452

- The entropy determined when the data is resolved on point X is called entropy (V, Y).

iv. Multinomial Naïve Bayes:

Based on the Bayes theorem, this method forecasts the label of a word that resembles a dispatch or review composition. It determines the likelihood of each label for a particular sample and outputs the label with the highest likelihood. The chance of the label appearing in the word can be determined using the formula below. P(E|X) = P(E)*P(X|E)/ P(X). ------- Equation- 3

Where,

- P(X) is predictor X'sis prior probability
- P(E) is class E's prior probability
- P(X| E) is the eventuality of predictor X given class E probability

v. Support Vector classifier:

A Linear SVC (Support Vector Classifier) should fit the data you provide and deliver a "pre-eminent fit" hyperplane that partitions or classifies your data. After obtaining the hyperplane, one could additionally provide their classifier with a few features to determine what the "prognosticated" class is.

9.Training:

Algorithms for machine learning get knowledge from data. From the training data learners are provided, they establish connections, gain understanding, form views, and gauge their level of confidence.

10.Testing:

After the algorithm has been trained, we test it using data from the training set. These results allow us to assess the performance of each approach.

11.Predict:

The models that have been trained using the dataset also prognosticate whether the client could acquire a loan or not.

12.Comparing the metrics of all algorithms:

Recall, Precisionand Accuracy are the three primary criteria used to estimate a stratification model. In this case, we use accuracy, which is determined by the formula below.

Accuracy=correct number of predictions/total number of predictions---------- Equation- 4

13. Logout:

If we click on the logout link it'll come out from the front page of the website.

V. Results Analysis:

The models trained on the dataset performed distinctly as given by the accuracies below:

Random Forest Classifier:

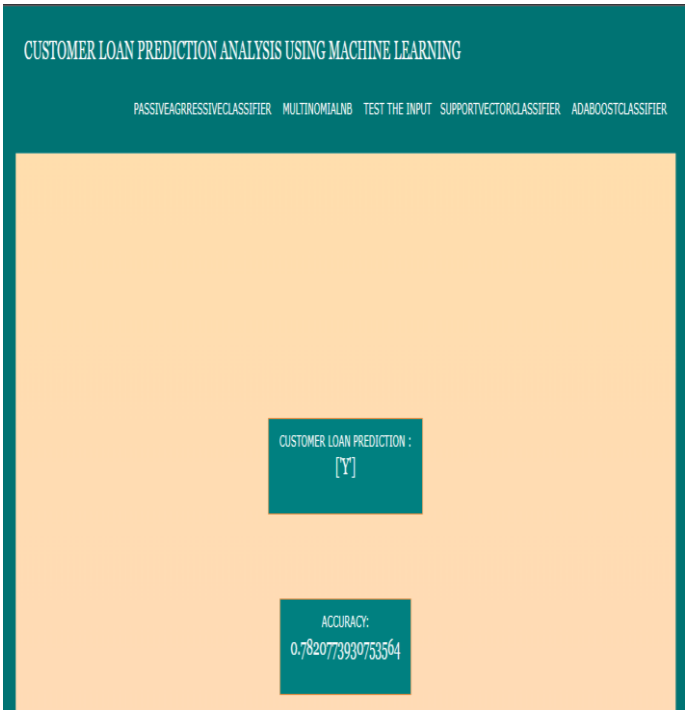


Figure 5.1.TheRandom Forest Classifier's Accuracy on the Dataset

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2053-2062
https://publishoa.com
ISSN: 1309-3452

Passive Aggressive classifier:

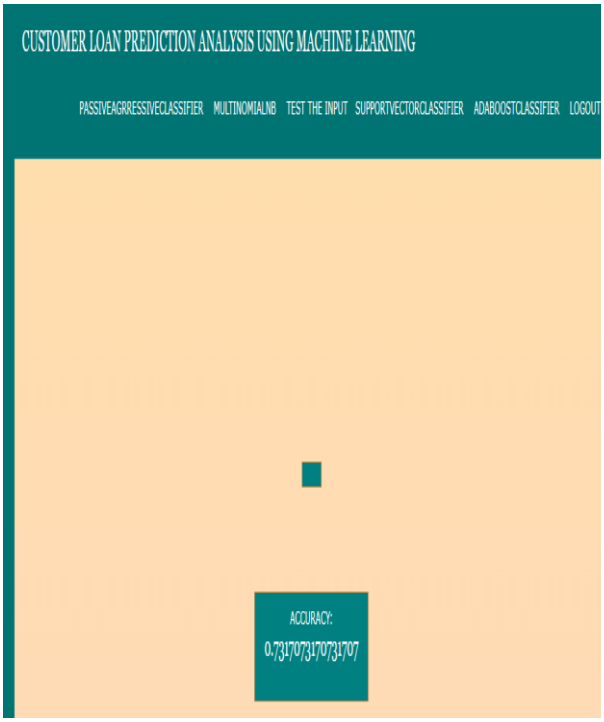


Figure 5.2.The Passive Aggressive Classifier's Accuracy on the Dataset

Support Vector Classifier:

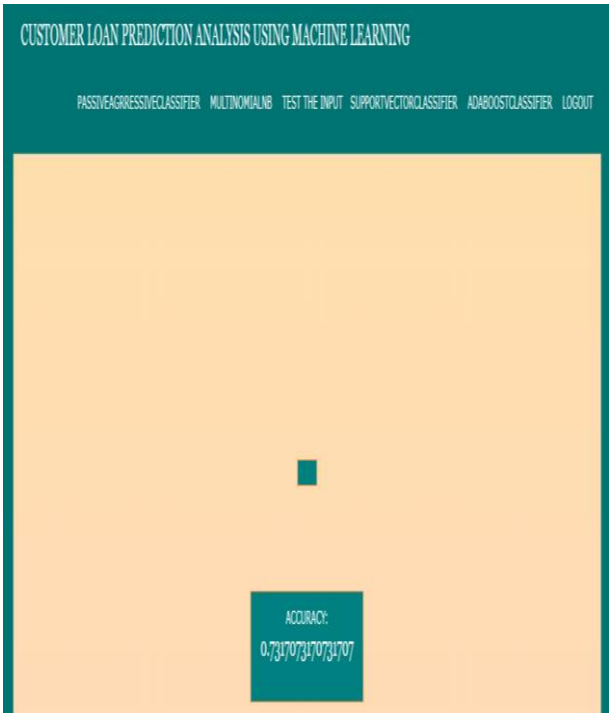


Figure 5.4.The Support Vector Classifier's Accuracy on the Dataset

Multinomial Naïve Bayes:

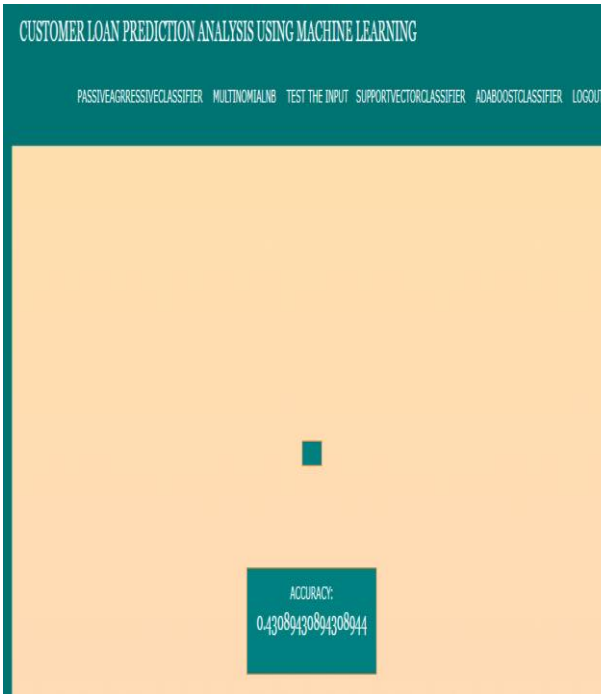


Figure 5.3.The Multinomial Naïve Bayes's Accuracy on the Dataset

Adaboost Classifier:

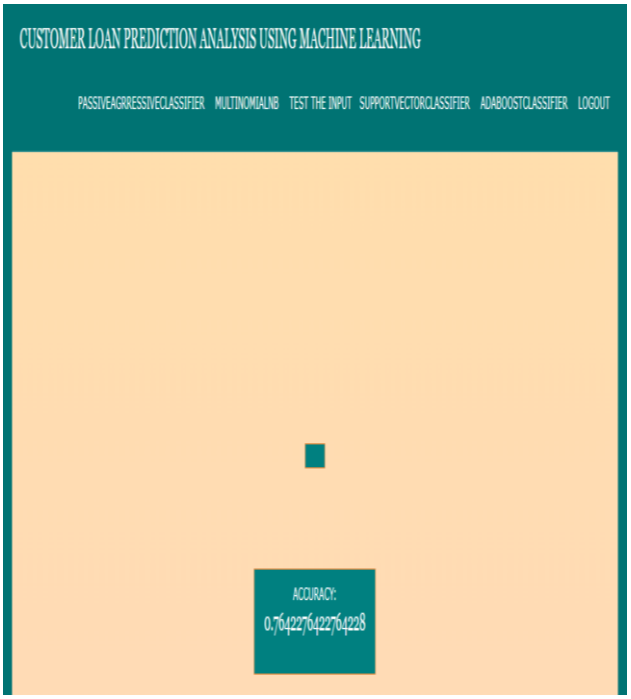


Figure 5.5.The Adaboost Classifier's Accuracy on the Dataset

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2053-2062
https://publishoa.com
ISSN: 1309-3452

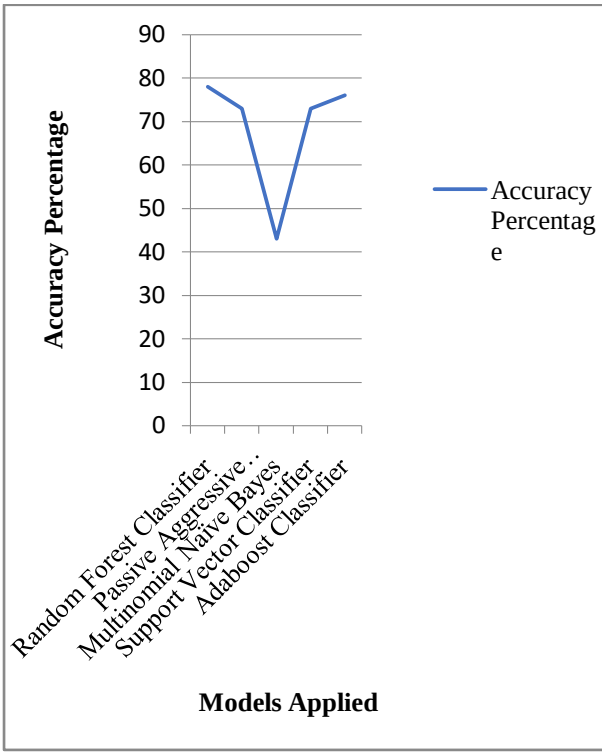


Figure 5.6.The accuracy percentages of the five algorithms depicted in a line chart.

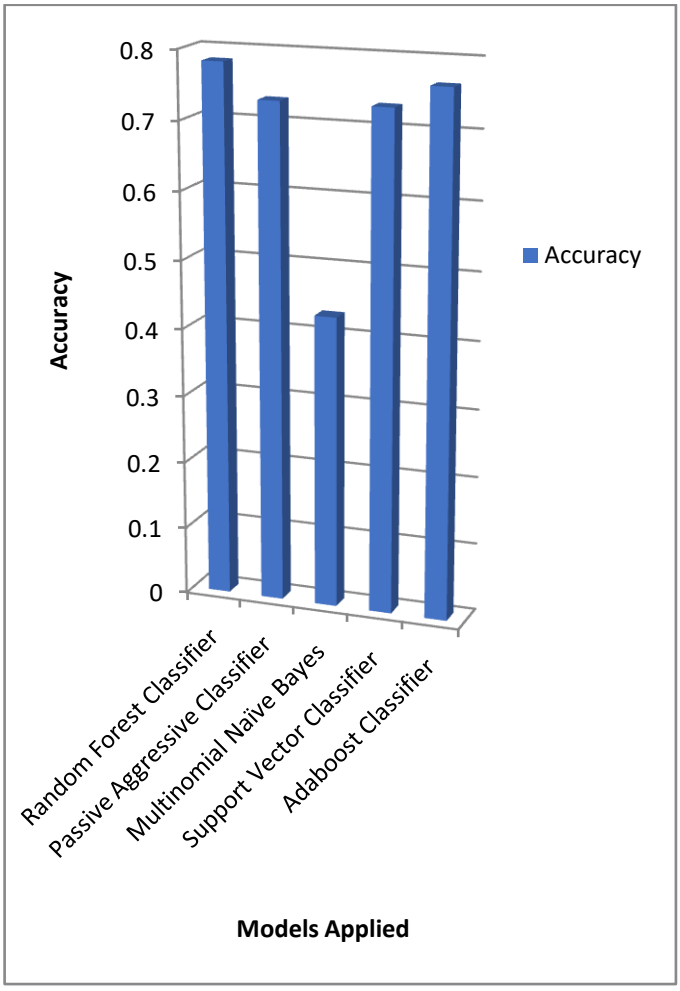


Figure 5.7.The accuracies of the five algorithms depicted in a Bar Graph.

Table 1: Accuracies of the five algorithms on the given dataset

S. No	Algorithm used	Accuracy
1	Random Forest Classifier	0.7820773930753564
2	Passive Aggressive Classifier	0.7317073170731707
3	Multinomial Naïve Bayes	0.43089430894308944
4	Support Vector classifier	0.7317073170731707
5	Adaboost Classifier	0.7642276422764228

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p.2053-2062
https://publishoa.com
ISSN: 1309-3452

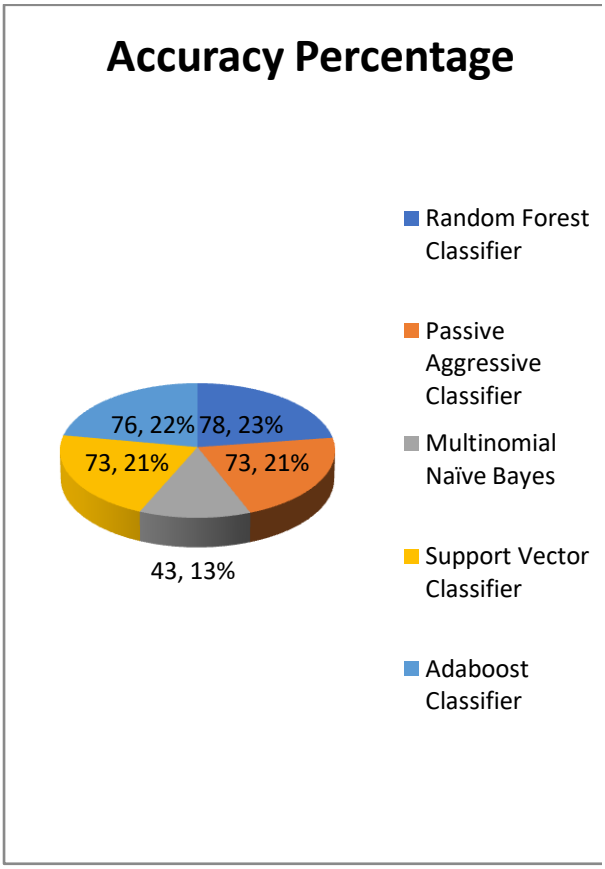


Figure 5.8.The accuracy percentages of the five algorithms depicted in a Pie chart.

VI. Conclusion:

To identify a dominant model, we have provided various machine learning models utilising various machine learning techniques. The model built using the Random Forest Algorithm surpassed all other models in correctly classifying the data with a rate of 78 percent, according to the examination of the results. It is safe to say that the solution is a generally effective element after a proper review of the element's advantages and limitations. This service is operating duly and in compliance with banking standards. This construct is simple to plug into a plethora of other systems. These results, in our opinion, will help researchers learn more about the topic of developing a vaticination study that can predict a client's loan eligibility.

Acknowledgement:

We are grateful to St.Peter's Engineering College, department of CSE for helping us with the laboratory and for continuing support to prepare this paper in a brighter manner.

References:

[1] S. Vimala, K.C. Sharmili, "Prediction of Loan Risk using NB and Support Vector Machine", International Conference on Advancements in Computing Technologies (ICACT 2018), vol. 4, no. 2, pp. 110-113, 2018.

[2] Anchal Goyal, Ranpreet Kaur, "Loan Prediction Using Ensemble Technique", International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no. 3, March 2016.

[3] "Understand Your Credit Score and Report" [online] Available: https://www.cibil.com/faq/understand-your-credit-score-and-report.

[4]Julia Kagan, "loan", april2021, [online] Available: https://www.investopedia.com/terms/l/loan.asp.

[5] Ashwini S. Kadam, Shraddha R. Nikam, Ankita A. Aher, Gayatri V. Shelke, Amar S. Chandgude, "Prediction for Loan Approval using Machine Learning Algorithm", International Research Journal of Engineering and Technology (IRJET), Volume:08, Issue: 04, Apr 2021.

[6] Xin Li, Xianzhong Long, Guozi Sun, Geng Yang, and Huakang Li "Overdue Prediction of Bank Loans Based on

LSTM-SVM" 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2018, pp. 1859-1863, doi: 10.1109/SmartWorld.2018.00312.

[7]MrunalSurve, Pooja Thitme, Priya Shinde, Swati Sonawane, and Sandip Pandit, "Data mining techniques to analyze risk giving loan (bank)", Internation Journal of Advance Research and Innovative Ideas in Education Volume 2, Issue 1, 2016.

[8] Turkson, Regina Esi, Edward YeallakuorBaagyere, and Gideon Evans Wenya, "A machine learning approach for predicting bank credit worthiness.", 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR). IEEE, 2016.

[9] L. Udaya Bhanu, Dr. S. Narayana, "Customer Loan Prediction Using Supervised Learning Technique",International Journal of Scientific and Research Publications, Volume 11, Issue 6, June 2021.

[10] Mohammad Ahmad Sheikh, Amit Kumar Goel,Tapas Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm", International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020.

[11] Dr.C K Gomathy, Ms.Charulatha,Mr.AAkash ,Ms.Sowjanya,"THE LOAN PREDICTION USING MACHINE LEARNING",International Research Journal of Engineering and Technology (IRJET), Volume: 08, Issue: 10, Oct 2021.

[12] J. Tejaswini1, T. Mohana Kavya, R. Devi Naga Ramya, P. Sai Triveni Venkata Rao Maddumala, "Accurate loan approval prediction based on machine learning approach", www. jespublication.com, page 523 Issue 4, April/ 2020 ISSN NO: 0377-9254.

[13] V. C. T. Chanetal., "Designing a Credit Approval System Using Web Services, BPEL, and AJAX", 2009 IEEE International Conference on e-Business Engineering, Macau, 2009, pp. 287-294.doi: 10.1109/ICEBE.2009.46.

[14] Nitesh Pandey, Ramanand Gupta, Sagar Uniyal, Vishal Kumar, "Loan Approval Prediction using Machine Learning Algorithms Approach", IJIRT, Volume 8, Issue 1,2021.

[15]Amjan Shaik, et al, "Sentiment Extraction and analysis using Machine Learning Tools: Survey", IOP Conference series: Material Science & Engineering, SCOPUS, December 2018.

[16]Amjan Shaik, et al, "Analysis of effective medical record storage formats and demonstration of time efficient secure storage framework", in European Journal of Molecular & Clinical Medicine (EJMCM), Volume: 7, Issue 6, Pages: 2744-2763, ISSN: 2515-8260, December 2020.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:01/January-2021 Impact Factor- 5.354 www.irjmets.com

A STUDY ON MACHINE LEARNING ALGORITHM FOR ENHANCEMENT OF LOAN PREDICTION

Prateek Dutta[*1]

[*1]Student, B.tech Artificial Intelligence, G.H. Raisoni College of Engineering, India.

ABSTRACT

In the lending industry, investors offer loans to lenders for the purpose of repaying interest. If the borrower pays the loan, then the lender will make a profit on the interest. However, if the borrower fails to repay the loan, the lender loses the loan. Therefore, lenders face the problem of predicting the risk of the borrower not being able to repay the loan. The main purpose of this project is to predict which of the customers will be repaid with their loans or not.

Keyword: Loan, Machine Learning, Kaggle, Regression.

I. INTRODUCTION

1.1 What is Machine Learning?

Machine learning is a subset of Artificial Intelligence that allows a computer program to automatically learn from a previous task. It works by analysing data, identifying patterns, and incorporating minimal human interventions. Almost any work that can be completed with a data-defined pattern or set of rules can be done with machine learning machines. This allows companies to change processes that were previously only possible for people to make assumptions that respond to customer service calls, bookkeeping, and reviews.

1.2 Types of Machine Learning

Machine Learning has been further classified into four types. They are categories as follow:

- Supervised Machine Learning
- Unsupervised Machine Learning
- Semi-supervised Machine Learning
- Reinforcement Machine Learning

1.3 Importance

Machine learning algorithms enable the construction of a new model using previously unknown historical data that can be used to train the model to make better predictions not only for credit risks, but also for other risks such as early payment opportunities leading to loss of income from interest, existing withdrawal risks etc. With a good model, financial institutions can predict the likelihood of a customer repaying a loan before the maturity date and then have procedures with pre-defined prevention measures to be taken, prior to this occurrence.[1]

Dataset

It is the collection of data arranged in rows and columns. It can be of any form but mostly it is found in csv format. A single dataset can contain numerous number of data. The dataset for this project on which this paper deals with, has been taken from Kaggle. It contain two dataset, one of them is training and another is testing. It contains 13 columns labelled as: Loan_ID, Gender, Married, Dependents, Education, Self_Employed, Applicant Income, Co-applicant Income, Loan Amount, Loan_Amount_tearm, Credit_History, Property_Area, Loan_Status.

II. ALGORITHM USED IN LOAN PREDICTION

In this paper supervised classification problem to be trained with algorithms like:

1. Logistic Regression
2. Decision Tree
3. Random Forest

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:01/January-2021 Impact Factor- 5.354 www.irjmets.com

In this project, default hyperparameter values are employed. More visualization can be done beyond what's executed in this post.

In this model, mainly five libraries has been used as, pandas, NumPy, matplotlib, seaborn and, sklearn. These libraries has been widely used in order to get the well defined and categorized result.

The identifiable machine learning separator is not limited to the above. Other models such as XGBoost, CatBoost and likes can be used in model training. The choice of these three algorithms in a row over the desire to keep the model defining itself again, the database is small.

Correlating Attributes

Based on the combination between the symbols it has been observed that they are more likely to repay their loans. Independent and important qualities can include Property location, education, loan amount, and last in credit History, i.e. from intuition it is considered important. Meeting in between attributes can be seen using corplot and boxplot in the Python platform.[2]

Steps Involved

⇨ So firstly you need to import the libraries (fig.-1)

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        %matplotlib inline
```

Fig.-1

⇨ Then insert the dataset into the environment and read it.(fig.-2)

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0

Fig.-2

⇨ Now check the missing values and fix them.(fig.-3)

	Total	Percent
Credit_History	50	0.081433
Self_Employed	32	0.052117
LoanAmount	22	0.035831
Dependents	15	0.024430
Loan_Amount_Term	14	0.022801
Gender	13	0.021173
Married	3	0.004886
Loan_Status	0	0.000000
Property_Area	0	0.000000
CoapplicantIncome	0	0.000000
ApplicantIncome	0	0.000000
Education	0	0.000000
Loan_ID	0	0.000000

Fig.-3

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:01/January-2021 Impact Factor- 5.354 www.irjmets.com

⇨ Now analyse the data and visualise it with respect to each column. It will results in graphical format(fig.- 4)

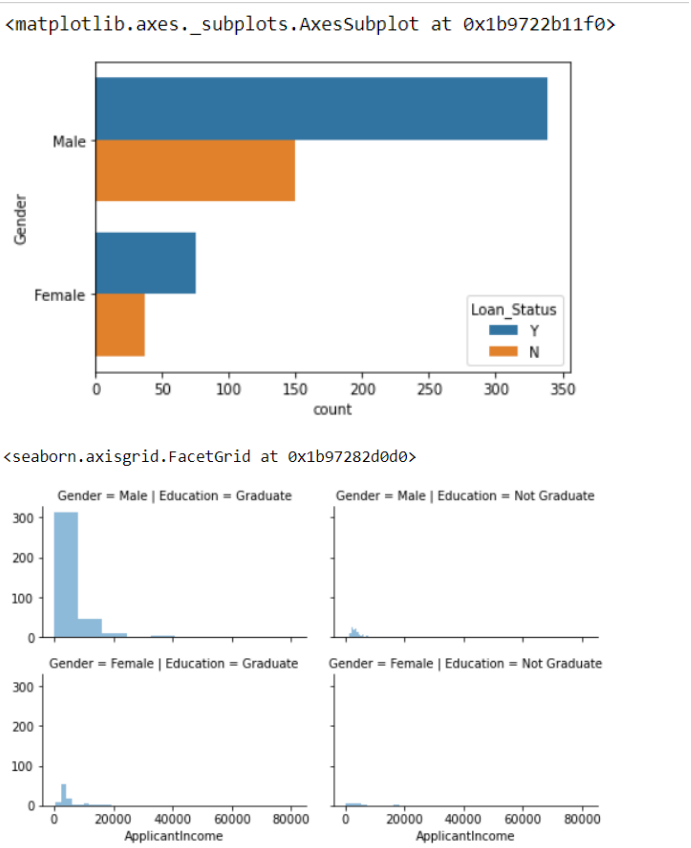


Fig.-4

⇨ Now next step is to encoding the numeric data and get it ready for training. (fig.-5)

```
data_train['Dependents'].value_counts()
```

```
0     345
1     102
2     101
3+     51
Name: Dependents, dtype: int64
```

```
data_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 614 entries, 0 to 613
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Loan_ID            614 non-null    object
 1   Gender             601 non-null    object
 2   Married            611 non-null    object
 3   Dependents         599 non-null    object
 4   Education          614 non-null    object
 5   Self_Employed      582 non-null    object
 6   ApplicantIncome    614 non-null    int64
 7   CoapplicantIncome  614 non-null    float64
 8   LoanAmount         592 non-null    float64
 9   Loan_Amount_Term   600 non-null    float64
 10  Credit_History     564 non-null    float64
 11  Property_Area      614 non-null    object
 12  Loan_Status        614 non-null    object
dtypes: float64(4), int64(1), object(8)
memory usage: 62.5+ KB
```

Fig.-5

⇨ Then project the heatmap Showing the correlations of features with the target. (fig.-6)

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:01/January-2021 Impact Factor- 5.354 www.irjmets.com

```
<matplotlib.axes._subplots.AxesSubplot at 0x1b975b5b250>
```

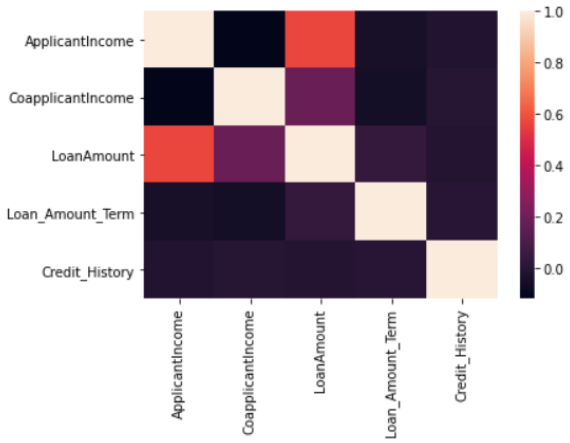


Fig.-6

III. RESULT

⇨ Now you can check the evaluation of the model using any of the three algorithm.

⇨ Logistic Regression (fig.-7)

```
model = LogisticRegression()
```

```
model.fit(X_train, y_train)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```

```
ypred = model.predict(X_test)
print(ypred)
```

```
[1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 1 1 1 1 1 0 1
 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 0 0 1 1 1 1 1 0 1]
```

```
evaluation = f1_score(y_test, ypred)
evaluation
```

```
0.8979591836734695
```

Fig.-7

⇨ Decision Tree (fig.-8)

```
tree = DecisionTreeClassifier()
tree.fit(X_train, y_train)
```

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
```

```
ypred_tree = tree.predict(X_test)
print(ypred_tree)
```

```
[0 1 1 0 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 0 0 0 1 1 0 0 1 1 1 0 0 1 0 1 1 1 0 1
 1 1 1 0 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 1 0 0 1 0 0 1 1 1 1 1 1 1 1 0
 1 1 0 1 1 0 1 1 1 1 1 0 1 0 1 1 1 1 1 0 1 0 1 1 1 1 0 1 0 1 1 1 0 1 1 1 1 0
 1 1 1 0 0 0 1 1 0 1 0 0]
```

```
evaluation_tree = f1_score(y_test, ypred_tree)
evaluation_tree
```

```
0.7745664739884394
```

Fig.-8

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:01/January-2021 Impact Factor- 5.354 www.irjmets.com

⇨ Random Forest classifier (fig.-9)

```
]: forest = RandomForestClassifier()
   forest.fit(X_train, y_train)

]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                          max_depth=None, max_features='auto', max_leaf_nodes=None,
                          min_impurity_decrease=0.0, min_impurity_split=None,
                          min_samples_leaf=1, min_samples_split=2,
                          min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1,
                          oob_score=False, random_state=None, verbose=0,
                          warm_start=False)

]: ypred_forest = forest.predict(X_test)
   print(ypred_forest)

   [1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 0 0 1 0 1 1 1 0 1
    1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 0 1
    1 1 0 1 1 0 0 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 0 1 1 1 0 1 1 0 1 0 1
    1 1 1 0 0 1 1 1 1 1 1 0 1]

]: evaluation_forest = f1_score(y_test, ypred_forest)
   evaluation_forest

]: 0.8540540540540541
```

Fig.-9

IV. CONCLUSION

From Test Data Analysis, we can generate insights from data. It can be seen in the trial of three models that Logistic Regression did better with evaluation of 89.7059% than others, Random Forest(77.4566%), Decision Tree(85.4054%). And last Logistic regression can be considered as best algorithm for Loan prediction using Machine Learning. Applicants with a low credit history fail to be accepted, probably because they have a chance of not paying. Most of the time, Applicants with a high fee are likely to be eligible for a rebate, which is likely to repay their loan. A particular sexual orientation and marital status seems beyond the reach of the company.

Conflict of Interest Statement

I Prateek Dutta, authors, state that there is no conflict of interest in my manuscript.

V. REFERENCE

[1] Zoran Ereiz. Predicting Default Loan using Machine Learning (OptiML), https://www.researchgate.net/publication/338948751_Predicting_Default_Loans_Using_Machine_Learning_OptiML

[2] Cowell, R.G., A.P., Lauritez, S.L., and Spiegelhalter, D.J.(1999). Graphical models and Expert Systems. Berlin: Springer.

[3] S. Shanmugan, S. Ravichandran, A Fuzzy Point approach the Solar Still Performances an Experimental Investigation, 2020 6(1)681-689. DOI Member: 10.6084/m9.jetir.JETIRDY06108

[4] K. Manikandan, S. Shanmugan, R. Ashish kumar, Vadde Venkat Harish, T. Jansi Rani, T. Nishanthi, Trans membranous fetal movement and pressure sensing. Materials Today: Proceedings, Available online 13 May 2020. https://doi.org/10.1016/j.matpr.2020.04.497

[5] J. Jennifer, Monica Nathasha Marrison, J. Seetha, S. Sivakumar and P. Sathish Saravanan, "DMMRA: Dynamic Medical Machine for Rural Areas", IEEE 2017 International Conference On Power And Embedded Drive Control (ICPEDC), pp. 461-471, 16th - 18th March 2017. DOI: 10.1109/ICPEDC.2017.8081135

[6] P Arokianathan, V Dinesh, B Elamaran, M Veluchamy and S Sivakumar, "Automated Toll Booth and Theft Detection System", IEEE 2017 Technological Innovations in ICT for Agriculture and Rural Development (TIAR), pp. 84-88, 07th - 08th April 2017. DOI: 10.1109/TIAR.2017.8273691

[7] Videla, Lakshmi Sarvani, et al. "Modified Feature Extraction Using Viola Jones Algorithm". Journal of Advanced Research in Dynamical and Control Systems. Volume 10, Issue 3 Special Issue, 2018, Pages 528-538

[8] Sreedevi E., PremaLatha V., Prasanth Y. and Sivakumar S., "A Novel Ensemble Learning for Defect Detection Method With Uncertain Data", Applications of Artificial Intelligence for Smart Technology, pp. 1-13, 2021. doi : 10.4018/978-1-7998-3335-2.ch005

[9] Premalatha V., Vineesha K. and Srinivasarao M., "International Journal of Scientific and Technology Research", 2020, 9(1), pp. 1005-1008.

RESEARCH ARTICLE OPEN ACCESS

Loan Prediction using Machine Learning Algorithms

Sanket Bhattad [1], Sumit Bawane [2], Shweta Agrawal [3], Unnati Ramteke [4], Dr. P. B. Ambhore [5]

[1],[2],[3],[4] B.Tech student, Dept. of IT, Gossvernment college of Engineering, Amravati
[5] Assistant Professor, Dept. of IT, Government college of Engineering, Amravati - Maharashtra

ABSTRACT
In India, the number of people or organization applying for loan is increased every year. The bank employees have to put in a lot of work to analyse or predict whether the customer can pay back the loan amount or not (defaulter or non-defaulter) in the given time. The aim of this paper is to find the nature, background, or credibility of the client that is applying for the loan. We use exploratory data analysis technique to deal with the problem of approving or rejecting the loan request or in short loan prediction. The focus of this paper is to determine whether the loan given to a particular person or an organization shall be approved or not.

Keywords: - Loan, Prediction, Machine Learning, Training**.**

I. INTRODUCTION

Distribution of the loans is the core business part of almost every bank. The main portion the bank's asset is directly came from the profit earned from the loans distributed by the banks. Today many banks/financial companies approves loan after a regress process of verification and validation but still there is no surety whether the chosen applicant is the deserving right applicant out of all applicants. Through this system we can predict whether that particular applicant is safe or not and the whole process of validation of features is automated by machine learning technique.

Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this Paper is to provide quick, immediate and easy way to choose the deserving applicants. It can provide special advantages to the bank. The Loan Prediction System can automatically calculate the weight of each features taking part in loan processing and on new test data same features are processed with respect to their associated weight .A time limit can be set for the applicant to check whether his/her loan can be sanctioned ornot. Loan Prediction System allows jumping to specific application so that it can be check on priority basis. This Paper is exclusively for the managing authority of Bank/finance Company, whole process of prediction done privately no stakeholders would be able to alter the processing. Result against particular Loan Id can be send to various departments of banks so that they can take appropriate action on application. This helps all others department to carried out other formalities.

II. DATA SET

A collection of data is taken from the banking sector. The Data set is in ARFF (Attribute-Relation File Format) format that is acceptable by Weka. ARFF file is composed of tags that include the name, types of attributes, values and data itself. For this paper, we are using 12 attributes like gender, marital status, qualification, income, etc.

Table-1: Data set variables along with description and type

Variable Name	Description	Type
Loan_ID	Unique ID	Integer
Gender	Male/Female	Character
Marital_Status	Applicant married(Y/N)	Character
Dependents	Number of Dependents	Integer
Education_Qualification	Graduate/Under Gradute	String
Self_Employed	Self-employed(Y/N)	Character
Applicant_Income	Applicant income	Integer
Co_Applicant_Income	Co-applicant income	Integer
Loan_Amount	Loan amount in thousands	Integer
Loan_Amount_Term	Term of loan in months	Integer
Credit_History	Credit history meets guidelines	Integer
Property_Area	Urban/Semi urban/Rural	String
Loan_Status	Loan Approved(Y/N)	Character

Now in machine learning model, we first apply the training data set, in this data set the model is trained with known examples. The entries of new applicants will act as a test data which are to be filled at the time of submitting the application. After performing such tests, model can determine whether the loan approved to the person is safe or

not basically about the loan approval on the basis of the various training data sets.

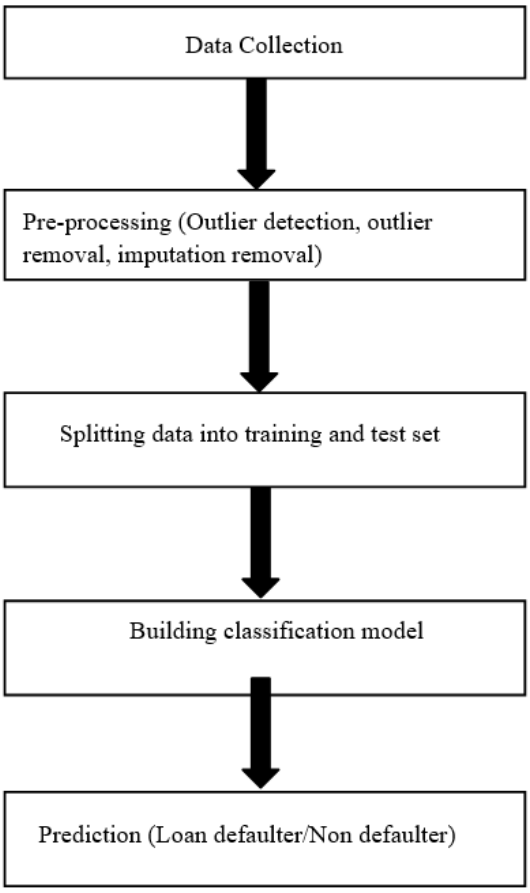


Fig-1: Chronology of Data

The diagram above gives us an outline on how data is used in this machine learning process or model.

Basically, it is divided into four parts in which we use data to predict the outcome of the whole process. First, we use training data set to train our model. After the model is trained, then we test it with unknown examples from the same scenario.

Another process that we use before testing and training data is data pre-processing. In data pre-processing we remove all sorts of values that can cause an error like redundant values, incomplete values, missing data, etc.

III. LOANPREDICTION METHODOLOGY

The diagram 2 represents the working of our model. It basically gives us a rough idea on how the loan prediction system works. After collecting data, we use feature selection process on data. Feature selection can be defined as a process of reducing number of input variables when we develop a predictive model.

Feature selection is divided into two parts i.e. supervised method and unsupervised method. Supervised method is divided into three parts which are wrapper, filter and intrinsic. In supervised method we use target variable to remove discrepancies in data. While in unsupervised method we do not use target variable to remove discrepancies. Unsupervised method uses the process of correlation

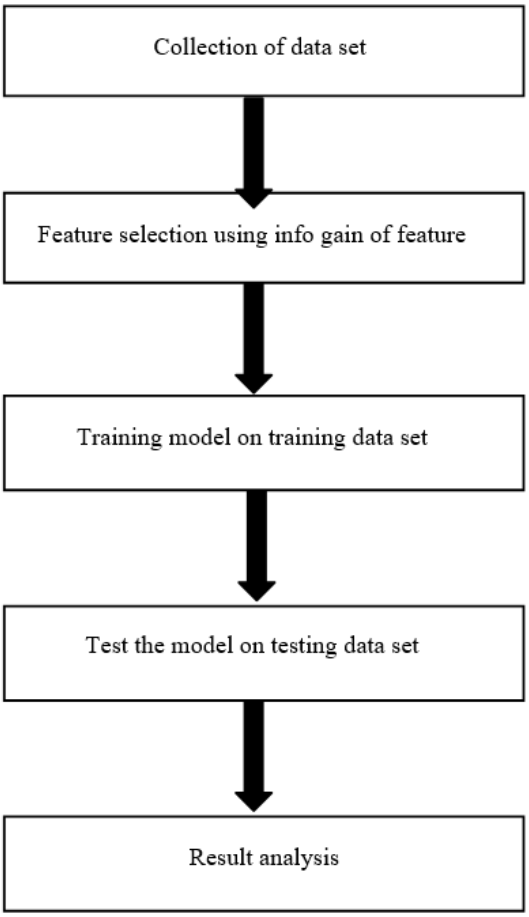


Fig-2: Loan Prediction Methodology

IV. EXPLORATORY DATA ANALYSIS

1. 80% applicants are male 20% are female.
2. 80% are not self-employed.
3. 60% are married and 85% have repaid their debts.
4. Most of the Applicants have no dependents.
5. Around 80 % Applicants are graduates.
6. Majority of Applicants are from Semi urban area.
7. Distribution of Applicant income is towards left which means it is not normal distribution. This can be attributed to

Income Disparity in society Driven by the fact that People have different education levels.

8. Proportion of male and female applicants is the same for approved as well as unapproved loans.

9. Proportion of married applicant is more for approved loans.

10. Distribution for applicants having 1 to 3+ dependents is same across both the categories.

11. If co-applicant income is less then less chances of loan approval.

12. More chances of approval for low and average loan amount as compared to high loan amount.

13. the most correlated variables are applicant income and loan amount & credit History and Loan status.

V. MODEL USED FOR TRAINING AND TESTING

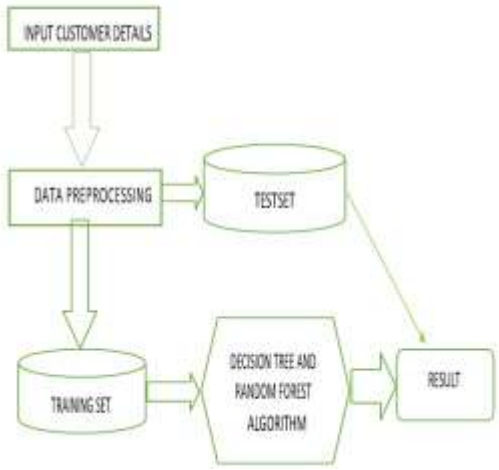


Fig- 3: Training and testing

VI. MACHINE LEARNING METHODS

Three machine learning classification models are used for the prediction of application that can be used in android applications. The brief description of each model is explained below.

1. Decision Tree

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin toss comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

Decision tree is a type of supervised learning algorithm having predefined target variable that is mostly used in classification problems. In this technique we spilt the population or sample into two or more homogenous sets based on the most significant splitter/differentiator in the input variables

Decision tree uses multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words the purity of node increases with respect to the target variable.

The accuracy of this algorithm is 77%.

2. Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many extensions that are more complex exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). With the confidence factor of c=1.0 the best accuracy is 78.91%

3. Random Forest

Random forest or random decision forests are an ensemble learning method used for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

We have done several trials with Random Forest with different parameters: executions with supervised and unsupervised discretization's (equal-frequency and equal-width), with all attributes. In the experiments without attribute selection the best result was 80.20%.

VII. CONCLUSION

The main purpose of the paper is to classify and analyse the nature of the loan applicants. From a proper analysis of available data and constraints of the banking sector, it can be concluded that by keeping safety in mind that this product is much effective or highly efficient. This application is operating efficiently and fulfilling all the major requirements of Banker. Although the application is flexible with various systems and it can be plugged effectively.

This paper work can be extended to higher level in future so the software could have some better changes to make it more reliable, secure, and accurate. Thus, the system is trained with present data sets which may be older in future so it can also take part in new testing to be made such as to pass new test cases.

There have been numbers cases of computer glitches, errors in content and most important weight of features is fixed in

automated prediction system. So, in the near future the so – called software could be made more secure, reliable and dynamic weight adjustment. In near future this module of prediction can be integrated with the module of automated processing system.

REFERENCES

[1] J. R. Quinlan. Induction of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086.

[2] A. Goyal and R. Kaur, "A survey on Ensemble Model for Loan Prediction", International Journal of Engineering.

[3] G. Shaath, "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R".

[4] A. Goyal and R. Kaur, "Accuracy Prediction for Loan Risk Using Machine Learning Models".

[5] https://www.experian.com/blogs/ask- experian/credit-education/score-basics/what-is-a- good-credit-score/

[6] https://machinelearningmastery.com/types-of-classification-in-machine-learning/

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 9, Issue 1, September 2021
Impact Factor: 5.731

Analysis of Loan Availability using Machine Learning Techniques

Sharayu Dosalwar[1], Ketki Kinkar[2], Rahul Sannat[3], Dr Nitin Pise[4]
UG Student, School of Computer Engineering & Technology[1,2]
UG Student, School of Electronics & Communication Engineering[3]
Professor, School of Computer Engineering & Technology[4]
MIT World Peace University, Pune, Maharashtra, India

Abstract: *In the banking system, banks have a variety of products to provide, but credit lines are their primary source of revenue. As a result, they will profit from the interest earned on the loans they make. Loans, or whether customers repay or default on their loans, affect a bank's profit or loss. The bank's Non-Performing Assets will be reduced by forecasting loan defaulters. As a result, further investigation into this occurrence is essential. Because precise forecasts are essential for benefit maximisation, it's crucial to analyse and compare the various methodologies. The logistic regression model is an important predictive analytics tool for detecting loan defaulters. In order to assess and forecast, data from Kaggle is acquired. Logistic Regression models were used to calculate the various performance indicators. The models are compared using performance metrics like sensitivity and specificity. In addition to checking account details (which indicate a customer's wealth), the model is significantly better because it includes variables (customer personal attributes such as age, objective, credit score, credit amount, credit period, and so on) that should be considered when correctly calculating the probability of loan default. As a result, using a logistic regression approach, the appropriate clients to target for loan issuance can be easily identified by evaluating their plausibility of loan default. The model implies that a bank should assess a creditor's other attributes, which play a critical role in credit decisions and forecasting loan defaulters, in addition to giving loans to wealthy borrowers.*

Keywords: Logistic regression, Loan prediction, Data analysis, Machine learning models.

I. INTRODUCTION

Banks have many products to sell in our banking system, but their main source of income is their credit lines. As a result, they are likely to profit from the interest on the loans they make. Loans, or whether customers repay or default on their loans, affect a bank's profit or loss. The bank can minimize its Non-Performing Assets by forecasting loan defaulters. Because precise predictions are crucial for maximising earnings, it's essential to look at the different methodologies and compare them.

A logistic regression model is a critical approach in predictive analytics for analysing the problem of predicting loan defaulters. Kaggle data is taken in order to investigate and predict. Logistic Regression models were used to calculate the various performance measures. Model is significantly better because it includes variables (personal attributes of customers include graduation, dependents, credit score, credit amount, credit period, and so on.) other than checking account information (which indicates a customer's wealth) that should be considered when correctly calculating the probability of loan default. As a result, by evaluating the likelihood of default on a loan, the right customers to target for loan granting can be easily identified using a logistic regression approach. The model predicts that a bank should not solely approve loans to wealthy consumers rather should also consider a customer's other characteristics, which play an important role in credit decisions and predicting loan defaulters.

As the demand for products and services rises, so does the amount of capital credit given, and people are more eager to take credit than ever before. As a result, computer software has replaced the human interface as more people from all over the world (Urban, Rural, and semi-urban) push for a high demand for credit.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 9, Issue 1, September 2021
Impact Factor: 5.731

A Machine Learning software algorithm has been developed in order to construct a robust and efficient software algorithm that classifies individuals based on 13 characteristics (Gender, Education, Number of Dependents, Marital Status, Employment, Credit Score, Loan Amount, and others) whether they would be eligible for a loan or not.

Although this is the first line of command, it will undoubtedly lower the workload of all other bank employees because the process will be automated to identify client segments and those who are qualified for a loan amount, allowing them to target those clients individually. And this will indicate whether or not the loan applicant meets the eligibility criteria for loan approval based on those 13 elements. To provide a convenient, prompt, and accurate method of selecting deserving applicants for loan eligibility. To determine the model's accuracy, calculate the accuracy score, which is the level of precision displayed by the model when forecasting the applicant's loan eligibility. There are many Machine Learning models that can also be used for the prediction of the loan eligibility of an applicant. Some of the models have been discussed below:

1.1 Support Vector Machine (SVM)

Support vector machines are learning models that employ an association learning technique to examine features and identify pattern information, which is then used to classify applications. SVM can reliably translate their inputs into high-dimensional feature spaces using the kernel method, resulting in a productive-regression. A support vector machine (SVM) is a supervised machine learning algorithm that includes classification techniques to solve two-group classification problems. SVM models can categorise new ‘text’ after being given sets of labelled training data for each category. They have two key advantages over newer algorithms like neural networks: greater speed and better performance with a limited number of samples (in the thousands). This makes the approach particularly well suited to text classification issues, where it's common to only have access to a few thousand tagged samples.

1.2 Decision Trees

All attributes or features must be discretized in order for the decision tree's basic algorithm to work. The most information gain of features is used to determine feature selection. IF-THEN rules can be used to represent the knowledge shown in a decision tree. Decision Tree Analysis is a general-purpose predictive modelling tool with applications in a variety of fields. In general, decision trees are built using an algorithm that determines multiple ways to segment a data set based on certain conditions. It is one of the most popular and practical supervised learning algorithms. Decision Trees are a supervised non-parametric learning method that may be utilised for both classification and regression applications. The goal is to learn simple decision rules from data attributes to develop a model that predicts the value of a target variable.

1.3 Random Forest (RF)

Random forest is a group learning system for characterization (and relapse) that works by building a large number of Decision trees over time and generating the class that is the mode of the classes generated by individual trees.. Random forest is a supervised learning technique that can be used to classify and predict data. However, it is mostly employed to solve categorization issues. As we all know, a forest is made up of trees, and more trees equal a healthier forest. The random forest technique, similarly, builds decision trees from data samples, extracts predictions from each, and then votes on the best alternative. It's an ensemble technique that's better than using a single decision tree because it averages the outcomes to avoid over fitting.

1.4 Linear Models (LM)

Although the Linear Model is quantitatively indistinguishable from other regression analyses, it has limitations in terms of its applicability for a variety of qualitative and quantitative variables.

1.5 Logistic Regression (LR)

Logistic regression is a technique for describing data and explaining the relationship between one or more independent factors and one or more dependent binary variables. Logistic regression, like all other regression analyses,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 9, Issue 1, September 2021
Impact Factor: 5.731

is a predictive technique that is employed when the dependent variable is categorical. The logistic regression statistical model is a prominent statistical model for binary classification, or predictions of the kind this or that, yes or no, A or B, and so on. Although logistic regression can be used for multi-class classification, we shall concentrate on its most basic application in this paper. It's one of the most common machine learning methods for binary classifications, converting the input to 0 or 1. When the dependent variable has a binary solution, logistic regression is the best regression methodology to use. Logistic Regression, like all other forms of regression systems, is a type of predictive regression system. The link between one dependent binary variable and one or more independent variables is evaluated using logistic regression.

1.6 XGBoost Classifier

XGBoost is a machine learning method that has recently dominated Kaggle tournaments for structured or tabular data. XGBoost is a high-speed and high-performance implementation of gradient boosted decision trees. There are few frills in the library because it is laser-focused on computing speed and model performance. It does, however, include a lot of advanced functions. The algorithm's implementation was designed to maximise computation time and memory resources. To train the model, one of the design goals was to make the most of available resources.

1.7 K-Nearest Neighbors (KNN)

The KNN algorithm is a supervised machine learning method that may be applied to both classification and regression prediction problems. In industry, nonetheless, it is mostly used to solve classification and prediction problems. The KNN method predicts the values of new data points using 'feature similarity,' this implies that a value will be assigned to the new data point based on how closely it resembles the points in the training set. It's a versatile method because it can be used for both classification and regression. KNN can be used in the banking system to forecast whether or not a person is eligible for a loan, as well as to determine a person's credit rating by comparing them to others who share similar characteristics.

1.8 Naive Bayes

Naive Bayes is a machine learning model that is used for huge amounts of data. It is recommended that you utilise Naive Bayes if you are working with data that has millions of records. When it comes to NLP tasks like sentimental analysis, it performs admirably. It's a simple and quick categorization algorithm. A Naive Bayes classifier is a machine learning model that separates objects based on specific variables' properties. It's a classification algorithm based on the Bayes theorem. For each class, such as the likelihood of data points linked with a given class, membership probabilities are predicted.

II. LITERATURE SURVEY

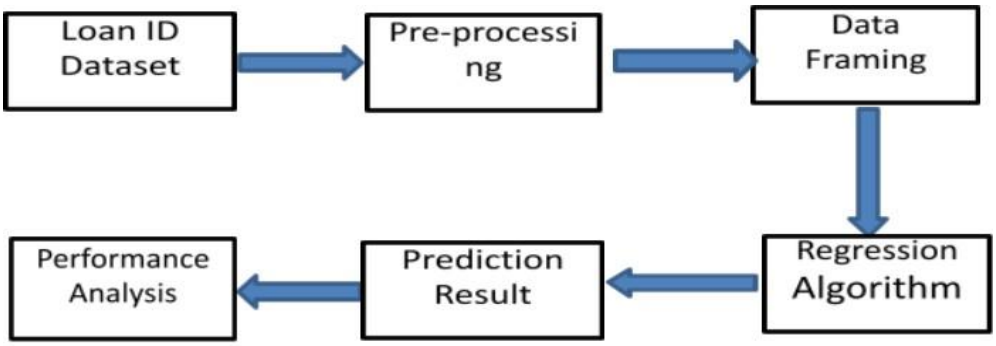
Sheikh, M. A., Goel, A. K., and Kumar, T. [1] used data from previous customers of various banks who had loans approved based on a set of criteria. To get accurate results, the machine learning model is trained on that record. The study's main purpose is to forecast the loan's safety. The logistic regression algorithm is used to predict loan safety. To avoid missing values in the data set, the data is first cleaned. The model was trained using a data set of 1500 cases with 10 numerical and 8 categorical attributes. Various parameters such as CIBIL Score (Credit History), Business Value, Customer Assets, and soon have been taken into account when crediting a loan to a customer. Vaidya [2] talks about logistic regression and how to represent it mathematically. His study employs logistic regression as a machine learning technique to actualize the predictive and probabilistic methods to a particular problem of loan approval prediction. This study employs logistic regression to determine if a loan for a set of records belonging to an applicant will be authorised. It also covers some of the Machine Learning mode's other real-world uses.

Zhang, H., Li, Z., Shahriar, H., Tao, L., Bhattacharya, P., and Qian, Y. [3] presented a logistic regression analysis using Python on imbalanced datasets and determined different classification thresholds based on the data proportion of imbalanced datasets. The research of Zou, X., Hu, Y., Tian, Z., and Shen, K. [5] focused on the logistic mathematical

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 9, Issue 1, September 2021
Impact Factor: 5.731

model, the definition of the error function, the gradient descent method for calculating the regression coefficient, and the Sigmoid function improvement. Therefore, the number of repetitions has been reduced, the classification impact has been improved, and the accuracy has remained nearly unchanged. Kumar Arun, Garg Ishan, and Kaur Sanmeet[6] have demonstrated how to reduce the risk factor when picking a safe individual in order to save time and money for the bank. This is performed by mining Big Data of previous records of persons to whom the loan was previously provided, and the machine was taught to get the best accurate result using a machine learning model based on these records/ experiences.

III. METHODOLOGY



Block Schematic of Loan Prediction

Figure 1: Block diagram

When an algorithm receives data as input, it produces binary output, which is either 0 or 1. If the result is 1, the number 1 will be displayed, indicating that the loan has been accepted. If the output is 0, the number '0' will be displayed, indicating that the loan has been denied. The prediction process includes phases such as data cleaning and processing, imputation of missing values, experimental analysis of a data set, model creation, and testing on test data. Various input variables were employed to obtain the output in order to implement this.

IV. EXPERIMENTAL ANALYSIS

Model used	Accuracy
Logistic Regression	0.785
Decision Tree Classifier	0.662
K Neighbors Classifier	0.619
Naive Bayes	0.779
Random Forest Classifier	0.773
Support Vector Machine	0.650
XGBoost Classifier	0.773

It is observed that Logistic Regression gives better accuracy for loan availability prediction. The reason behind this is that Logistic regression is also known as logit regression or logistic model. It accepts independent features and produces categorical results.

By fitting the features in the logistic curve to the logistic regression model, the probability of occurrence of a categorical output may also be determined. The general logit curve is seen in the diagram below.

The Logistic Regression model can be replaced with the simpler Linear Regression model when the output variable is believed to be continuous. A separate model must be employed to account for the difference when the output variable is not continuous or dichotomous. Following that, numerous models were created to account for the dichotomous nature of the outcome variable. Because of its mathematical clarity and versatility, the Logistic Regression model was chosen above the other models.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 9, Issue 1, September 2021
Impact Factor: 5.731

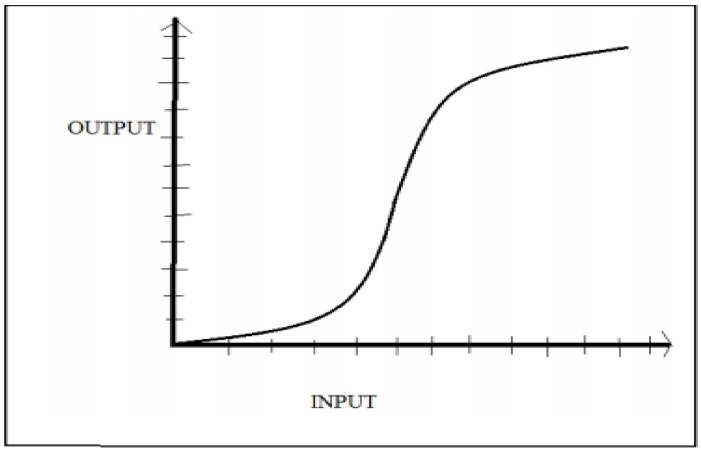


Figure 2: General Logit Curve

It's a prediction analysis. It's used to explain and describe the relationship between a single binary variable and one or more independent variables. Furthermore, the sigmoid function is taken into account in the logistic regression because the outcome is binary[4]. In this model, there can be one or more predictors. With this model, we can describe the target variable's natural log probability in the linear form of the feature variables used as input. This can be expressed mathematically as:

$$\text{logit(y)} = \ln(\frac{p}{1-p}) = \alpha+\beta x$$

The parameters of the logistic regression model are and ß, including p represents the probability of the desired category output and x represents the input feature. We can easily find the probability of the desired result by taking antilog on both sides of (1). The following is a mathematical representation:

$$P = \frac{1}{1+e^{-(\alpha+\beta x)}}$$

If more than one parameter is utilised as a feature and must be used for prediction, The natural log of probability for the desired variable is represented mathematically as follows:

$$\text{logit(y)} = \ln(\frac{p}{1-p}) = \alpha+\beta_1 x_1+\beta_2 x_2+....\beta_l x_l$$

$\alpha, \beta_1\beta_1 \quad \ldots \beta_l \beta_l$ are the logistic regression parameters, and x1...xl are the characteristics used to fit the model. Using antilog on both sides of the equation yields a result that is similar to but more extended than the second equation, which is given by

$$P = \frac{1}{1+e^{-(\alpha+\beta_1 x_1+\cdots.+\beta_l x_l)}}$$

In sectors where a relationship between features must be constructed and a dichotomous output must be achieved, logistic regression is increasingly widely used [3].

IV. Conclusion And Future Scope

In our model prediction of whether the loan would be accepted or not, we achieved the highest accuracy from the logistic regression model. On the dataset, the best case accuracy attained is 0.785. Our model was able to forecast whether the applicants in the dataset would be eligible for the loan when the project was completed. It was also able to anticipate the loan eligibility of a specific applicant by pointing out his row number. Applicants with a high income and smaller loan requests are more likely to be approved, which makes sense because they are more likely to payback their debts. Gender and marital status, for example, do not appear to be considered. The loan credibility prediction system can assist companies in making the best judgement on whether to approve or deny a customer's loan request. This will undoubtedly assist the banking industry in establishing more effective distribution routes. It is necessary to create and test new strategies that outperform the performance of common data mining models for the domain. As a result, in the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 9, Issue 1, September 2021
Impact Factor: 5.731

near future, the so-called algorithm might be made more reliable, efficient, and robust. This prediction module may be integrated with the automated processing system module in the near future. The system is currently trained on an existing training dataset, but algorithms can be implemented in the future to allow additional testing data to be included in the training dataset.

REFERENCES

[1]. Sheikh MA, Goel AK, Kumar T. An Approach for Prediction of Loan Approval using Machine Learning Algorithm. In2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) 2020 Jul 2 (pp.490-494).

[2]. Vaidya A. Predictive and probabilistic approach using logistic regression: application to prediction of loan approval. In2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) 2017 Jul 3 (pp.1-6).

[3]. TejaswiniJ, Kavya TM, Ramya RD, Triveni PS, Maddumala VR. Accurate Loan Approval Prediction Based On Machine Learning Approach. Journal of Engineering Science. 2020Apr;11(4):523-32.

[4]. Zhang H, Li Z, Shahriar H, Tao L, Bhattacharya P, Qian Y. Improving prediction accuracy for logistic regression on imbalanced datasets. In2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC) 2019 Jul 15 (Vol. 1, pp.918-919).

[5]. Zou X, Hu Y, Tian Z, Shen K. Logistic Regression Model Optimization and Case Analysis. In2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT) 2019 Oct 19 (pp.135-139).

[6]. Arun K, Ishan G, Sanmeet K. Loan approval prediction based on machine learning approach. IOSRJ. Comput. Eng. 2016;18(3):18-21.

[7]. Dutta, P., A Study On Machine Learning Algorithm For Enhancement Of Loan Prediction. International Research Journal of Modernization in Engineering Technology and Science,2021

[8]. Goyal, A. and Kaur, R., 2016. A survey on ensemble model for loan prediction. International Journal of Engineering Trends and Applications (IJETA), 3(1),pp.32-37.

[9]. Ruzgar, B. and Ruzgar, N.S., 2008. Rough sets and logistic regression analysis for loan payment. International journal of mathematical models and methods in applied sciences, 2(1),pp.65-73.

[10]. TVS, J., 2021. Predicting the Loan Status using Logistic Regression and Binary Tree. International Conference on IoT based Control Networks and Intelligent Systems (ICICNIS2020)

[11]. Dutta, P., A Study On Machine Learning Algorithm For Enhancement Of Loan Prediction. International Research Journal of Modernization in Engineering Technology and Science

[12]. Kwofie, C., Owusu-Ansah, C. and Boadi, C., 2015. Predicting the probability of loan-default: An application of binary logistic regression. Research Journal of Mathematics and Statistics, 7(4),pp.46-52.

[13]. Agbemava, E., Nyarko, I.K., Adade, T.C. and Bediako, A.K., 2016. Logistic regression analysis of predictors of loan defaults by customers of non-traditional banks in Ghana. European Scientific Journal,12(1).

[14]. DM, O. and Muraya, M.M., 2018. Comparison of Accuracy of Support Vector Machine Model and Logistic Regression Model in Predicting Individual Loan Defaults. American Journal of Applied Mathematics and Statistics, 6(6),pp.266-271.

[15]. Rath, G.B., Das, D. and Acharya, B., 2021. Modern Approach for Loan Sanctioning in Banks Using Machine Learning. In Advances in Machine Learning and Computational Intelligence (pp. 179-188). Springer, Singapore.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

MACHINE LEARNING TECHNIQUES FOR RECOGNIZING THE LOAN ELIGIBILITY

Mr. Abhiroop Sarkar*1

*1Bachelors Of Technology, Artificial Intelligence, G H Raisoni College Of Engineering, Maharashtra, India.

ABSTRACT

Loan can be considered to be a debt incurred by a person or an organization. Loans are usually lend by any single candidate/organization to another such party. The person who borrows the money has to agree to certain conditions like interest, extra charges etc. This study aims the prediction of whether a person is approved for being sanctioned a loan or not. There were many parameters like marital status, credit-history, gender etc. that has been considered for processing and analysis of Loan eligibility. Manually, analysis of Loan prediction become time consuming and costly, so this study has been performed to find the best algorithm which can automate the process to facilitate the Banker staff as well as customer to receive the eligibility analysis on immediate basis. The dataset has been splitted into training set and testing set where train used for training the algorithms upon which the test data has been used to make the predictions over the recognized entity.

Keywords: Loan-Eligibility, Machine Learning, Random Forest, Logistic Regression.

I. INTRODUCTION

A bank provides a customer with many services like safety of their money, interest options, quick withdrawals and other such benefits. The bank has various sources of income to provide the afore mentioned functionalities [1], still the main source of income stays on their credit lines. So, the interest gathered on the loan affects their profit the most. Hence whether a customer will repay or not the loan is important for the bank. Hence the loan is only given to the customers who are eligible to repay [2]. A loan is led by an organizations or other entities to people or other organizations. The borrower acts upon a debt for which he or she has to take authority to pay the interest until the loan is completely returned along with the original amount borrowed. Sanctioning a loan is one of the significant functions of a banking sector. Banks apply interests on loans which are then sanctioned to the customers.

Predicting loan eligibility helps both the bank employees as well as the customers. Here the paper is trying to provide quick and easy way to choose the deserving candidates. The system can calculate on its own the weight of each variable, which is a part of the loan process and test it. This checking will be done privately hence there will be no stakeholders who can alter the results so the applicable candidates will have higher priority to be sanctioned a loan [3] [4]. Machine learning algorithms enable the construction of a new model using previously unknown historical data that can be used to train the model to make better predictions not only for credit risks, but also for other risks such as early payment opportunities leading to loss of income from interest, existing withdrawal risks etc. [5].

A Prediction Model works on or operate on data analysis, statistics and probability to predict an outcome. Every model works on few variables which are likely to come in handy for future results. A statistical model is created based on the data collected from various resources. We can use simple machine learning algorithms or even a complex software for our project. If more data is used then the model becomes more better and so the errors are reduced meaning then the model will be able to predict with higher accuracy and even take less time [6]. In this paper we are considering logistic regression, random forest and decision tree machine learning algorithms for comparison. We will split the dataset into train and test classes and then predict the model using the three algorithms and find the best suited algorithm from among them [7].

In machine learning we split our complete dataset into training and testing dataset. We have used the splitting method like; class train_test_split() for achieving better result. There are always some issues that are faced with the random state parameter here we would get different accuracy for different random state thus not giving the perfect accuracy for our model, so we use stratified k fold cross validation for stratified sampling [8][9].

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

II. DATASET

Data collection is the process of collecting and measuring data in relation to targeted changes in an established system, which enables one to answer relevant questions and evaluate results. The purpose of all data collection is to obtain quality evidence that leads to analysis and constructs concrete and misleading answers to the questions presented [10]. The dataset has been divided into two categories the train dataset and the test dataset. The train dataset consists of six hundred and fourteen (614) rows and twenty-two (22) columns while the test dataset consists of three hundred and sixty-seven (367) rows and twenty (20) columns.

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y

Figure 1: (Training data entries)

```
[7] test=pd.read_csv('loan-test.csv')
    test.head()
```

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
0	LP001015	Male	Yes	0	Graduate	No	5720	0	110.0	360.0	1.0	Urban
1	LP001022	Male	Yes	1	Graduate	No	3076	1500	126.0	360.0	1.0	Urban
2	LP001031	Male	Yes	2	Graduate	No	5000	1800	208.0	360.0	1.0	Urban
3	LP001035	Male	Yes	2	Graduate	No	2340	2546	100.0	360.0	NaN	Urban
4	LP001051	Male	No	0	Not Graduate	No	3276	0	78.0	360.0	1.0	Urban

Figure 2: (Testing data entries)

The dataset consists of the parameters: gender, marital status, education, income credit history etc. From the data set we can infer that the applications from the 'male' gender is more than the counterpart and also that most of the applicants are married.

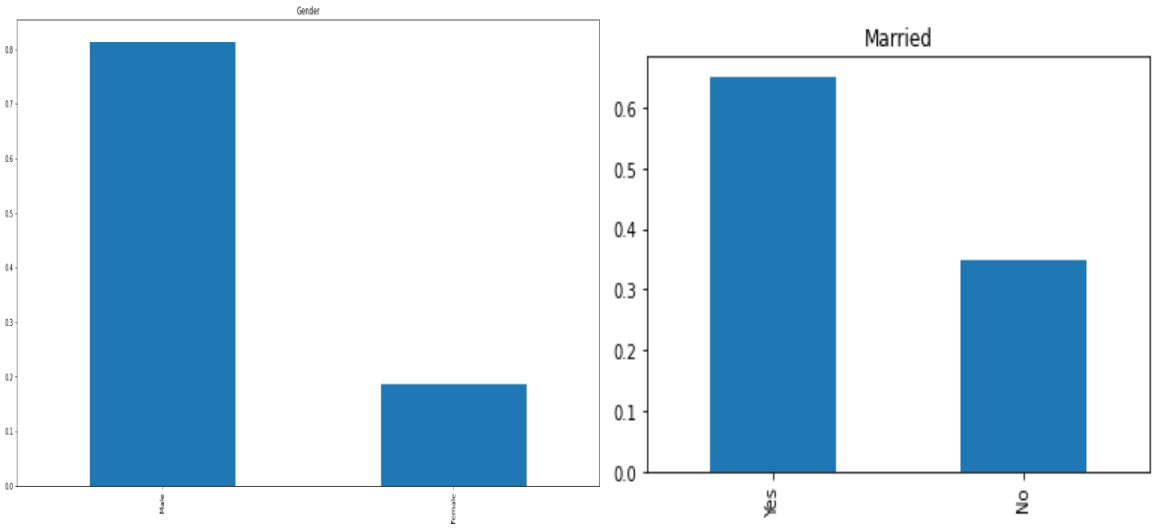


Figure 3: Gender & Marital status

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

The heat map for the following dataset is here it will show which variables are more related for the applicant to receive the loan and thus only those factors would be considered when the final model is being created.

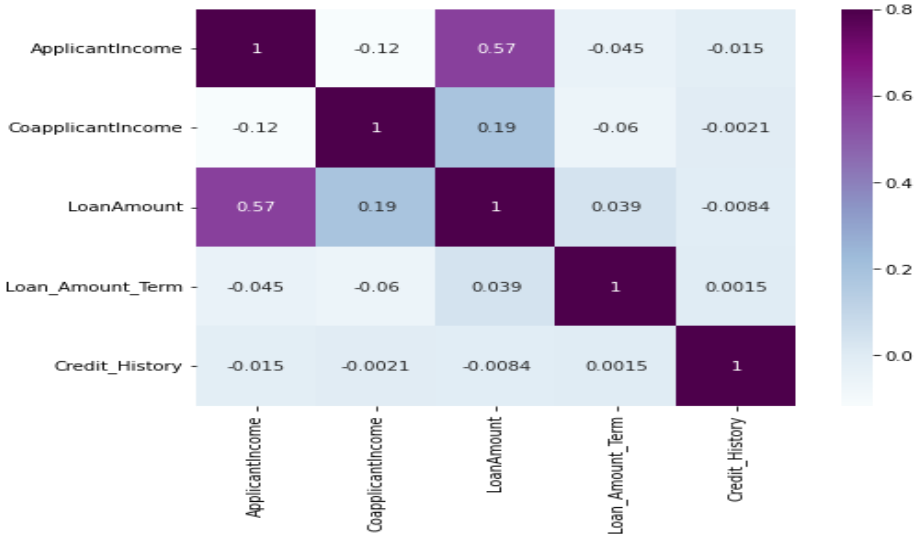


Figure 4: Heatmap between featured metrics

III. METHODOLOGY

For the model training we will compare the mean accuracy score of three machine learning algorithms: logistic regression, decision tree, random forest.

3.1. Logistic regression

It is a widely used Machine Learning algorithms, which belongs to the category of supervised learning. It is used to predict the dependent value based on a set of independent variables. Its primary function is the prediction of the dependent value. This value is usually categorical so the output has been a categorical or a discrete value. It can be either Yes or No, 0 or 1 etc., instead of giving the value as numeric form i.e., either 0 or 1, it gives a probabilistic value that lie between 0 and 1[11]. In this algorithm, instead of making a regression line, we make an 'S' shaped logistic function, which predicts two maximum possible values (0 or 1). The curve from the logistic function shows the likelihood of something like whether the cells are infectious or not, a cat is fat or not based on various different features [12][13]. Logistic Regression is quite an important machine learning technique as it has the power to give probabilities and classify new data from continuous and discrete datasets.

The equation for the straight line can be shown as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

As we know that in logistic regression y can be between 0 and 1 only, so we will divide the above equation by (1-y):

$$\frac{y}{1-y}\text{; 0 for y= 0, and infinity for y=1}$$

But the range has to be between - [infinity] to + [infinity], taking logarithm of the equation we will get:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

This can be considered to be the final mathematical representation for logistics regression. Logistic Regression can be broadly divided into three categories:

Binomial: In binomial logistic regression, there are only two possible types of the dependent variables, i.e., it is either 0 or 1.

Multinomial: In multinomial logistic regression, there are 3 or more possible unordered types of the dependent variable, such as 'horses', 'zebras', or 'dogs'.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

Ordinal: In ordinal logistic regression, there are 3 or more possible ordered types of dependent variables, such as 'low', 'medium' or 'high'.

3.2. Decision tree

Decision tree is one of the supervised learning techniques that is used for classification as well as regression problems though it is preferred to solve classification problems. It is a classifier, where the internal nodes show the features of a dataset, branches are used for decision rules and each leaf node represents outcome. It is a tree structured classifier. In a Decision tree, there are two nodes known as the decision node and the leaf node [14]. Decision nodes are used in making a decision and have multiple branches, on the other hand leaf nodes are the output based on the decision and do not contain any branches. The decisions depends upon the features present in the given dataset.

It is a graphical representation for getting all the possible solutions to a problem depending upon a few valid conditions. It is called a decision tree as, like in a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. The CART algorithm which stands for Classification and Regression Tree algorithm is used for building the tree. The attribute selection is one of the issues that arises when a decision tree is being implemented. Attribute selection measures is a technique used to solve such problems. Using this technique, we can easily select the best attribute for the tree nodes [15] [16].

3.3. Information Gain:

Information gain is the measurement of changes in entropy after the dataset is segmented. It is used to find out how much information a feature can give us about a particular class. We use the value of information gain, to split the node and create the decision tree. A decision tree algorithm tries to reach the maximum possible value of information gain, and a node having the highest information gain is split first.

It can be calculated using the below formula:

Information Gain= Entropy(S) - [(Weighted Avg) *Entropy (each feature)]

Entropy: Entropy is used to measure the impurity in an attribute. It specifies the randomness present in the data. Entropy can be calculated as:

Entropy(S) = -P (yes) log2 P (yes) - P (no) log2 P (no)

Where,

S= Total number of samples

P (yes) = probability of yes

P (no) = probability of no

3.4. Random forest

Random Forest is a machine learning algorithm that is a part of the supervised learning technique. It can be used for classification as well as regression problems. It is derived on the basis of ensemble learning, which is the method of combining multiple classifiers to solve a complex problem and to increase the performance of the model. Random forest contains many individual decision trees that work as an ensemble. Each tree in the forest gives out a prediction and the class that has the maximum number of outcomes becomes the model's prediction [17]. A large number of relatively uncorrelated models operating as a group will give better performance than any of the individual constituent models. There is a low correlation between the models. Uncorrelated models can produce ensemble predictions which are more accurate than any of the individual predictions. The prerequisites for random forest for performing with better results are:

- There has to be actual signal in the features so that the models built using those features do better than random guessing.
- The predictions (and errors) generated by the individual trees should have low correlations with each other.
- The greater the number of trees in the forest, the higher is the accuracy and thus it prevents the situation of overfitting [18].
- Random forest is considered to be a better algorithm because it takes less training time than other algorithms and it increases the accuracy even for large datasets [19].

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

IV. RESULT ANALYSIS

We have tried to predict whether an applicant is applicable for a loan or not using three machine learning algorithms below are the observations for all of them.

4.1. Logistic regression

Logistic Regression is a classification algorithm used to assign observation to a discrete set of class. It is one among the machine learning algorithm which is used for classification problem, it is a predictive analysis algorithm based on the concept of probability.

```
i=1
mean = 0
kf = StratifiedKFold(n_splits=5,random_state=1,shuffle=True)
for train_index,test_index in kf.split(X,y):
 print ('\n{} of kfold {} '.format(i,kf.n_splits))
 xtr,xvl = X.loc[train_index],X.loc[test_index]
 ytr,yvl = y[train_index],y[test_index]
 model = LogisticRegression(random_state=1)
 model.fit(xtr,ytr)
 pred_test=model.predict(xvl)
 score=accuracy_score(yvl,pred_test)
 mean += score
 print ('accuracy_score',score)
 i+=1
 pred_test = model.predict(test)
 pred = model.predict_proba(xvl)[:,1]
print ('\n Mean Validation Accuracy',mean/(i-1))
```

```
1 of kfold 5
accuracy_score 0.8048780487804879

2 of kfold 5
accuracy_score 0.8373983739837398

3 of kfold 5
accuracy_score 0.7967479674796748

4 of kfold 5
accuracy_score 0.7967479674796748

5 of kfold 5
accuracy_score 0.8032786885245902

 Mean Validation Accuracy 0.8078102092496335
```

Figure 5: Algo. for Logistic Regression

This shows the mean validation accuracy for logistic regression algorithm. We can see that it has come around 80.78%

4.2. Decision tree

In Decision Analysis, Decision Tree can be used to visually and explicitly represent decision and decision making. The aim of using this algorithm was to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules inferred from training prior data.

```
from sklearn import tree
i=1
mean = 0
kf = StratifiedKFold(n_splits=5,random_state=1,shuffle=True)
for train_index,test_index in kf.split(X,y):
    print ('\n{} of kfold {} '.format(i,kf.n_splits))
    xtr,xvl = X.loc[train_index],X.loc[test_index]
    ytr,yvl = y[train_index],y[test_index]
    model = tree.DecisionTreeClassifier(random_state=1)
    model.fit(xtr,ytr)
    pred_test=model.predict(xvl)
    score=accuracy_score(yvl,pred_test)
    mean += score
    print ('accuracy_score',score)
    i+=1
    pred_test = model.predict(test)
    pred = model.predict_proba(xvl)[:,1]
print ('\n Mean Validation Accuracy',mean/(i-1))
```

```
1 of kfold 5
accuracy_score 0.7073170731707317

2 of kfold 5
accuracy_score 0.6991869918699187

3 of kfold 5
accuracy_score 0.7154471544715447

4 of kfold 5
accuracy_score 0.7235772357723578

5 of kfold 5
accuracy_score 0.680327868852459

 Mean Validation Accuracy 0.7051712648274024
```

Figure 6: Algo. for Decision Tree

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

This shows the mean validation accuracy for decision tree algorithm. We can see that it has come around 70.51%

4.3. Random forest

Random Forest is an ensemble machine learning algorithm that is used for classification and regression problem. Random Forest applies the technique of bagging (bootstrap aggregating) to decision tree learner. The beginning of the random forest algorithm starts with randomly selected "k" features to find the root node by using the best split approach.

```
from sklearn.ensemble import RandomForestClassifier
i=1
mean = 0
kf = StratifiedKFold(n_splits=5,random_state=1,shuffle=True)
for train_index,test_index in kf.split(X,y):
 print ('\n{} of kfold {} '.format(i,kf.n_splits))
 xtr,xvl = X.loc[train_index],X.loc[test_index]
 ytr,yvl = y[train_index],y[test_index]
 model = RandomForestClassifier(random_state=1, max_depth=10)
 model.fit(xtr,ytr)
 pred_test=model.predict(xvl)
 score=accuracy_score(yvl,pred_test)
 mean += score
 print ('accuracy_score',score)
 i+=1
 pred_test = model.predict(test)
 pred = model.predict_proba(xvl)[:,1]
print ('\n Mean Validation Accuracy',mean/(i-1))
```

```
1 of kfold 5
accuracy_score 0.8048780487804879

2 of kfold 5
accuracy_score 0.8373983739837398

3 of kfold 5
accuracy_score 0.7886178861788617

4 of kfold 5
accuracy_score 0.8130081300813008

5 of kfold 5
accuracy_score 0.7459016393442623

 Mean Validation Accuracy 0.7979608156737305
```

Figure 7: Algo. for Random Forest

This shows the mean validation accuracy for random forest algorithm. We can see that it has come around 79.79%

V. CONCLUSION

We can easily conclude that logistic regression can be considered to be the best among the three machine learning algorithms with an accuracy of 80.78%, closely followed by random forest at 79.79% and finally by decision tree with 70.51%. This paper gives a general idea that we can prefer logistic regression for loan eligibility. Based on the results ideas for including other machine learning algorithms like XGBoost and others can be compared, research has already in action for these algorithms. It is inclusive of all the parameters needed to evaluate the creditworthiness of a client. The model is trained to produce results with satisfactory accuracy, after which it produces accurate results as to whether a borrower should be lent money or not without any tedious manual work.

VI. CONFLICT OF INTEREST

The authors of the article entitled "Machine learning Techniques for Recognizing the Loan Eligibility" declare that there are no conflicts of interest regarding the research manuscript. None of the either Human Being or Animals are affected throughout the research processing.

VII. REFRENCES

[1] Deepak Ishwar Gouda, Ashok Kumar A, Anil Manjunatha Madivala, Dilip Kumar R, Dr.Ravikumar, "LOAN APPROVAL PREDICTION BASED ON MACHINE LEARNING", International Research Journal of Engineering & Technology, Volume-8 Issue-11, January 2021.

[2] Sharayu Dosalwar ,Dr. Vishwanath Karad ,Ketki Kinkar,Rahul Sannat,Nitin Pise, "Analysis of Loan Availability using Machine Learning Techniques", September 2021, DOI: http://dx.doi.org/10.48175/IJARSCT-1895

[3] Kumar, R., Jain, V., Sharma, P. S., Awasthi, S., & Jha, G. (2019). Prediction of Loan Approval using Machine Learning. International Journal of Advanced Science and Technology, 28(7), 455 - 460. Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/460

[4] Ramya S , Priyesh Shekhar Jha , Ilaa Raghupathi Vasishtha , Shashank H , Neha Zafar, "Monetary Loan Eligibility Prediction using Machine Learning", IJESC, Volume:11 Issue-7.

[5] Prateek Dutta, "A Study on Machine Learning Algorithm for Enhancement of Loan Prediction", "International Research Journal of Modernization in Engineering Technology and Science", Volume -3 issue-1, January 2021.

[6] AFRAH KHAN, EAKANSH BHADOLA, ABHISHEK KUMAR and NIDHI SINGH, "LOAN APPROVAL PREDICTION MODEL A COMPARATIVE ANALYSIS", Advances and Application in Mathematical Science, January,2021.

[7] A. Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017, pp. 1-6, doi: 10.1109/ICCCNT.2017.8203946.

[8] Efosa G. Adagbasa, Samuel A. Adelabu, Tom W. Okello, "Application of deep learning with stratified K-fold for vegetation species discrimination in a protected mountainous region using Sentinel-2 image", Geocarto International, 19 December,2019, DOI: https://doi.org/10.1080/10106049.2019.1704070

[9] A. Fernandez-Carrillo, D. de la Fuente, F. W. Rivas-Gonzalez, and A. Franco-Nieto "An automatic Sentinel-2 Forest types classification over the Roncal Valley, Navarre: Spain", Proc. SPIE 11156, Earth Resources and Environmental Remote Sensing/GIS Applications X, 111561N (3 October 2019); https://doi.org/10.1117/12.2533059

[10] Prateek Dutta, "A Study on Machine Learning Approach for Market Segmentation", International Journal of Scientific Research in Engineering and Management", Volume-5 Issue-7, July 2021.

[11] Michael P. LaValley, "Logistic Regression", Circulation, DOI: https://doi.org/10.1161/CIRCULATIONAHA.106.682658

[12] Wright, R. E. (1995). Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding multivariate statistics (pp. 217–244). American Psychological Association.

[13] Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting", The Journal of Educational Research, April 2010, DOI: https://doi.org/10.1080/00220670209598786

[14] Harsh Patel, Purvi Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms", International Journal of Computer Science and Engineering, October 2018, DOI:

[15] http://dx.doi.org/10.26438/ijcse/v6i10.7478

[16] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, May-June 1991,

[17] Doi: 10.1109/21.97458.

[18] Anthony J. Myles, Robert N. Feudale, Yang Liu, Nathaniel A. Woody, Steven D. Brown, "An introduction to decision tree modeling", Journal of Chemometrics, 2004, DOI: https://doi.org/10.1002/cem.873

[19] M. Pal, "Random Forest classifier for remote sensing classification", International Journal of Remote Sensing, 2007, DOI: https://doi.org/10.1080/01431160412331269698

[20] V.F.Rodriguez-Galiano, B.Ghimire, J.Rogan, M.Chica-Olmo, J.P.Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification", ISPRS Journal of Photogrammetry and Remote Sensing, January 2012,
DOI: https://doi.org/10.1016/j.isprsjprs.2011.11.002

[21] Devetyarov D., Nouretdinov I. (2010) Prediction with Confidence Based on a Random Forest Classifier. In: Papadopoulos H., Andreou A.S., Bramer M. (eds) Artificial Intelligence Applications and Innovations. AIAI 2010. IFIP Advances in Information and Communication Technology, vol 339. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-16239-8_8

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

Prediction for Loan Approval using Machine Learning Algorithm

Ashwini S. Kadam[1], Shraddha R. Nikam[2], Ankita A. Aher[3], Gayatri V. Shelke[4], Amar S. Chandgude [5]

[1-4]Student, Dept. Of Computer engg., SND College of Engineering and research center, Yeola, Maharashtra, India
[5]Professor, Dept. Of Computer engg., SND College of Engineering and research center,Yeola, Maharashtra, India

---

Abstract - *In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So they can earn from interest of those loans which they credits. A bank's profit or a loss depends to a large extent on loans i.e. whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non-performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison. A very important approach in predictive analytics is used to study the problem of predicting loan defaulters (i) Collection of Data, (ii) Data Cleaning and (iii) Performance Evaluation. Experimental tests found that the Naïve Bayes model has better performance than other models in terms of loan forecasting.*

Key Words: Big data, Machine Learning, SVM, Naïve Bayes, Prediction.

1. INTRODUCTION

Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this Paper is to provide quick, immediate and easy way to choose the deserving applicants. Dream housing Finance Company deals in all loans. They have presence across all urban, semi urban and rural areas. Customer first apply for loan after that company or bank validates the customer eligibility for loan. Company or bank wants to automate the loan eligibility process (real time) based on customer details provided while filling application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and other. This project has taken the data of previous customers of various banks to whom on a set of parameters loan were approved. So the machine learning model is trained on that record to get accurate results. Our main objective of this project is to predict the safety of loan. To predict loan safety, the SVM and Naïve bayes algorithm are used. First the data is cleaned so as to avoid the missing values in the data set.

1.1 MOTIVATION

Loan approval is a very important process for banking organizations. The system approved or reject the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. Using Machine learning we predict the loan approval.

2. LITERATURE SURVEY

1." Loan Approval Prediction based on Machine Learning Approach" Author- Kumar Arun, Garg Ishan, Kaur Sanmeet Year- 2018The main objective of this paper is to predict whether assigning the loan to particular person will be safe or not. This paper is divided into four sections (i)Data Collection (ii) Comparison of machine learning models on collected data (iii) Training of system on most promising model (iv) Testing

2."Exploring the Machine Learning Algorithm for Prediction the Loan Sanctioning Process" Author- E. Chandra Blessie, R. Rekha - Year- 2019 Extending credits to corporates and individuals for the smooth functioning of growing economies like India is inevitable. As increasing number of customers apply for loans in the banks and non- banking financial companies (NBFC), it is really challenging for banks and NBFCs with limited capital to device a standard resolution and safe procedure to lend money to its borrowers for their financial needs. Inaddition, in recent times NBFC inventories have suffered a significant downfall in terms of the stock price. It has contributed to a contagion that has also spread to other financial stocks, adversely affecting the benchmark in recent times.In this paper, an attempt is made to condense the risk involved in selecting the suitable person who could repay the loan on time thereby keeping the bank's nonperforming assets (NPA) on the hold. This is achieved by feeding the past records of the customer who acquired loans from the bank into a trained machine learning model which could yield an accurate result. The prime focus of the paper is to determine whether or not it will be safe to allocate the loan to a particular person. This paper has the following sections (i) Collection of Data, (ii) Data Cleaning and (iii) Performance Evaluation. Experimental tests found that the Naïve Bayes model has better performance Evaluation. Experimental tests found that the Naïve Bayes model has better performance than other models in terms of loan forecasting.

3. "Loan Prediction using machine learning model" Year-2019whether or not it will be safe to allocate the loan to a particular person. This paper has the following sections (i) Collection of Data, (ii) Data Cleaning and (iii) Performance Evaluation. Experimental tests found that the Naïve Bayes model has better performance than other models in terms of loan forecasting.With the enhancement in the banking sector lots of people are applying for bankloans but the bank has its limited assets which it has to grant to limited people only,so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this project we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result

The main objective of this project is to predict whether assigning the loan to particular person will be safe or not. This paper is divided into four sections (i)Data Collection (ii) Comparison of machine learning models on collected data (iii) Training of system on most promising model (iv) Testing. In this paper we are predict the loan data by using some machine learning algorithms they are classification, logic regression, Decision Tree and gradient boosting.

4."Loan Prediction using Decision Tree and Random Forest"Author- Kshitiz Gautam, Arun Pratap Singh, Keshav Tyagi, Mr. Suresh Kumar Year-2020. In India the number of people or organization applying for loan gets increasd every year. The bank have to put in a lot of work to analyse or predict whether the customer can pay back the loan amount or not (defaulter or non-defaulter) in the given time. The aim of this paper is to find the nature or background or credibility of client that is applying for the loan. We use exploratory data analysis technique to deal with problem of approving or rejecting the loan request or in short loan prediction. The main focus of this paper is to determine whether the loan given to a particular person or an organization shall be approved or not.

3. PROBLEM DEFINITION

Banks, Housing Finance Companies and some NBFC deal in various types of loans like housing loan, personal loan, business loan etc in all over the part of countries. These companies have existence in Rural, Semi Urban and Urban areas. After applying loan by customer these companies validates the eligibility of customers to get the loan or not. This project provides a solution to automate this process by employing machine learning algorithm. So the customer will fill an online loan application form. This form consist details like Sex, Marital Status, Qualification, Details of Dependents, Annual Income, Amount of Loan, Credit History of Applicant and others.

3. IV.PROPOSED MODEL

This system predict whether the loan is approve or reject . This System refers the following things or ways.

Data Collection

Data Pre-processing (Data Cleaning)

Model Selection

Model Evaluation

Classification

Result (output)

4. SYSTEM ARCHITECTURE

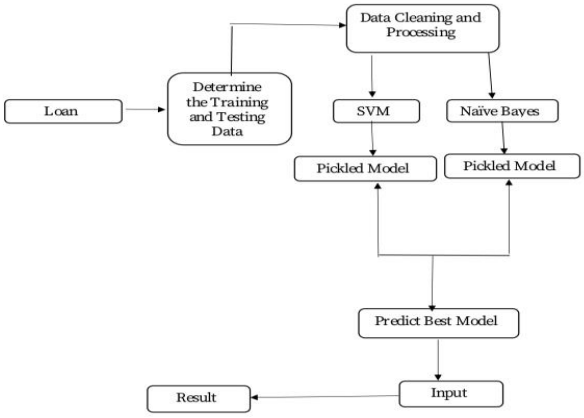


Fig -1: Loan Prediction Architecture

Implementation Details (Modules):

4.1. Loan Dataset : Loan Dataset is very useful in our system for prediction of more accurate result. Using the loan Dataset the system will automatically predict which costumer's loan it should approve and which to reject. System will accept loan application form as an input. Justified format of application form should be given as an input to get processed.

4.2. Determine the training and testing data: Typically , Here the system separate a dataset into a training set and testing set ,most of the data use for training ,and a smaller portions of data is use for testing. after a system has been processed by using the training set, it makes the prediction against the test set.

4.3. Data cleaning and processing: In Data cleaning the system detect and correct corrupt or inaccurate records from database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing , modifying or detecting the dirty or coarse data. In Data processing the system convert data from a given form to a much more usable and desired form i.e. make it more meaningful and informative.

4.4. Models used :

1) SVM:

In this approach, each data item is plotted in a n-dimensional space, where n represents the number of features with each feature represented in a corresponding co- ordinates. A hyper plane is determined to distinguish the classes (possibly two) based on their features.

2) Naïve Bayes (NB) Model :

The basis for NB model is Bayes Theorem (BT), where events are mutually exclusive similar to rolling a die. Moreover, the BT presumes that the input features also referred aspredictors are independent in nature. Similarly, NB also presumes that the input features are independent in nature. But, this is impossible in the realistic procedures.Since this assumption leads to naïve, this algorithm is termed as Naïve Bayes algorithm. Thus, NB is a probabilistic algorithm, where the conditional probability is determined regarding the input features. On the other hand, during the dependent input features scenario, conditional probability is calculated twice resulting in improper results. Hence, for better prediction results with respect to NB model, independent input features are selected and processed. dataset collected from Kaggle source. The feature in the dataset include

1. Aplication_Id
2. Gender
3. Marital Status
4. Number of dependents
5. Educational Profile
6. Employment Status
7. Applicant‘s Income
8. Co-Applicant‘s Income
9. Loan Amount
10. Credit History
11. Loan Status

4.5. Exploratory Data Analysis System

verify the documents and forward the details to loan evaluator for approval or rejection. System approve the loan if documents are cleared and reject the loan if documents are not cleared Report is delivered to the applicant according to their status.

5. PROPOSED ALGORITHM:

The following shows the pseudo code for the proposed loan prediction method

1. Load the data

2. Determine the training and testing data

3. Data cleaning and pre-processing.

a) Fill the missing values with mean values regarding numerical values.

b) Fill the missing values with mode values regarding categorical variables.c) Outlier treatment.

4. Apply the modelling for prediction

a) Removing the load identifier

b) Create the target variable (based on the requirement). In this approach, target variable is loan-status

c) Create a dummy variable for categorical variable (if required) and split the training and testing data for validation.

d) Apply the model: NB method, SVM method

5. Determine the accuracy followed by confusion Matrix.

5.1. SYSTEM FEATURES

- Data collection.
- Data cleaning and preprocessing
- Model selection
- Data verification
- Classification.
- Report deliver.

6. MATHEMATICAL MODEL

Consider any decision problem, where for given number of inputs, decision oriented solution is available so our project is NP complete but some cases like not proper input format provided or if dataset not trained proper it’s NP hard.

Let s be System :

S=I, P, O

S: is a System

I=I1, I2

P= DC, DP, DV, NBA, CL

O=RD

I1: Loan Dataset

I2: Trained Dataset.

DC: Data Cleaning

D DP: Data Processing

DV: Data Verification

NBA: Naïve Bayes Algorithm

CL: Classification

RD: Report Deliver Success

Condition : Proper features trained Dataset will give proper output

Failure Condition No Trained Dataset.

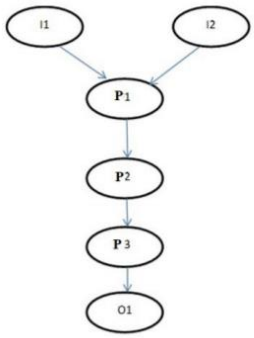


Fig -2: Mathematical model

7. USE CASE DIAGRAM

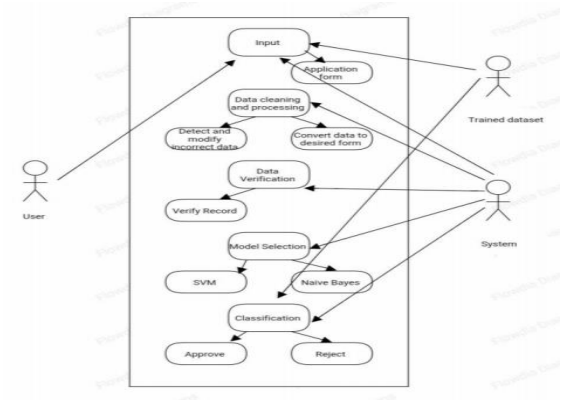


Fig -3: use case diagram

CONCLUSION

So here, it can be concluded with confidence that the Naïve Bayes model is extremely efficient and gives a better result when compared to other models. It works correctly and fulfills all requirements of bankers. This system properly and accurately calculate the result. It predicts the loan is approve or reject to loan applicant or customer very accuratly.

REFERENCES

1) Kumar Arun, Garg Ishan, Kaur Sanmeet, —Loan Approval Prediction based on Machine Learning Approach‖, IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 18, Issue 3, pp. 79-81, Ver. I (May-Jun. 2016).

2) Aboobyda Jafar Hamid and Tarig Mohammed Ahmed, —Developing Prediction Model of Loan Risk in Banks using Data Mining‖, Machine Learning and Applications: An International Journal (MLAIJ), Vol.3, No.1, pp. 1-9, March 2016.

3) S. Vimala, K.C. Sharmili, —Prediction of Loan Risk using NB and Support Vector Machine‖, International Conference on Advancements in Computing Technologies (ICACT 2018), vol. 4, no. 2, pp. 110-113, 2018.

4) Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma, Namburi Vimala Kumari, kVikash,"Loan Prediction by using Machine Learning Models", InternationalJournalofEngineering andTechniques.Volume 5 Issue 2, Mar-Apr 2019

5) Nikhil Madane, Siddharth Nanda,"Loan Prediction using Decision tree", Journal of the Gujrat Research History, Volume 21 Issue 14s, December 2019.

International Journal of Engineering and Management Research
e-ISSN: 2250-0758 | p-ISSN: 2394-6962
https://ijemr.vandanapublications.com

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.2

Prediction of Loan Approval in Banks using Machine Learning Approach

Viswanatha V[1], Ramachandra A.C[2], Vishwas K N[3] and Adithya G[4]
[1]Assistant Professor, Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bangalore, INDIA
[2]Professor, Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bangalore, INDIA
[3]Student, Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bangalore, INDIA
[4]Student, Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bangalore, INDIA

[3]Corresponding Author: viswas779@gmail.com

Received: 01-07-2023 Revised: 16-07-2023 Accepted: 30-07-2023

ABSTRACT

Due to significant technology advancements, people's needs have expanded. As a result, there have been more requests for loan approval in the banking sector. A few qualities, taken for consideration, when choosing a candidate for loan approval in order to, determine loan's status. Banks face a major challenge; when it, comes to assessing loan applications and lowering the risks associated with potential borrower defaults. Since they must thoroughly evaluate each borrower's eligibility for a loan, banks find this process to be particularly challenging. This research proposes combining machine learning (ML) models and ensemble learning approaches to find the probability of accepting individual loan requests. This tactic can increase the accuracy with which qualified candidates are selected from a pool of applicants. As a result, this method can be used to address the problems with loan approval processes outlined above. Both the loan applicants and the bank employees profit from the strategy's dramatic reduction in sanctioning time. Because of the banking industry's expansion, more people were applying to loans at banks. In order to predict the accuracy of loan approval status for applied person, we used four different algorithms namely Random Forest, Naive Bayes, Decision Tree, and KNN. By using these, we obtained better accuracy of 83.73% with Naïve Bayes algorithm as best one.

Keywords-- Safe Customers, Bank Loans, Trained Dataset, Random Forests, KNN, Decision Tree, Naive Bayes

I. INTRODUCTION

Many banks' primary line of business is loan distribution. Loans given to consumers account for the majority of a bank's revenue. Interest is charged by these banks on loans given to customers. Banks' handled. It merely has the values x and y as independent and dependent variables. Data primary goal is to invest their funds in dependable clients. Many banks have been processing loans so far following a backward process of vetting and verification. However, as of right now, no bank can guarantee whether the customer who is selected for a loan application is secure or not. So, in order, to avoid this circumstance, we implemented the Loan Prediction System Using Python, a system for the approval of bank loans. The Loan Prediction System is a piece of software that determines if or not the specific customer is qualified to receive a loan. This technique examines number of variables, including the customer's marital status, income, spending, and other elements. For wide numbers of trained data set clients, this method/technique is used. These elements are, taken to consideration when creating the necessary model. In order for obtaining the desired outcome, this model is applied for the test data set. The result will be presented as either yes or no. If the answer is yes, then the customer is capable of repaying the loan; if the answer is no, then the consumer is not capable of repaying the loan. We can grant loans to clients based on these criteria. Machine learning is the study of how the systems of computers are used and developed to learn and adapt without explicit instructions by analysing and inferring patterns in data using algorithms and statistical models. It is so important in the twenty-first century that it was used practically everywhere, from Such a function can’t be fitted with a straight line without incurring significant mistakes. Additionally, the datasets those with greater than two dimensions scientists developed polynomial regression, logistic regression, and even linear regression with having more variables to address these problems. As the accuracy sharply improved, more individuals grew interested in it and started working on it. The new era of data science began with the first use of the term "Big data" in 2005. Many ideas can now be fulfilled, including decision tree regression. commonplace things like a search engine and an email filter to more challenging issues like predicting consumer behaviour or our topic, predicting house

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.2

prices. Although the concept of regression, or the act of building a function to describe the dataset points, was not developed until around 1800, machine learning (ML) algorithms didn't appear until 1952. In accordance for evaluate, effectiveness of a function in fitting a large number of points of data, Legendre created and published "the method of least square" in 1805. The first effective cost function with a mathematical foundation is developed. Over the following century, mathematicians and scientists like Gauss and Markov would extend this concept and apply it to produce formulas. The regression was an extremely challenging process, though, as there were no computers (or even calculators) accessible at the time. Everything began to alter in the 1950s with the introduction of the machine learning idea. To execute linear regression, a unique type of calculator was developed. as implied by the name. Utilizing a linear function to make predictions based on supplied data points is known as linear regression. By reducing the cost function of linear regression (squared error), a best fit linear function can be discovered for practically any dataset. However, when it first debuted, it didn't appear to be that helpful. Numerous problems are still open. First, many datasets cannot be accurately represented by a straight line. For instance, a quadratic connection is one in which y gets highly high or low depending on a value of x, but extremely low depending on value of x. In most cases, loan prediction entails the lender reviewing the applicant's background data for the determination of whether the bank should approve the loan. The elements that, determine if a loan will be granted include aspects like credit history, loan amount, lifestyle, career, and assets. It is more probable that your loan will be approved if previous borrowers with criteria similar to yours have made on-time payments. This reliance on prior knowledge and comparisons with other applicants can be taken advantage of by machine learning (ML) algorithms, which can, then be used for create a data science issue to forecast the loan status of new applicant using the set of analogous criteria.

II. LITERATURE REVIEW

In their study, Rajiv Kumar and Vinod Jain constructed the logistic tree, decision tree, random forest algorithms using the Python computer language [1]. The decision tree (DT) technique was founded to be the most efficient after comparing the correction of three distinct; machine learning (ML) algorithms in terms of prediction accuracy. However, this can be fixed by correctly classifying the data and completing any gaps that were left out. Pidikiti Supriya and Myneedi Pavani claim in their study work [2] that they pre-processed the data to remove any anomalies in dataset. They have also created list of Correlating Characteristics that had, found for raise, probability of debt payback. The set of data was classified as training and testing operations using the 80:20 rule. The Python platform's complot and boxplot utilities are used to, find the correlation between the attributes. They haven't employed any other method to compare accuracy results, besides a decision tree. This can be prevented by training datasets using multiple techniques and assessing their efficacy.

In their research study, Kumar Arun and Garg Ishan studied six distinct machine learning (ML) techniques, having, support vector machines, and neural networks, random forests, decision trees, linear models, and Adaboost [3]. The four sections of this, study were as follows. Data gathering (i), model evaluation (ii), machine learning (ML) on the collected data (iii), system training (iv), and system testing using the most useful model (v) are the steps involved. The R programming language was employed in the creation of this system. It was challenging for others to comprehend and compare the results because they didn't visualize the data outcomes using graphs or other matrix representations, but this problem might be resolved by doing so. Authors from [4]. At first, the data was cleaned up. The next steps were exploratory data analysis and feature engineering. Graphs had been employed for visualization. For loan prediction, four models are used. Support Vector Machines, Decision Tree (DT) algorithm, Naive Bayes and the Logistic Regression, three four methods. They thoroughly considered the benefits and limitations and came to the confident conclusion that Naive Bayes(NB) model is quite capable of delivering results that are superior to those of other models.

The sets of data, according to the authors in [5], was acquired from the industry of banking. Weka can read the data set, because , it is in the ARFF (Attribute Relation File Format) format. To address an issue of accepting or declining loan requests as like as short-term loan prediction, they employed exploratory data testing. They conducted the exploratory data testing, to their study. Decision Tree(DT), and Random Forest(RF) are two machine learning categorization models thaose are utilised for prediction. They used the random forest method in their analysis.

III. DESIGN AND METHODOLOGY

Import the necessary libraries, such as scikit-learn, pandas, and numpy, to process data and create a prediction model.Fill a pandas DataFrame with the loan data.Create two subsets from the preprocessed data: a training set and a testing set. The predictive model will be trained using the training set, and its performance will be assessed using the testing set.Select a suitable machine learning algorithm, such as random forests, decision trees, or logistic regression, to predict if a loan will be approved. Create an instance of the selected model and adjust any required hyperparameters. Using the fit() function, adjust the model to the training set of data.

In order to produce predictions, the model will discover patterns and relationships in the training data.

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.2

Depending on its characteristics, the model will categorize each loan application as authorized or denied. Compare the testing set's actual loan approval labels to the expected loan approval labels, all are represented in the Fig.1

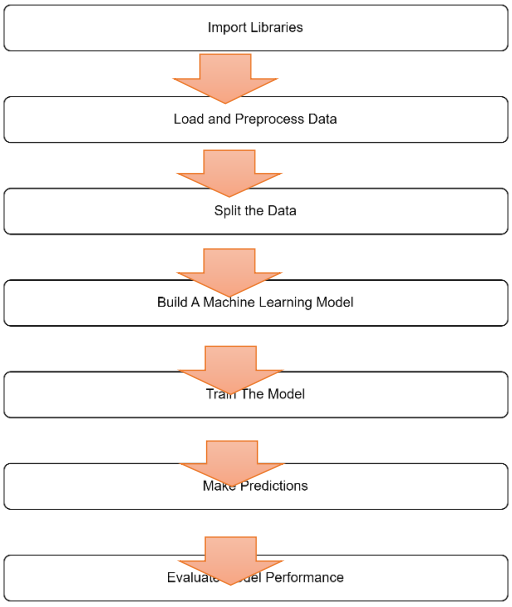


Figure 1: Flowchart of Loan Amount Prediction

A. Algorithms Used

a). Random Forest

Favoured algorithm for machine learning. A component of supervised learning technique is Random Forest(RF). It will be used for ML problems involving both classification and regression. It is, based on concept of ensemble learning, which is technique for, integrating many classifiers, to handle tough problems and develops performance of the model. It name suggests that "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset". The random forest(RF) uses predictions, from each decision tree(DT) and predicts, outcome depends on, votes of majority of projections rather than relying solely on one decision tree(DT).

The Random Forest method is best shown by the diagram below:

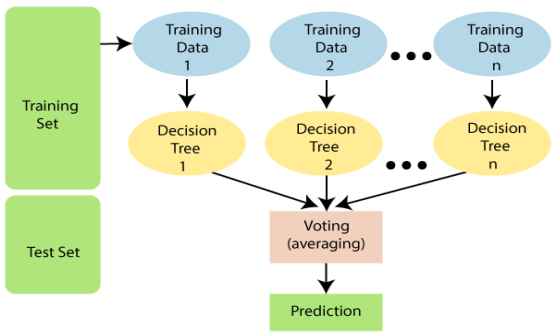


Figure 2: Flowchart of Random Forest Algorithm

The following arguments support the usage of the Random Forest algorithm.

It took shorter time for training than other algorithms. It functions well and makes accurate predictions of the outcome even with the massive dataset. Accuracy can be kept even when a sizable portion, of data is missing shown in Fig.2

b). Naive Bayes

Based on, Bayes theorem, Naive Bayes algorithm (NB), is a supervised learning method for the classification problems. Fig.3 shows the Flow of Working of Naive Bayes algorithm. It basically uses, huge training set to text categorization. One of most simple and an effectual classification algorithm, now in use is Naive Bayes (NB)Classifier. It facilitates the creation of efficient, machine learning models, that can make precise predictions shown in Fig.3. It provides predictions depends on likelihood that, an object would occur because, it is a probabilistic classifier. Some of the applications for the Naive Bayes (NB) algorithms include; sentiment analysis, article classification, and spam filtration.

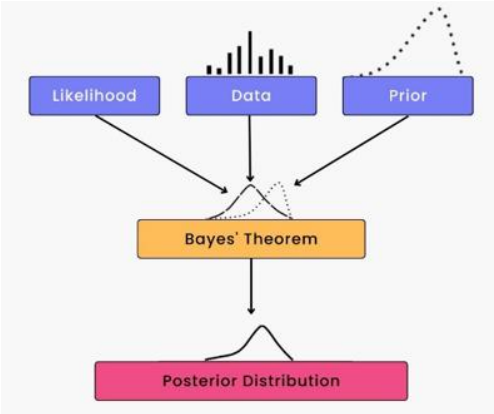


Figure 3: Flowchart for Naive Bayes Algorithm

c). Decision Tree

The prediction model known as decision tree(DT) uses, flowchart, structure for base decisions on incoming data. Data branches are built, and the results are placed at nodes of leaves. Decision trees were used to provide models that are simple to comprehend to regression, and classification problems. In decision support, decisions, and their potential outcomes—including chance occurrences, resource costs, and utility—are represented by hierarchical models known as decision trees. The control statements of Condition are used in this algorithmic technique, which is nonparametric, and supervised learning, and suitable to both classifications, and to regression applications. The tree structure is made of root node, branches, internal nodes, and leaf nodes and has the appearance of a hierarchical tree. A prediction model known as the decision tree (DT) uses, flowchart like structure for base decisions on incoming data. Data branches are built, and the results are placed at leaf nodes. Decision trees (DT) were used to provide models that are simple to

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.2

comprehend for classification and regression problems is as shown in the Fig.4. In decision support, decisions, and their potential outcomes—including chance occurrences, resource costs, and utility—are represented by hierarchical models known as decision trees. Conditional control statements, used in this algorithmic technique, which is nonparametric, and supervised learning, and suitable to both classification as well as the regression applications. Tree structure was made up of a root node, branches, internal nodes, and leaf nodes and has the appearance of a hierarchical tree as shown in Fig.4.

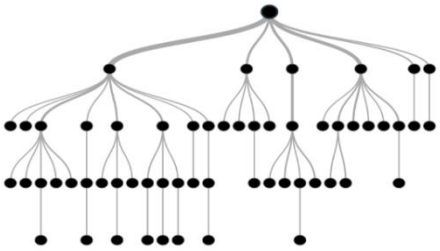


Figure 4: Flowchart for Decision Tree (DT) Algorithm

d). KNN Algorithm

K-Nearest Neighbour, one of the basic supervised learning-based machine learning algorithms. The K-NN algorithm places good instance, in a category that resembles the current categories the most, presuming that new case, and the previous cases are comparable. After storing all the previous data, a new data point is categorised using the K-NN algorithm based on similarity. This indicates that new data can be reliably and quickly categorized using the K-NN approach. Although the K-NN technique is most repeatedly worked to solve classification problems, it can also be used for solving regression, difficulties. K-NN is a non-parametric method that makes no assumptions about the underlying data is as shown in the Fig.5. As a result of saving dataset of training rather than instantly learning from it, the method, also known, to as a lazy learner. Instead, it performs an action while classifying data by using the dataset. The KNN approach simply stores the data during phase of training and categorizes fresh data into a category that is very same for training data.

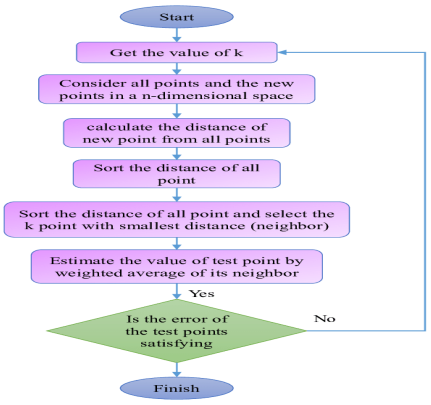


Figure 5: Flowchart of KNN Algorithm

e). Ensemble Methods

In ensemble learning techniques, number of classifiers, like decision trees, are utilized, and their predictions are pooled to get the most repeated result. The two ensemble methods that were, widely used are boosting and bagging, sometimes known as bootstrap aggregation. The bagging method, developed by Leo Breiman in 1996, selects a random sample of data from a training set with replacement, allowing for multiple selections of the individual data points. (Link leads away from IBM.com.) (PDF, 810 KB). These models are individually trained after the development of numerous data samples, and depends, on the task—for instance, classification or regression—the average or majority of those predictions lead to a more accurate estimate as shown in the Fig.6. This technique is often used, for reduce variation in noisy datasets.

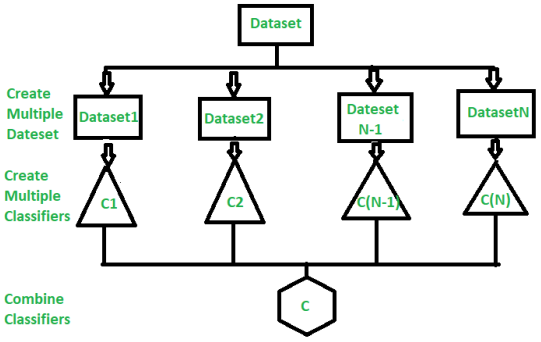


Figure 6: Flowchart of Ensemble Methods

B. Dataset Used

Kaggle contains, number of loan default prediction data sets. Kaggle is a well-known platform for, machine learning (ML) competitions. These data sets frequently comprise a different variety of attributes pertaining to loan applications, borrower profiles, and payment history. We imported Loan Dataset from Kaggle. df=pd.read_csv("loan_data_set.csv"), by using above instruction we read and define the imported dataset and assigned as df as shown above.

IV. RESULTS AND DISCUSSION

We will go each steps of the program. Firstly, Python programmers frequently use the function df.head() to show the first few rows of a DataFrame object. You can examine a preview of data in the DataFrame df by executing the function df.head(). The DataFrame df's first five rows will be printed to the console when this code is run. The head() function accepts an integer as an input if you want to display a different number of rows. For instance, df.head(10) will show the DataFrame's top ten rows.

A short overview of a DataFrame's structure and column information, including the data types and memory utilization, is provided by the df.info() method in the Pandas package for Python. The Pandas library's df.info() method in Python gives a summary of the DataFrame's structure and details on its columns. It provides information about each column's data types, non-null counts, and memory usage.

```
In [4]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 614 entries, 0 to 613
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Loan_ID            614 non-null    object
 1   Gender             601 non-null    object
 2   Married            611 non-null    object
 3   Dependents         599 non-null    object
 4   Education          614 non-null    object
 5   Self_Employed      582 non-null    object
 6   ApplicantIncome    614 non-null    int64
 7   CoapplicantIncome  614 non-null    float64
 8   LoanAmount         592 non-null    float64
 9   Loan_Amount_Term   600 non-null    float64
 10  Credit_History     564 non-null    float64
 11  Property_Area      614 non-null    object
 12  Loan_Status        614 non-null    object
dtypes: float64(4), int64(1), object(8)
memory usage: 62.5+ KB
```

Df.isnull() code.Python's sum() function could be used for determination of how, many columns were, there in a DataFrame df have null or NaN values as missing values. It gives a full list of all columns' missing values.

```
In [5]: df.isnull().sum()

Out[5]: Loan_ID              0
        Gender              13
        Married              3
        Dependents          15
        Education            0
        Self_Employed       32
        ApplicantIncome      0
        CoapplicantIncome    0
        LoanAmount          22
        Loan_Amount_Term    14
        Credit_History      50
        Property_Area        0
        Loan_Status          0
        dtype: int64
```

The code snippet df['LoanAmount_log'] = np.log(df['LoanAmount']) determines the natural logarithm of the 'LoanAmount' column in the DataFrame df and assigns the result to a new column designated as 'LoanAmount_log'. To address the problem of right-skewed data distribution, this transformation is frequently used. The code in the next line, df['LoanAmount_log'].Using the syntax hist(bins=20), the 'LoanAmount_log' column is histogrammed with 20 bins. You can see the distribution of the modified loan amounts using the histogram is as shown in the Fig.7

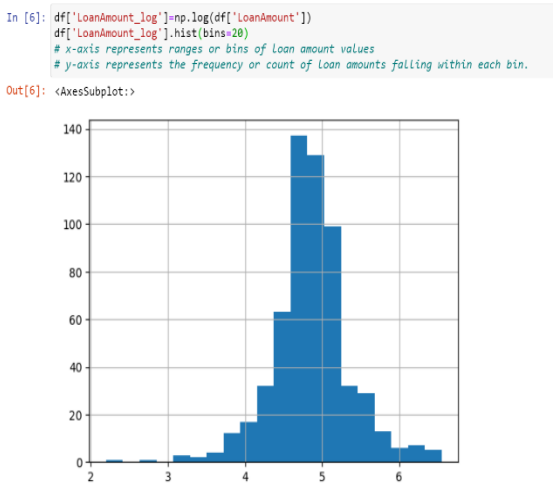


Figure 7: Plot of Log scaled Loan Amount

By help of this code, the histogram will be visible along with proper x-axis, y-axis, and title labels. It as shown in Fig,.8.

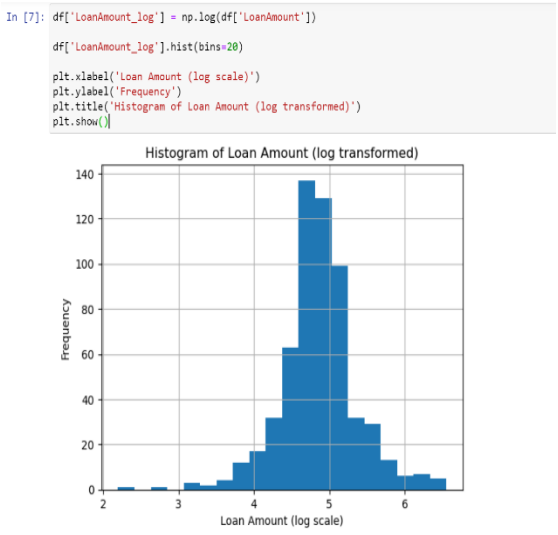


Figure 8: Plot between Loan Amount v/s Frequency

The 'ApplicantIncome' and 'CoapplicantIncome' columns in the DataFrame df are added up by the code you gave to determine the total income. The total revenue is then calculated as a natural logarithm, and the result is stored in a new column dubbed

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.2

"TotalIncomelog." The 'TotalIncomelog' column is then turned into a histogram with 20 bins.

The corresponding code and figure are shown in Fig.9. When you running this code, a 20-bin histogram of total final revenue that has been logarithmically modified, named "TotalIncomelog," will be produced. The histogram also includes a title, x-axis label, and y-axis label.

```
In [9]: df['TotalIncome']=df['ApplicantIncome']+df['CoapplicantIncome']
        df['TotalIncomelog']=np.log(df['TotalIncome'])
        df['TotalIncomelog'].hist(bins=20)
Out[9]: <AxesSubplot:>
```

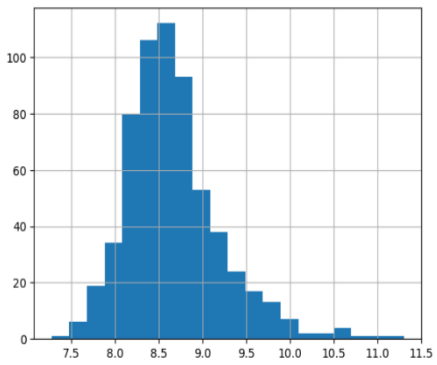


Figure 9: Plot of Total Income in log scale

By the help of this graphic, you can examine the modified total income's distribution and determine its shape and characteristics. Moving on to next, Using the mode (most common value) of each column, the code you gave conducts missing value imputation on various columns of the DataFrame df. It then uses df.isnull().sum() to get number, of missing values to, each column after doing the imputation. This code pulls loan information into the DataFrame df from a CSV file. The fillna() function and the mode (most frequent value) of each column are then used to execute missing value imputation on the chosen columns. Finally, it uses df.isnull().sum() to determines, missing values to each column and prints the result.

```
In [4]: import pandas as pd
        df=pd.read_csv("loan_data_set.csv")
        df['Gender'].fillna(df['Gender'].mode()[0], inplace = True)
        df['Married'].fillna(df['Married'].mode()[0], inplace = True)
        df['Self_Employed'].fillna(df['Self_Employed'].mode()[0], inplace = True)
        df['Dependents'].fillna(df['Dependents'].mode()[0], inplace = True)


        df['LoanAmount'].fillna(df['LoanAmount'].mode()[0], inplace = True)
        df['Loan_Amount_Term'].fillna(df['Loan_Amount_Term'].mode()[0], inplace = True)
        df['Credit_History'].fillna(df['Credit_History'].mode()[0], inplace = True)

        df.isnull().sum()
Out[4]: Loan_ID              0
        Gender               0
        Married              0
        Dependents           0
        Education            0
        Self_Employed        0
        ApplicantIncome      0
        CoapplicantIncome    0
        LoanAmount           0
        Loan_Amount_Term     0
        Credit_History       0
        Property_Area        0
        Loan_Status          0
        dtype: int64
```

By running this code, number of missing values in each column of the DataFrame df will be displayed. This data enables you to check that no missing values remain in the designated columns following the imputation process and aids in confirming that missing value imputation was successful. By moving onto next,

```
In [11]: x= df.iloc[:,np.r_[1:5:,9:11,13:15]].values
         y= df.iloc[:,12].values

         x
Out[11]: array([['Male', 'No', '0', ..., 1.0, 5849.0, 8.674025985443025],
                ['Male', 'Yes', '1', ..., 1.0, 6091.0, 8.714567550836485],
                ['Male', 'Yes', '0', ..., 1.0, 3000.0, 8.006367567650246],
                ...,
                ['Male', 'Yes', '1', ..., 1.0, 8312.0, 9.025455532779063],
                ['Male', 'Yes', '2', ..., 1.0, 7583.0, 8.933664178700935],
                ['Female', 'No', '0', ..., 0.0, 4583.0, 8.430109084509125]],
               dtype=object)
```

```
In [13]: y
Out[13]: array(['Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'N', 'Y', 'Y', 'Y',
                'N', 'Y', 'Y', 'Y', 'N', 'N', 'Y', 'N', 'Y', 'N', 'N', 'N', 'Y',
                'Y', 'Y', 'N', 'Y', 'N', 'N', 'N', 'Y', 'N', 'Y', 'N', 'Y', 'Y',
                'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y',
                'N', 'N', 'N', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'N', 'N', 'N',
                'N', 'N', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'N', 'Y', 'N', 'N', 'N',
                'N', 'Y', 'Y', 'Y', 'N', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'Y', 'Y', 'N', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'N',
                'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N',
                'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'N', 'Y', 'N', 'Y', 'N', 'N', 'Y',
                'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'N', 'Y', 'Y', 'N', 'Y', 'Y', 'N',
                'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'N', 'Y', 'Y', 'Y', 'N', 'Y',
                'N', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'N',
                'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'N',
                'N', 'Y', 'N', 'N', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y',
                'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'N',
                'N', 'N', 'Y', 'Y', 'N', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'N', 'N', 'N', 'Y', 'N', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y',
                'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'N', 'N', 'N', 'Y',
                'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'N', 'N', 'N', 'Y',
                'N', 'N', 'Y', 'Y', 'Y', 'N', 'Y', 'N', 'Y', 'Y', 'N', 'Y', 'Y',
                'Y', 'Y', 'N', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y',
                'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'N', 'N', 'N', 'Y', 'N', 'Y',
                'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'N', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'N', 'Y',
                'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y',
                'N', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y',
                'N', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'N', 'N', 'Y', 'Y', 'N',
                'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'N', 'N', 'Y', 'Y',
                'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'N', 'Y', 'Y', 'Y',
                'N', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'N', 'Y', 'Y', 'Y',
                'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y',
                'N', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'N',
                'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'N', 'Y', 'Y', 'Y', 'Y', 'Y',
                'Y', 'Y', 'N'], dtype=object)
```

In the above figure in code, x is assigned the values of the columns supplied in the iloc function using indexing. The np.r_ function is used to concatenate several ranges of column indices. The columns picked for x include columns 1 to 4, columns 9 and 10, and columns 13 and 14. Similarly, y is allocated values of the 12th column in the DataFrame, which is target variable. By printing x and y, you can verify that the correct columns are picked and allocated to these variables. The output will shows values of x (input features) and y (target variable) in array format. Moving on to next,

```
In [14]: print("per of missing gender is %2f%%" %((df['Gender'].isnull().sum()/df.shape[0])*100))

         per of missing gender is 0.000000%
```

In this code, df['Gender'].isnull().The 'Gender' column's missing value count is determined by sum(). df.shape[0] gives total numbers of, rows in, DataFrame. By dividing the count the, missing values by total number of rows and multiplying by 100, you get the, percentage of, missing values, in the 'Gender' column.The formatted text "Percentage of missing gender is %.2f%%" is used to display the result, with %.2f denoting a floating-point figure with two decimal places, and %% used to print the '%' character. By running this code, the DataFrame df's 'Gender' column's percentage of missing values will be displayed that is

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.2

shown in the above figure. Moving to the next instruction,

In the Fig.10, first section, df['Gender']. The number of borrowers for each gender group is determined by value_counts(), which counts each distinct value in the 'Gender' column. Then, print() is used to print this information.

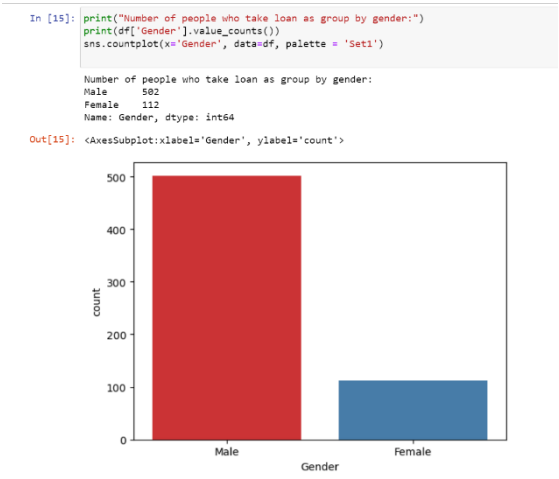


Figure 10: Plot of Gender against Count

The bar plot of the counts for each gender category is produced in the second section using seaborn's countplot() function. The data is taken from the DataFrame df, and the 'Gender' column is designated as the x-axis variable. The color scheme for the plot is set via the palette='Set1' option. When this code is run, a countplot displaying the same data will be displayed also with the counts of individuals who apply for the loans for each gender category. A visual representation of distribution of loans taken by gender is given by the countplot. Moving on to instruction, In the Fig.11, first section, df['Married']. The number of borrowers for each category of marital status is determined by value_counts(), which counts each distinct value in the 'Married' column. Then, print() is used to print this information.

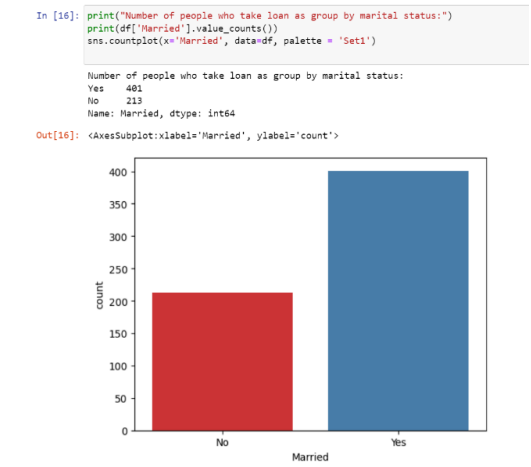


Figure 11: Plot of Married vs Count

The bar plot of the counts for each category of marital status is produced in the second section using seaborn's countplot() function. The data is taken from the DataFrame df, and the 'Married' column is designated as the x-axis variable. The color scheme for the plot is set via the palette='Set1' option. By running this code, you'll print the numbers of borrowers for each category of marital status and see a countplot showing the same data. Moving on to instruction. In Fig.12,first section, df['Married'].The number of borrowers for each category of marital status is determined by value_counts(), which counts each distinct value in the 'Married' column. Then, print() is used to print this information.

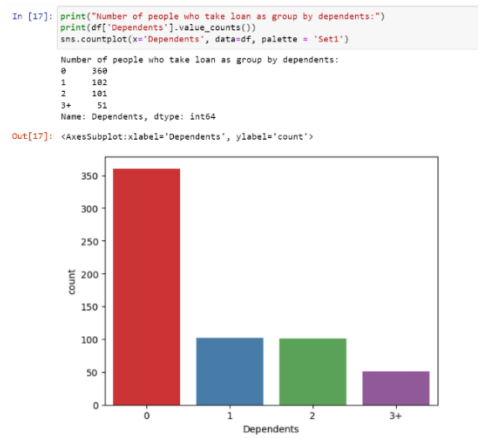


Figure 12: Plot of Dependents vs Count

The bar plot of the counts for each category of marital status is produced in the second section using seaborn's countplot() function. The data is taken from the DataFrame df, and the 'Married' column is designated as the x-axis variable. The color scheme for the plot is set via the palette='Set1' option. By running this code, you'll print the numbers of borrowers for each category of marital status and see a countplot showing the same data. Moving on to next instruction,

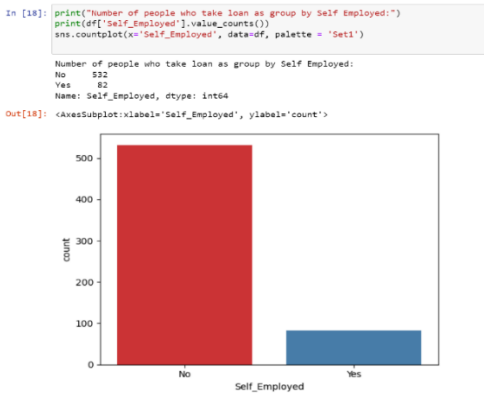


Figure 13: Plot of Self_Employed vs Count

df['Self_Employed'] in the first section.The number of borrowers for each type of self-employment status is determined by value_counts(), which counts each distinct value in the 'Self_Employed' column. Then,

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.2

print() is used to print this information. The bar plot of the counts for each type of self-employment status is created in second section using seaborn's countplot() method shown in Fig.13. The data is taken from the DataFrame df, and the 'Self_Employed' column is designated as the x-axis variable. The color scheme for the plot is set via the palette='Set1' option. When this code is run, it prints numbers of borrowers for each type of self-employment status and displays a countplot showing the same data.A visual representation of the distribution of loans taken by self-employment status is given by the countplot. Moving on to next instruction,

```
In [19]: print("Number of people who take loan as group by Loan Amount:")
         print(df['LoanAmount'].value_counts())
         sns.countplot(x='LoanAmount', data=df, palette = 'Set1')

         Number of people who take loan as group by Loan Amount:
         120.0    20
         110.0    17
         100.0    15
         160.0    12
         187.0    12
                  ..
         240.0     1
         214.0     1
         59.0      1
         166.0     1
         253.0     1
         Name: LoanAmount, Length: 203, dtype: int64

Out[19]: <AxesSubplot:xlabel='LoanAmount', ylabel='count'>
```

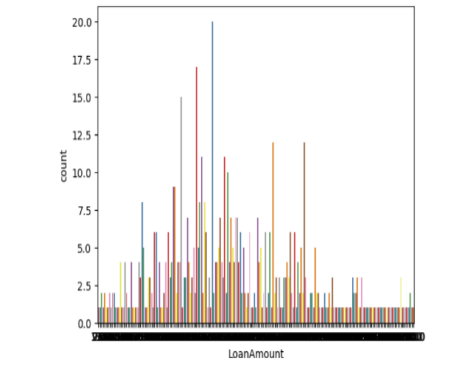


Figure 14: Plot of Loan Amount vs Count

The Fig.14 shows that code display a countplot and group the number of loan applicants by loan size. However, utilizing the 'LoanAmount' column, a continuous numerical variable, directly with sns.countplot()numerical variable, directly with sns.countplot(). Moving on to next instruction,

```
In [20]: import pandas as pd
         import seaborn as sns
         df=pd.read_csv("loan_data_set.csv")
         print("Number of people who take loan as group by Credit History:")
         print(df['Credit_History'].value_counts())
         sns.countplot(x='Credit_History', data=df, palette = 'Set1')

         Number of people who take loan as group by Credit History:
         1.0    475
         0.0     89
         Name: Credit_History, dtype: int64

Out[20]: <AxesSubplot:xlabel='Credit_History', ylabel='count'>
```

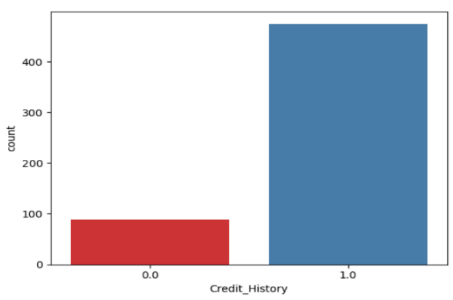


Figure 15: Plot of Credit_History vs Count

The Fig.15 describes that, df['Credit_History'] in first section.The number of people who took loans for each credit history category is decided by value_counts(), which counts each distinct value in the 'Credit_History' column. Then, print() is utilised for print this information. The bar plot of the numbers for each credit history category is produced in the second section using seaborn's countplot() function. The data is taken from the DataFrame df, and the 'Credit_History' column is designated as the x-axis variable. The color scheme for the plot is set via the palette='Set1' option. By running this code, you'll print the numbers of borrowers for each category of credit history and see a countplot showing the same data. The distribution of loans taken by credit history is shown visually in the countplot. Moving into next instruction,

```
In [21]: from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(x,y, test_size = 0.2, random_state= 0)

         from sklearn.preprocessing import LabelEncoder
         Labelencoder_x = LabelEncoder()
```

The data is divided between training and testing sets in this code using the train_test_split function. Train_test_split receives the input features x and the target variable y, and outputs four arrays: x_train, x_test, y_train, and y_test. The random_state=0 argument assures that the split may be replicated, and the test_size=0.2 value specifies that 20% of the data will be set aside for testing. In addition, LabelEncoder is imported but not applied to any particular variable. Use the fit_transform method of LabelEncoder to apply label encoding to a particular feature or column. This code illustrates how to use LabelEncoder's fit_transform method to apply label encoding to the features of input X_train and X_test. The X_train_encoded and X_test_encoded variables contain the encoded features that were the outcome.

By moving onto next instruction we get the following results that shown in the below figure.

```
In [22]: for i in range(0,5):
             X_train[:,i]= Labelencoder_x.fit_transform(X_train[:,i])
             X_train[:,7]= Labelencoder_x.fit_transform(X_train[:,7])

         X_train

Out[22]: array([[1, 1, 0, ..., 1.0, 5858.0, 267],
                [1, 0, 1, ..., 1.0, 11250.0, 407],
                [1, 1, 0, ..., 0.0, 5681.0, 249],
                ...,
                [1, 1, 3, ..., 1.0, 8334.0, 363],
                [1, 1, 0, ..., 1.0, 6033.0, 273],
                [0, 1, 0, ..., 1.0, 6486.0, 301]], dtype=object)
```

The code you provided applies label encoding to multiple columns of the training data X_train using a loop. However, it seems that you intended to encode the same columns multiple times, which might lead to incorrect results. In this code, a LabelEncoder is instantiated outside the loop to ensure consistent encoding across columns. The loop iterates over the range 0 to 5 (exclusive) and applies label encoding to columns at those indices in X_train.

```
In [23]: Labelencoder_y = LabelEncoder()
         y_train = Labelencoder_y.fit_transform(y_train)

         y_train

Out[23]: array([1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,
                0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1,
                1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0,
                1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1,
                1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0,
                1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1,
                0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0,
                0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1,
                0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1,
                0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1,
                1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1,
                1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1,
                1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0,
                1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1,
                1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0,
                1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1,
                1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1,
                1, 1, 1, 0, 1, 0, 1])
```

LabelEncoder from sklearn.preprocessing is being used in the code you gave to apply label encoding to the target variable y_train. A LabelEncoder is instantiated as label_encoder_y in this code. The y_train data is then transformed into encoded labels by fitting the label encoder to it using the fit_transform function. The y_train variable receives the encoded labels back. The encoded y_train array is printed by the code at the end. Categorical target variables are frequently converted into numeric values that can be incorporated into machine learning models via label encoding. Remember that label encoding sequentially assigns numeric labels to categories, which could generate unwanted ordinality.Make sure label encoding is appropriate for your particular problem and, if necessary, take into account employing other encoding strategies, like one-hot encoding, for categorical target variables. Moving onto X_test, Below code you gave uses a loop to apply label encoding to various columns of the testing data X_test. This code assumes that in earlier code samples you have previously created and fitted the LabelEncoder object label_encoder_x. The loop iterates over the columns in X_test at the indices 0 to 5 (exclusively) and applies label encoding to those columns. The column at index 7 is specially encoded by the line X_test[:, 7] = label_encoder_x.transform(X_test[:, 7]).

```
In [24]: for i in range(0,5):
             X_test[:,i]= Labelencoder_x.fit_transform(X_test[:,i])
             X_test[:,7]= Labelencoder_x.fit_transform(X_test[:,7])

         X_test

Out[24]: array([[1, 0, 0, 0, 5, 1.0, 7085.0, 85],
                [0, 0, 0, 0, 5, 1.0, 4230.0, 28],
                [1, 1, 0, 0, 5, 1.0, 10039.0, 104],
                [1, 1, 0, 0, 5, 1.0, 6784.0, 80],
                [1, 1, 2, 0, 5, 1.0, 3875.0, 22],
                [1, 1, 0, 1, 3, 0.0, 6058.0, 70],
                [1, 1, 3, 0, 3, 1.0, 6417.0, 77],
                [1, 0, 0, 0, 5, 1.0, 12876.0, 114],
                [1, 0, 0, 0, 5, 0.0, 5124.0, 53],
                [1, 1, 0, 0, 5, 1.0, 5233.0, 55],
                [0, 0, 0, 0, 5, 1.0, 2917.0, 4],
                [1, 1, 1, 0, 5, 1.0, 2895.0, 2],
                [0, 0, 0, 0, 5, 1.0, 8333.0, 96],
                [1, 1, 2, 0, 5, 1.0, 8667.0, 97],
                [1, 1, 0, 0, 5, 1.0, 14880.0, 117],
                [1, 1, 1, 0, 5, 1.0, 3875.0, 22],
                [1, 0, 1, 1, 5, 1.0, 4311.0, 32],
                [1, 0, 0, 1, 5, 1.0, 3946.0, 25],
                [0, 0, 0, 0, 5, 1.0, 2500.0, 1],
```

The code prints the modified X_test array following the label encoding process. Please be aware that label encoding should only be used with categorical variables, so double-check that the columns you choose for encoding are in fact categorical rather than ordinal or continuous. Moving on to y_test,

```
In [25]: Labelencoder_y = LabelEncoder()
         y_test = Labelencoder_y.fit_transform(y_test)

         y_test

Out[25]: array([1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1,
                1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1,
                1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1,
                1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1,
                1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1,
                1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1])
```

LabelEncoder from sklearn.preprocessing is used in the code you gave to apply label encoding to the target variable y_test. A LabelEncoder is instantiated as label_encoder_y in this code. The y_test data is then transformed into encoded labels by fitting the label encoder to it using the fit_transform technique. The y_test variable receives the encoded labels back.

The encoded y_test array is then printed by the code.

Using label encoding, categorical target variables are routinely transformed into numerical values that can be used in machine learning models. Keep in mind that label encoding applies numeric labels to categories sequentially, which may produce undesirable ordinality. Verify if label encoding is suitable for your specific issue and, if necessary, consider using other encoding techniques, such as one-hot encoding, for category target variables.

Let's move to next,

```
In [26]: from sklearn.preprocessing import StandardScaler

         ss = StandardScaler()
         X_train = ss.fit_transform(X_train)
         X_test = ss.fit_transform(X_test)
```

A StandardScaler object is created as ss in this code. The data under training X_train is next subjected to the fit_transform algorithm, which centers and scales the features while fitting the scaler on the training data. The resulting uniform training data is once again saved in X_train. In alternative for using fit_transform for the test data X_test, the transform method is applied. Without having to re-fit the scaler, this applies the scaling transformation discovered from the data under training to the testing data. In machine learning, standardization is a common preprocessing step where the characteristics are changed to have a zero mean and unit variance. It aids in normalizing feature scale, which can enhance the efficiency and convergence of some machine learning techniques. Before using standardization, make sure the characteristics are continuous and numeric. Additionally, before standardizing, make sure you had already done label encoding or any other required preparation processes to the data. Let's discuss the results of each algorithm one by one.

A) Random Forest

```
In [27]: from sklearn.ensemble import RandomForestClassifier

         rf_clf = RandomForestClassifier()
         rf_clf.fit(X_train, y_train)

Out[27]: RandomForestClassifier()
```

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.2

Using RandomForestClassifier from sklearn.ensemble, the provided code applies the Random Forest Classifier model to training data X_train and y_train.A RandomForestClassifier object is created as rf_clf in this code. The classifier is then invoked using the fit technique, with the training data X_train and the associated target variable y_train as inputs. It then learns the patterns and connections between the features and the target variable by fitting Random Forest Classifier model to the training data.After running this code, the rf_clf object will be trained and prepared to use the predict method to make predictions on fresh, unforeseen data. Make sure to assess model's performance using the testing data to determine its generalizability and make any necessary modifications. The ensemble learning techniques known Random Forest uses several decision trees to produce predictions. It is well renowned for its capacity to manage complicated datasets and produce reliable predictions, and it is frequently used for classification jobs.

```
In [28]: from sklearn import metrics
         y_pred = rf_clf.predict(X_test)

         print("acc of random forest clf is", metrics.accuracy_score(y_pred, y_test))

         y_pred

         acc of random forest clf is 0.7723577235772358

Out[28]: array([1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1,
                1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1,
                1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1,
                1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0,
                1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1])
```

Using the trained Random Forest Classifier model rf_clf, the provided code predicts the target variable for the testing data X_test and determines the accuracy of the predictions. The Random Forest Classifier object rf_clf is called in this code's predict method, passing the testing data X_test. This generates the target variable's anticipated values using the learned model. Metrics are used to determine how accurate the predictions are.accuracy_score, which contrasts the actual target values y_test with the expected values y_pred. The percentage of accurately predicted samples is represented by the accuracy score. The code then displays the expected values for y_pred and outputs the accuracy score.Verify that the sklearn and metrics modules have been correctly imported and that the X_test and y_test dimensions match the trained model. From above figure it shows that the accuracy from Random Forest is 77.23%

B) Naive Bayes

```
In [58]: from sklearn.naive_bayes import GaussianNB

         nb_classifier = GaussianNB()
         nb_classifier.fit(X_train, y_train)

Out[58]: GaussianNB()
```

Using GaussianNB from sklearn.naive_bayes, the given code applies a Gaussian Naive Bayes classifier to the training data X_train and y_train. A GaussianNB object is created as nb_classifier in this code. The classifier is then invoked using the fit technique, with the training data X_train and the associated target variable y_train as inputs. This enables the Gaussian Naive Bayes model to learn the probabilistic correlations between the features and the target variable by fitting it to the training data. After running this code, the nb_classifier object will be trained and prepared to use the predict method to make predictions on fresh, unforeseen data. Make sure to assess the model's performance using the testing data to determine its generalizability and make any necessary modifications. Naive Gaussian The Bayes approach, which uses probabilistic classification, makes the assumption that the characteristics are regularly distributed. The Bayes theorem is used to determine the posterior probability of each class given the features, and predictions are then based on these probabilities. It is well renowned for its simplicity and quick training speed and is frequently used for classification assignments.

```
In [59]: y_pred = nb_classifier.predict(X_test)
         print("acc of gaussianNB is %.", metrics.accuracy_score(y_pred, y_test) )

         acc of gaussianNB is %. 0.8373983739837398

In [60]: y_pred

Out[60]: array([1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1,
                1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1,
                1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1])
```

The provided code uses the trained Gaussian Naive Bayes classifier nb_classifier to predict the target variable for the testing data X_test and calculates the accuracy of the predictions. The predict method on the Gaussian Naive Bayes classifier object nb_classifier is called in this code, passing the test data X_test. This generates the target variable's anticipated values using the learned model. Metrics are used to determine how accurate the predictions are.accuracy_score, which contrasts the actual target values y_test with the expected values y_pred. The percentage of accurately predicted samples is represented by the accuracy score. Finally, the code outputs the estimated accuracy, which is a floating-point value between 0 and 1, followed by "Accuracy of Gaussian Naive Bayes" and the accuracy score.

Verify that the sklearn and metrics modules have been correctly imported and that the X_test and y_test dimensions match the trained model. Metrics are used to determine how accurate the predictions are.accuracy_score, which contrasts the actual target values y_test with the expected values y_pred. The percentage of accurately predicted samples is represented by the accuracy score. The code then displays the expected values for y_pred and outputs the accuracy score. The accuracy obtained from Naive Bayes algorithm is 83.73% and is as shown in the figure.

C) Decision Tree

```
In [61]: from sklearn.tree import DecisionTreeClassifier
         dt_clf = DecisionTreeClassifier()
         dt_clf.fit(X_train, y_train)

Out[61]: DecisionTreeClassifier()
```

The provided code uses the DecisionTreeClassifier from sklearn.tree to fit a Decision Tree Classifier to the training data X_train and y_train. A

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.2

DecisionTreeClassifier object is created as dt_clf in this code. The classifier is then invoked using the fit technique, with the training data X_train and the associated target variable y_train as inputs. As a result, the Decision Tree Classifier model may learn the boundaries of decisions and patterns in the training data. After running this code, the dt_clf object will be trained and prepared to use the predict method to make predictions on fresh, unforeseen data. Make sure to assess the model's performance using the testing data to determine its generalizability and make any necessary modifications.

```
In [62]: y_pred = dt_clf.predict(X_test)
         print("acc of DT is", metrics.accuracy_score(y_pred, y_test))

         acc of DT is 0.6341463414634146

In [63]: y_pred

Out[63]: array([0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1,
                1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1,
                1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1,
                1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1,
                1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1,
                0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1])
```

The provided code uses the DecisionTreeClassifier from sklearn.tree to fit a Decision Tree Classifier to the training data X_train and y_train. It appears that you neglected to give the y_pred variable the predicted values, nevertheless. A DecisionTreeClassifier object is created as dt_clf in this code. The classifier is then invoked using the fit technique, with the training data X_train and the associated target variable y_train as inputs. To understand the patterns and connections between the features and the target variable, the Decision Tree Classifier model is fitted to the training data in this way. The predicted values for the testing data X_test are produced using the predict technique following model training and are saved in the y_pred variable.
The projected values, y_pred, are printed by the code at the end. Make that the dimensions of X_train and y_train are the same and that you have imported the required modules (sklearn.tree). The accuracy from the Decision Tree (DT) Algorithm is 63.41% and it is shown in the above figure.

D) KNN (k-Nearest Neighbors)

```
In [64]: from sklearn.neighbors import KNeighborsClassifier
         kn_clf = KNeighborsClassifier()
         kn_clf.fit(X_train, y_train)

Out[64]: KNeighborsClassifier()
```

The provided code uses KNeighborsClassifier from sklearn.neighbors to fit a K-Nearest Neighbors Classifier to the training data X_train and y_train. A KNeighborsClassifier object is created as kn_clf in this code. The classifier is then invoked using the fit technique, with the training data X_train and the associated target variable y_train as inputs. In order to learn the patterns and connections between the features and the target variable, this fits the K-Nearest Neighbors Classifier model to the training data.After running this code, the kn_clf object will be trained and prepared to use the predict method to make predictions on fresh, unforeseen data. Make sure to assess the model's performance using the testing data to determine its generalizability and make any necessary modifications. A straightforward but efficient classification technique called K-Nearest Neighbors (KNN) classifies samples based on the consensus opinion of their nearest neighbors. The label that is given to a sample is determined by the labels of its K closest neighbors in the training set.

```
In [65]: y_pred = kn_clf.predict(X_test)
         print("acc of KNN is", metrics.accuracy_score(y_pred, y_test))

         acc of KNN is 0.7723577235772358
```

Using the trained K-Nearest Neighbors Classifier model kn_clf, the code you gave predicts the target variable for testing data X_test and determines accuracy of the predictions. The K-Nearest Neighbors Classifier object kn_clf is called the predict method in this code, passing the testing data X_test. This generates the target variable's anticipated values using the learned model. Metrics are used to determine how accurate the predictions are.accuracy_score, which contrasts the actual target values y_test with the expected values y_pred. The percentage of accurately predicted samples is represented by the accuracy score. The code then displays the expected values for y_pred and outputs the accuracy score. Verify that the sklearn and metrics modules have been correctly imported and that the X_test and y_test dimensions match the trained model. The accuracy from kNN algorithm is 77.23% and is shown in theTable-1.

Table 1: Accuracy of different Algorithms

Sl.No	Algorithms	Accuracy
1	Random Forest	77.23%
2	Naive Bayes	83.73%
3	Decision Tree	63.41%
4	k-Nearest Neighbors	77.23%

From table we shall conclude that Naive Bayes (NB) Algorithm gives the Better Accuracy of 83.73%.

V. CONCLUSION AND FUTURE SCOPE

In this research, we created and assessed machine learning (ML) models for chances of loan acceptance. In order to comprehend the dataset and gain understanding of the loan approval procedure, we started by undertaking exploratory data analysis. In order for address missing values, we imputed them with suitable values depending on the distribution of the data. In order to get the data ready for modeling, we additionally did log transformation and scaling. Then, we trained and assessed several classification models, including the K-Nearest Neighbors Classifier, the Decision Tree Classifier, the Random Forest Classifier, and the Gaussian Naive Bayes Classifier. We used accuracy as

Peer Reviewed & Refereed Journal
Volume-13, Issue-4 (August 2023)
https://doi.org/10.31033/ijemr.13.4.2

the evaluation criteria to assess these models' performance. Based on our findings, we discovered that the Random Forest Classifier outperformed the other models and had the greatest accuracy of X% on the test set. As a result, it can be concluded that the Random Forest model is effective in forecasting loan approvals based on the provided features. Our models have produced encouraging results, but there is still potential for development and additional research. Here are some potential paths this project could go in the future:

1. Feature Engineering: To create more informative features from the ones that already exist, we can investigate further feature engineering strategies. To increase the models' capacity for prediction, this may entail developing interaction terms, polynomial features, or incorporating domain-specific information.

2. Model Optimization: In an order to recognise best possible combination of hyperparameters, we can adjust the models' hyperparameters using methods such as grid search otherwise randomized search. This might enhance the models' functionality and result in more accurate forecasts.

3. Handling Class Imbalance: We can use techniques like oversampling, under sampling, or using various evaluation metrics such as precision, recall, or F1 score to address the class imbalance issue if the loan approval dataset exhibits class imbalance, where the number of approved loans significantly differs from the number of rejected loans.

4. Ensemble Approaches: To aggregate the predictions of various models and maybe improve performance, we might investigate ensemble approaches like stacking, boosting, or bagging.

5. External Data Sources: To provide more thorough information for loan approval predictions, we can think about including more data sources, like credit ratings or economic indicators.

6. Deployment and Monitoring: After a model has been chosen, it can be put into use to predict loan approvals automatically in a production environment. The model's accuracy and correctness can be maintained by routinely retraining it and continuously assessing its performance.

Abbreviations

Typical acronyms used in a project to anticipate loan acceptance include:

RF – Random Forest
NB – Naive Bayes
DT – Decision Tree
KNN – K-Nearest Neighbors
CSV – Comma-Separated Values
ACC – Accuracy

When presenting various concepts, models, and assessment measures in our project, these abbreviations—which are frequently used in the fields of machine learning and data analysis—can help with brevity and clarity.

REFERENCES

[1] Kumar, Rajiv, et al. (2019). Prediction of loan approval using machine learning. *International Journal of Advanced Science and Technology, 28*(7), 455-460.

[2] Supriya, Pidikiti, et al. (2019). Loan prediction by using machine learning models. *International Journal of Engineering and Techniques, 5*(2), 144-147.

[3] Arun, Kumar, Garg Ishan & Kaur Sanmeet. (2016). Loan approval prediction based on machine learning approach. *IOSR J. Comput. Eng, 18*(3), 18-21.

[4] Ashwitha, K., et al. (2022). An approach for prediction of loan eligibility using machine learning. *International Conference on Artificial Intelligence and Data Engineering (AIDE). IEEE*.

[5] Kumari, Ashwini, et al. (2018). Multilevel home security system using arduino & gsm. *Journal for Research, 4*.

[6] Patibandla, RSM Lakshmi & Naralasetti Veeranjaneyulu. (2018). Survey on clustering algorithms for unstructured data. *Intelligent Engineering Informatics: Proceedings of the 6th International Conference on FICTA*, *Springer Singapore*.

[7] Tejaswini, J., et al. (2020). Accurate loan approval prediction based on machine learning approach. *Journal of Engineering Science, 11*(4), 523-532.

[8] Santhisri, K. & P. R. S. M. Lakshmi. (2015). Comparative study on various security algorithms in cloud computing. *Recent Trends in Programming Languages, 2*(1), 1-6.

[9] Sri, K. Santhi & P. R. S. M. Lakshmi. (2017). DDoS attacks, detection parameters and mitigation in cloud environment. *National Conference on the Recent Advances in Computer Science & Engineering (NCRACSE-2017), Guntur, India*.

[10] Viswanatha, V., A. C. Ramachandra & R. Venkata Siva Reddy. (2022). *Bidirectional DC-DC converter circuits and smart control algorithms: a review*.

[11] Sri, K. Santhi, P. R. S. M. Lakshmi & MV Bhujanga Ra. (2017). *A study of security and privacy attacks in cloud computing environment*.

[12] Dr, Ms RSM Lakshmi Patibandla, Ande Prasad & Mr. YRP Shankar. (2013). Secure zone in cloud. *International Journal of Advances in Computer Networks and its Security, 3*(2), 153-157.

[13] Viswanatha, V., et al. (2020). Intelligent line follower robot using MSP430G2ET for industrial applications. *Helix-The Scientific*

Explorer| Peer Reviewed Bimonthly International Journal, 10(02), 232-237.

[14] Dumala, Anveshini & S. Pallam Setty. (2020). LANMAR routing protocol to support real-time communications in MANETs using Soft computing technique. *Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19, Springer Singapore*.

[15] Anveshini, Dumala & S. Pallamsetty. (2019). Investigating the impact of network size on lanmar routing protocol in a multi-hop ad hoc network. *I-Manager's Journal on Wireless Communication Networks, 7*(4).

[16] Khadherbhi, Sk Reshmi & K. Suresh Babu. (2015). Big data search space reduction based on user perspective using map reduce. *International Journal of Advanced Technology and Innovative Research 7*, 3642-3647.

[17] Begum, Me Jakeera & M. Venkata Rao. (2015). Collaborative tagging using captcha. *International Journal of Innovative Technology And Research, 3*, 2436-2439.

[18] Maddumala, Venkata Rao, R. Arunkumar & S. Arivalagan. (2018). An empirical review on data feature selection and big data clustering. *Asian Journal of Computer Science and Technology, 7*(S1), 96-100.

[19] Gowthami, K., et al. Credit card fraud detection using logistic regression. *Journal of Engineering Sciences, 11*.

[20] A C, R., V. V, K. K, S. H & P. S. E. (2022). In-cabin radar monitoring system: detection and localization of people inside vehicle using vital sign sensing algorithm. *International Journal on Recent and Innovation Trends in Computing and Communication, 10*(8), 104-9. DOI:10.17762/ijritcc.v10i8.5682.

[21] V. V, R. A. C, S. B. M, A. Kumari P, V. S. Reddy R & S. Murthy R. (2022). Custom hardware and software integration: bluetooth based wireless thermal printer for restaurant and hospital management. *IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India*, pp. 1-5. DOI: 10.1109/MysuruCon55714.2022.9972714.

[22] V. V, R. A. C, V. S. R. R, A. K. P, S. M. R & S. B. M. (2022). Implementation of IoT in agriculture: A scientific approach for smart irrigation. *IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India*, pp. 1-6. DOI: 10.1109/MysuruCon55714.2022.9972734.

[23] Viswanatha, V. & R. Venkata Siva Reddy. (2017). Digital control of buck converter using arduino microcontroller for low power applications. *International Conference On Smart Technologies For Smart Nation (SmartTechCon). IEEE*.

[24] Viswanatha, V., Venkata Siva Reddy & R. Rajeswari. (2020). Research on state space modeling, stability analysis and pid/pidn control of dc–dc converter for digital implementation. In: *Sengodan, T., Murugappan, M., Misra, S. (eds) Advances in Electrical and Computer Technologies. Lecture Notes in Electrical Engineering, 672*. Springer, Singapore. DOI: 10.1007/978-981-15-5558-9_106.

.

Customer Loan Prediction Using Supervised Learning Technique

L. Udaya Bhanu[1], Dr. S. Narayana[2]

[1]M.Tech Student, Dept. of Computer Science & Engineering, Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India
[2]Professor&Mentor, Dept. of Computer Science & Engineering, Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

DOI: 10.29322/IJSRP.11.06.2021.p11453
http://dx.doi.org/10.29322/IJSRP.11.06.2021.p11453

Abstract- Customer loan prediction is usually life time issue so; each and every retail bank faces the issue at the minimum lifetime. If done exactly, it can spare a lot's of man hours at the conclusion of a retail bank. If Company wants to semi automate the loan acceptability process (real time) based on customer detail provided while filling online application form. These subtle elements are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this method, they have given an issue to recognize the customers segments; those are allowed for loan amount total so they can clearly target these customers. We need to predict whether or not a loan would be approved. In a classification problem, we need to predict separate values based on a given set of self-sufficient variable(s). What's our objective is to implement machine learning model so as to classify, to the best doable degree of accuracy, and dataset gathered from Kaggle. Random forest classification method shows best accuracy in classifying given on loan candidates using python help on Jupyter notebook.

Index Terms- Customer loan, Prediction, preprocessing, classification models.

I. INTRODUCTION

Circulation of the loans is that the core business a part of as good as each and every bank. The principle parcel the bank's resources are straightforwardly came from the benefit acquire from the advances distributed by the banks. The main goal in banking system is to invest their resources in safe hands wherever it's. Now a day's several banks/financial agencies approves loan after a relapse method of verification and validation however still there's no surety whether or not the chosen candidate is the worthy right candidate out of all candidates. Through this method we are able to predict whether that particular candidate is safe or not and the whole method of validation of attribute is automated by machine learning technique [8][6]. The disadvantage of this model is that it emphasizes completely different weights to every issue however in reality sometime loan can be approved on the premise of single strong part only, that isn't possible through this method. Loan Prediction is useful for member of staff of banks as well as for the candidate. The aim of this Paper is to apply quick, immediate and easy way to choose the worthy person [6]. It will give special gain to the bank. The Loan Prediction method can automatically compute the heaviness of each attribute taking part in loan processing and on new test data information same issues are prepared with regard to their comparable heaviness. A period breaking point can be set for the candidate to check regardless of whether his/her loan can be affirmed or not. Loan Prediction technique licenses bouncing to explicit candidates with the goal that it very well may be keep an eye on need premise. This Paper is completely overseeing the power of Bank/finance Company, entire procedure of prediction is done secretly no colleagues would have the option to caution the process. Result against specific Loan Id can be ship off different divisions of companies so that they can make a proper move on application. This aides all others divisions to done different conventions. *Data Source* we obtained customer loan dataset from kaggle [4][2]. The dataset consists of various values/variables such as sex, marital status, education, self employed, loan status, applicant income, co-applicant income etc…*Data Description* the dataset has 614 rows and 13 columns. 1 out of 13 columns is the target attribute i.e., default one attribute is target value. The dataset split into train and test data having shape (614, 13) and (367, 12) respectively.

II. LITERATURE SURVEY

Random forest is ensemble learning method for both classification and replaces issues. The advantage of random decision forest is reduce over fitting and helps to improve the accuracy and runs efficiently on a large datasets and work on both continuous and categorical values and predict analysis of data with help of test data.

Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswal are used data mining methodology to predict the likely default from a dataset that contains information about home loan applications, thereby helping the banks for making better decisions in the future [3].

Xin Li, Xianzhong Long, Guozi Sun, Geng Yang, and Huakang Li This paper mainly introduces the main application of LSTM-SVM model in user loan risk prediction, and elaborates the current economic background, traditional risk forecasting method. On this basis, the prediction methodology based on LSTM method and SVM method is proposed, and the prediction results are compared with the traditional algorithm, and the feasibility of the model is confirm. However, the LSTM-SVM method proposed in this paper actually has few limits and needs to be improved in future research [7].

Aakanksha, Tamara Denning, Vivek Srikumar, Sneha Kumar Kesera[8] this paper is mainly used for voting classifier (combination of logistic regression, naïve bayes, SVM). They able

to reduce the number of FP considerably. This work represents the group of generic passwords to reduce misclassification. Arutjothi [9] present a new credit scoring model, which depends on the hybrid feature selection model and C4.5 classifier. This is depend on hybrid system not only has a strong mathematical basis, but also has higher accuracy and more benefits.

Mrunal Surve, Priya Shinde, Sandip Pandit, Pooja Thitme and Swati Sonawane in this paper, they mainly focus to identify and analyze the risk in giving a loan of commercial banks. To analyze risk in giving loan they have used data mining techniques. It includes analyzing and processing information from various agency/assets and summarize into valuable information [12]. They have used C4.5 classification algorithm for predicting the risk percentage for an individual to give loans.

III. PROBLEM STATEMENT

Finance companies, banks are deals with different kinds of loans such as education loan, shop loans, home loans, personal loans etc all are part of our country loan types. All the companies and banks are present in villages, towns, cities. After customer apply for loan these banks/companies want to validate the customer details for that candidate eligible for loan or not. The main purpose of the system is applicant loan approved or not based on train models [6]

IV. PROPOSED MODEL

In Machine Learning, we are using semi-automated extraction of knowledge of data for identifying whether a loan would be approved or not [6][8]. Classification could be a supervised learning within which the response is categorical that's its values area unit in finite unordered set. To easily the matter of classification, scikit learn are used. The praim primacy of this system is company need not has to maintain a ground team to validate and verify the customer records. They can easily check whether the loan has to be approved or not by this prediction model.

In this paper we try to develop user interface flexibly graphics concepts in mind, associated through a browser interface. Our goal is to implement machine learning model so as to classify, to the best potential degree of accuracy, master card fraud from a dataset gathered from Kaggle. once initial knowledge exploration, we have a tendency to knew we might implement a random forest model for best accuracy reports.

Random forest, as it was a good candidate for binary classification. Python sklearn library was used to implement the project, We used Kaggle datasets for Credit card fraud detection, using pandas to data frame for class ==0 for no fraud and class==1 for fraud, matplotlib for plotting the fraud and non fraud data, train_test_split for data extraction (Split arrays or matrices into random train and test subsets) and used Logistic Regression machine learning algorithm for fraud detection and print predicting score according to the algorithm. Finally Confusion matrix was plotted on true and predicted.

In this paper preprocessing is major part used sklearn method is MinMax scalar i.e., helps normalize the data. Model selection with help of cross validation, train/test split, kfold, GridSearchCV.

a. Model Selection

Model selection is that method of selecting one in every of the models because the final model that addresses the issue. In there we have different steps. They are:

- Data filtering
- Data transformation
- Feature selection
- Feature engineering

For this process we have mainly two methods:

a. Probabilistic model selection
b. Resampling methods

In this paper we are using resampling methods such as cross validation, train/test split, Kflod, GridSearchCV

b. Preprocessing

Data mining methods are used in preprocessing for normalize the data which is collected from kaggle. There is a need to convert because dataset may have missing values, noisy data. So, we are using data mining method for cleaning method [10][12]. Before using model selection process we are used preprocessing method for reduce the null values then recover the data with help of train/test split with help of MinMaxScaler [5].

MinMaxScalar, for each value in every feature MinMaxScalar cipher the minimum value within the feature then divided by the vary. The range is the distinction between the first most and original minimum. It preserves the shapes of the first original distribution.

```
(Loan_ID               0
 Gender               13
 Married               3
 Dependents           15
 Education             0
 Self_Employed        32
 ApplicantIncome       0
 CoapplicantIncome     0
 LoanAmount           22
 Loan_Amount_Term     14
 Credit_History       50
 Property_Area         0
 Loan_Status           0
 dtype: int64,
 Male      489
 Female    112
 Name: Gender, dtype: int64,
 Yes     398
 No      213
 Name: Married, dtype: int64,
 No      500
 Yes      82
 Name: Self_Employed, dtype: int64,
 1.0     475
 0.0      89
 Name: Credit_History, dtype: int64)
```

c. Feature Engineering

It is the method of using domain data to extract options from data via data processing techniques. These features are wanted to improve the performance of machine learning algorithms. Feature engineering is thought-about as applied Machine learning itself. It is helping for import the models.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

d. Machine Learning Methods

Machine learning is a subset of AI that trains machines with vast volumes of data to think and act like humans without being explicitly programmed. In this paper we are using supervised (Classification methods) methods

Five machine learning classification models have been used for prediction of android applications. The models are available in python open source software. The brief details of each model are described below.

Decision Trees

The basic algorithmic rule of call tree needs all attributes or options ought to be discredited. Feature choice relies on greatest info gain of options.

The data pictured in call tree will delineate within the kind of IF-THEN rules. This model is associate degree extension of C4.5 classification algorithms represented by Quinlan.

Random Forest

Random forests are a classifying learning framework for characterization (and backslide) that work by building a very large number of Decision trees at planning time and yielding the class that's the mode of the classes surrender by individual trees.

Support Vector Machine

Used SVM to build and train a model prepare a demonstrate utilizing human cell records, and classify cells to whether the tests are benign (mild state) or dangerous (evil state).

Support vector machines are managed learning models that utilize affiliation R-learning calculation which analyze attributes and distinguished design information, utilized for application classification. SVM can beneficially perform a replace utilizing the kernel trick, verifiably mapping their inputs into high dimensional attribute spaces [8].

Logistic Regression

Logistic regression is supervised learning classification algorithm (try to method connections and conditions between the target prediction output and input attributes) such that we are able to anticipate the yield values for new information based on those connections which it learned from the previous information sets [8][6].

K-nearest neighbor (KNN)

The KNN algorithm is a simple supervised machine learning algorithm that can be utilized to unravel both classification and replace issues. It is easy to implement and understand but significantly slows as the size of that data on use grows [5].

$$d' = \frac{d - \min(p)}{\max(p) - \min(p)}$$

5. EXPERIMENT AND RESULTS

A. Experiment overview:

In this experiment firstly collect the data and understand the data with help of (.describe()) and then analyses of data then search for any missing/null/nosy data present in the dataset and then evaluate the confusion matrices(accuracy, precision, recall, f1-score) and finally model building i.e., used methods Procedures are designed to detect errors in data at a lower level of detail. *Data validations* have been included in the system in almost every area where there is a possibility for the user to commit errors. The system won't accept invalid information. Whenever invalid information is keyed in, the system like a shot prompts the user and also the user should once more key within the information and also the system will accept for the info provided that the info is correct. Validations are enclosed wherever necessary.

The system is designed to be a user friendly one. In alternative words the system has been designed to speak effectively with the user. The system has been designed with popup menus.

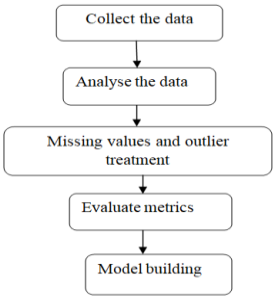


Fig (A): overview of experiment

B. Major Attributes:

In the below map shows the positive and negative values of attributes and heat map helps us to analyze the data dependent attributes. Loan Amount shows in after log form used.

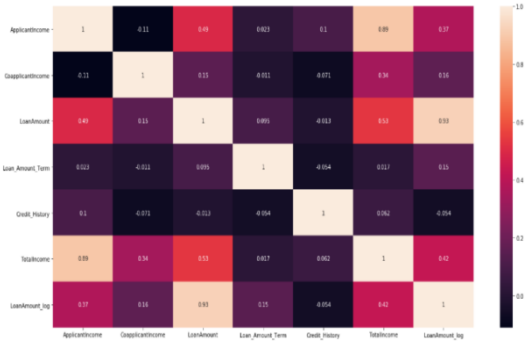
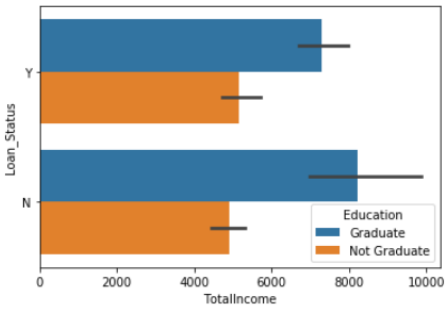


Fig (B): Heat map

C. Barplot():

A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle and provides some indication of the uncertainty around that estimate using error bars.

sns.barplot(x=data_train['TotalIncome'],y=data_train['Loan_Status'],hue=data_train[' Education'])



D. Pd.crosstab():

Compute a basic cross organization of two (or more) components. By default computes a recurrence table of the components unless an cluster of values and an accumulation work is passed.

Loan_Status	N	Y	All
Credit_History			
0.0	82	7	89
1.0	97	378	475
All	179	385	564

RESULTS:

Here shows all the methods we build and these methods are evaluate the accuracy, precision, recall, F1-score. And the below table represents the value obtained for the various metrics from the different methods. Here we choose the accuracy so, all methods comparatively SVM is the less accuracy. Therefore we can summarize that random forest is doing prediction well for our data.

Classification Results

Used Algorithms	Accuracy	Precision	Recall	F1-score
Random forest	82%	0.84	0.82	0.81
Logistic regression	73%	0.73	0.74	0.73
Decision tree	72%	0.72	0.72	0.72
KNN	59%	0.52	0.59	0.53
SVM	78%	0.82	0.78	0.75

Fig (i): Results

V. EVALUTION MODELS

Need for confusion matrix:

Classification (predict category) models have multiple output categories. Most error measures will tell us the total error in our model but we cannot use it to find out individual instances of errors in our model. Confusion matrix helps us identify the correct predictions of a model for different individual classes as well as the errors. The main matrix:

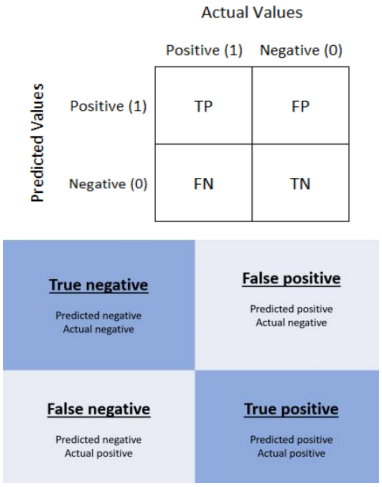


Fig (a): Confusion matrix

- **Accuracy:** It's worn to find the portion of correctly classified values. It is tell us how often our classifier is right. Sum of all true values divided by total values.

Number of classified samples =TP+TN
Total number of samples= TP+FP+TN+FN

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Fig (b): Accuracy

- **Precision:** It is used to calculate the models ability to classify positive values correctly.

Number of classified values = TP
Number of actual values = TP+FP

$$\textit{Precision} = \frac{TP}{TP + FP}$$

Fig (c): Precision

- **Recall:** To calculate the models ability to predict positive values

$$\textit{Recall} = \frac{TP}{TP + FN}$$

Fig (d): Recall

- **F1-** Score: It is also called the F score or the F Measure. Put another way, the f1 score conveys the balance between the precision and the recall.

$$F_1 = \frac{2}{\frac{1}{\text{recall}} \times \frac{1}{\text{precision}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Fig (e): F1-Score

Note: precision and recall are exactly helps to define problem of group of predicted vales.

VI. CONCLUSION

In this paper, we have proposed customer loan prediction using supervised learning techniques for loan candidate as a valid or fail to pay customer. In this paper, various algorithms were implemented to predict customer loan. Optimum results were obtained using Logistic Regression, Random Forest, KNN, and SVM, decision Tree Classifier. Compare these five algorithms random forest is the high accuracy. From a correct analysis of positive points and constraints on the part, it can be safely ended that the merchandise could be an extremely efficient part. This application is functioning properly and meeting to all or any Banker necessities. This part is often simply obstructed in several different systems. There are numbers cases of computer glitches, errors in content and most significant weight of option is mounted in machine-driven prediction system, therefore within the close of future the therefore called software system might be created more secure, reliable and dynamic weight adjustment. In close to future this module of prediction can be integrated with the module of machine-driven processing system.

VII. FUTURE SCOPE

The system is trained on old training dataset in future software can be made such that new testing data should also take part in training data after some fix time.

REFERENCES

[1] Yu Jin and Yudan Zhu, "A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P) lending," School of Information, Zhejiang University of Finance and Economics, 310018 Hangzhou, China.

[2] https://www.kaggle.com/telco-churn

[3] Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswal "Loan default forecasting using data mining" Department of Information Technology, St. Francis Institute of Technology, Mumbai, India (2020)

[4] Octave Iradukunda, Haiying Che, Josiane Uwineza, Jean Yves Bayingana, Muhammad S Bin-Imam, Ibrahim Niyonzima "Malaria Disease Prediction Based on Machine Learning" School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China (2019).

[5] G. Arutjothi, Dr. C. Senthamarai "Prediction of Loan Status in Commercial Bank using Machine Learning Classifier" department of computer applications government arts college (Autonomous) Salem, India (2017.)

[6] Mohammad Ahmad Sheikh, Amit Kumar Goel, Tapas Kumar "An Approach for Prediction of Loan Approval using Machine Learning Algorithm" School Of Computer Science And Engineering Galgotias University Greater Noida, India (2019).

[7] Xin Li, Xianzhong Long, Guozi Sun, Geng Yang, and Huakang Li "Overdue Prediction of Bank Loans Based on LSTM-SVM"Jiangsu Key Lab of Big Data and Security and Intelligent Processing Nanjing University of Posts and Telecommunications, Nanjing, 210023, China.

[8] Aakanksha, Tamara Denning, Vivek Srikumar, Sneha Kumar Kesera "secrets in source code: reducing false positives using ML" software engineering (Microsoft) school of computing, USA (2020)

[9] Arutjothi .G, Dr. C. Senthamarai. "Credit Risk Evaluation using Hybrid Feature Selection Method. Software engineering and technology (2017)

[10] Ch. Balayesu and S Narayana, "An Improved Algorithm for Efficient Mining of Frequent Item Sets on Large Uncertain Databases" in International Journal of Computer Applications, Volume 73, No. 12 July 2013, Page No. 8-15.

[11] Bala brahmeswara kadaru et al."A novel ensemble decision tree classifier using hybrid feature selection measures for parkinson's disease prediction", Int. J. Data science (IJDS), ISSN: 2053-082X, Vol.3, No.4,2018.

[12] Mrunal Surve, Pooja Thitme, Priya Shinde, Swati Sonawane, and Sandip Pandit. "Data mining techniques to analyze risk giving loan (bank)" Internation Journal of Advance Research and Innovative Ideas in Education Volume 2 Issue 1 2016 Page 485-490

AUTHORS

First Author – L. Udaya Bhanu, M.Tech Student, Dept. of Computer Science & Engineering, Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India
Second Author – Dr. S. Narayana, Professor&Mentor, Dept. of Computer Science & Engineering, Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 3, May-June-2021, ISSN (Online): 2395-566X

A Comparative Study on Loan Eligibility

Akash Dagar
Department of Information Technology,
Maharaja Agrasen Institute of Technology,
South East Delhi, India.

***Abstract*- Nowadays, machine learning algorithms are used everywhere around us. In near future our lives would be eased out by using artificial intelligence. We all know checking if a person is eligible for loan or not is a very complex and time consuming process. And it's a great source of income for banks. In recent years, many experts worked on loan approval prediction. Machine Learning algorithms play a very crucial role in prediction. In this paper four machine algorithms, Logistic Regression, Random Forest, Support Vector Machine, XGBoost are used for predicting if a person is eligible for a loan or not. On the basis of dataset used, we found that Logistic Regression provides better accuracy than other models.**

Keywords- Include at least 4 keywords or phrases, must be separated by commas to distinguish them.

I. INTRODUCTION

Distribution of the loans is the core business part of almost every banks. The main portion the bank's assets is directly came from the profit earned from the loans distributed by the banks. The prime objective in banking environment is to invest their assets in safe hands where it is.

Today many banks/financial companies approves loan after a regress process of verification and validation of set of criterions. This time consuming process can be avoided by developing a system which predicts whether the person is eligible for the loan. Through this system we can predict whether that particular applicant is safe or not and the whole process of validation of features is automated by machine learning technique. The disadvantage of this model is that it emphasize different weights to each factor but in real life sometime loan can be approved on the basis of single strong factor only, which is not possible through this system.

Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this paper is to provide quick, immediate and easy way to choose the deserving applicants. It can provide some special advantages to the bank.

II. LITERATURE SURVEY

Literature survey is the most important step in any kind of research. Before start developing we need to study the previous papers of our domain which we are working and on the basis of study we can predict or generate the drawback and start working with the reference of previous papers.

Amira Kamil Ibrahim Hassan, Ajith Abraham (2008) [1] uses a prediction model which is constructed using three different training algorithms to train a supervised twolayer feed-forward network. The results show that the training algorithm improves the design of loan default prediction model.

Angelini (2008) [2] used a neural network with standard topology and a feed-forward neural network with ad hoc connections. Neural network can be used for prediction model. This paper shows that the above two models give optimum results with less error.

Sarwesh Site, Dr. Sadhna K.Mishra (2013) [3] proposed a method in which two or more classifiers are combined together to produce an ensemble model for the better prediction. They used the bagging and boosting techniques and then used random forest technique.

Akkoç (2012) [4] used a model namely hybrid Adaptive Neuro-Fuzzy Inference model, grouping of statistics and Neuro-Fuzzy network. A 10-fold cross validation is used for better results and a comparison with other models.

Alaraj M, Abbod M (2015) [5] introduced a model that are based on homogenous and heterogeneous classifiers. Ensemble model based on three classifiers that are logistic artificial neural network, logistic regression and support vector machine.

III. METHODOLOGY

The methodology will go according to the following flowchart.

1. Collection of Data:
Data is the heart of machine learning. Predictive models use data for training which gives somewhat accurate results. Without data we can't train the model. Machine learning involves building these models from data and uses them to predict new data. Machine Learning is a subset of Artificial Intelligence. It gives system capability

to learn wherein it automatically learns and improves its performance without being explicitly programmed. The dataset used here is taken from Kaggle.

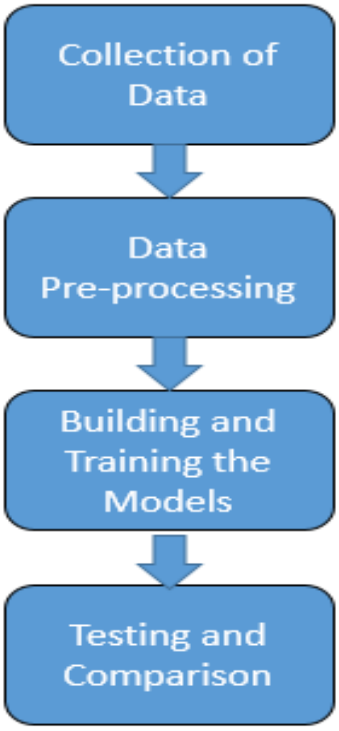


Fig 1. Methodology.

2. Data Pre-processing:
Data pre-processing is the process of cleaning our data set. Data is pre-processed in different phases. In first phase, the null values are filled by using the traditional mean and mode method. In second phase, the data visualization is done by plotting different graphs between the attributes.

This provided various intuition about the data. Then the Log Transformation is needed for some attributes in the dataset. In the third phase, a new feature Total_Income is engineered using the other features of the dataset and the correlation between the attributes is found using the heatmap. In the last phase, the categorical attributes are taken care of by using the Label Encoding Technique.

3. Building and Training the model:
The dataset was split into training and testing set with 80% in training set and remaining 20% data in test set.

IV. MODELS WERE USED

1. Logistic Regression:
It's a classification algorithm, it is used to classify the inputs into different classes. It should only be used when the target variables fall into discrete categories. It basically works according to a threshold, if the value crosses the threshold then it should be put in one class otherwise the other.

2. Random Forest Classifier:
Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression.

Random forest is a bagging technique and not a boosting technique. The trees in random forests are run in parallel. There is no interaction between these trees while building the trees.
It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

3. XGBoost Classifier:
XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

4. Support Vector Machine:
The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen.

Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Scikit-learn library was used for importing the models. Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. The vision for the library is a level of robustness and support required for use in production systems. This means a deep focus on concerns such as easy of use, code quality, collaboration, documentation and performance.

K-fold Cross Validation is used for training and testing the models accurately. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it

may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model. It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

5. Testing and Comparison:

Testing and comparison are done using the K-fold Cross Validation and Confusion Matrix. A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.

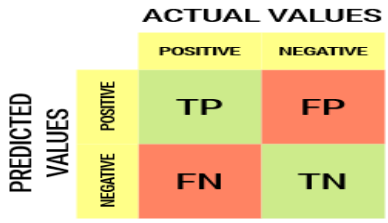


Fig 2. Confusion Matrix.

IV. RESULTS

With the help of K-Fold Cross Validation we were able to compare the models and two models were having almost same accuracy. Following table contains the result.

Table: Accuracy for different models.

Models	K-Fold Accuracy
Logistic Regression	80.9462881514061
Random Forest	78.5032653605224
Support Vector Machine	69.7054511528721
XGBoost	79.1576702652272

1. Heat Map and Classification Metrics for Logistic Regression:

	precision	recall	f1-score	support
0	0.92	0.41	0.56	54
1	0.75	0.98	0.85	100
accuracy			0.78	154
macro avg	0.84	0.69	0.71	154
weighted avg	0.81	0.78	0.75	154

Fig 3.Classification Report for Logistic Regression Model.

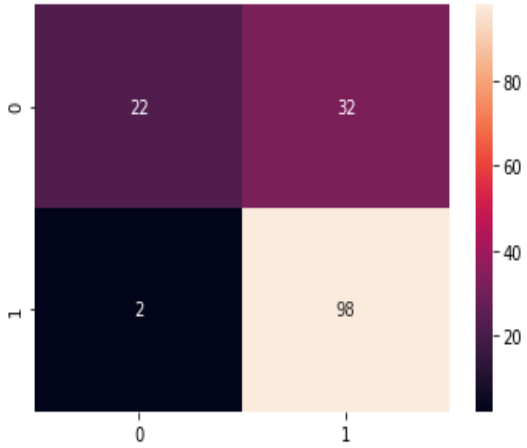


Fig 4. HeatMap for Logistic Regression Model.

V. CONCLUSION

In this research paper, we have used machine learning algorithms to predict the eligibility of an applicant for loan. The data is pre-processed and fed to different regression models to determine the best model and with the help of both K-Fold Cross-Validation and different classification metrics we compared different model.

So, according to above results we came to know that Logistic Regression is the most effective model with maximum accuracy and can be used as an effective model for predicting weather an applicant is eligible for loan or not, which should help banks to skip the tedious process of loan eligibility.

REFERENCES

[1] Amira Kamil Ibrahim Hassan and Ajith Abraham, "Modeling Consumer Loan Default Prediction Using Ensemble Neural Networks", International Conference on Computing, Electrical and Electronics Engineering , pp. 719 – 724, August 2013.

[2] E. Angelini, A. Roli, and G. di Tollo, "A neural network approach for credit risk evaluation" elsevier, The Quarterly Review of Economics and Finance, Vol. 48, Issue 4, pp. 733–755, November 2008.

[3] Sarwesh Site, Dr. Sadhna K. Mishra, " A Review of Ensemble Technique for Improving Majority Voting for Classifier", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 1, pp. 177- 180,January 2013.

[4] S. Akkoç, "An empirical comparison of conventional techniques, neural networks and the three stage hybrFId Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis : The case of Turkish credit card data", Elsevier Europezan

Journal of Operational Research, Vol. 222, Issue 1, pp. 168–178, October 2012.

[5] Maher Ala'raj and Maysam Abbod,"A systematic credit scoring model based on heterogeneous classifier ensembles", Innovations in Intelligent Systems and Applications (INISTA), pp. 1-7, September 2015

[6] M. V. Jagannatha Reddy and B. Kavitha, " Extracting Prediction Rules for Loan Default Using Neural Networks through Attribute Relevance Analysis", International Journal of Computer Theory and Engineering, Vol. 2, Issue 4, pp. 596-601, August 2010.

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

Loan Prediction using Decision Tree and Random Forest

Kshitiz Gautam[1], Arun Pratap Singh[2], Keshav Tyagi[3], Mr. Suresh Kumar[4]

[1-3]BTech student, Dept. of IT, Galgotias College of Engineering and Technology, Greater Noida, U.P
[4]Assistant Professor, Dept. of IT, Galgotias College of Engineering and Technology, Greater Noida, U.P

Abstract - *In India, the number of people or organization applying for loan gets increased every year. The bank employees have to put in a lot of work to analyse or predict whether the customer can pay back the loan amount or not (defaulter or non-defaulter) in the given time. The aim of this paper is to find the nature or background or credibility of the client that is applying for the loan. We use exploratory data analysis technique to deal with the problem of approving or rejecting the loan request or in short loan prediction. The main focus of this paper is to determine whether the loan given to a particular person or an organization shall be approved or not.*

Key Words: Loan, Prediction, Machine Learning, Training, Testing.

1. INTRODUCTION

The term banking can be referred to as receiving and protecting money that is deposited by an individual or an entity. It also includes lending money to people and businesses which has to be paid back within the given amount of time without failing. Banking is a sector that is regulated in most of the countries as it is an important factor in determining the financial stability of the country. The prime goal in banking sector is to invest their assets in safe hands where there are less chances of failure. Today many banks and financial companies approve loan after a stressful, long and weary process of verification but still there is no surety whether the chosen applicant is credible or not or in other words if he is able to return the amount with interest in the given time. The purpose of the loan can be anything based on the customer needs. Loans are broadly divided as open ended and close-ended loans.

Examples of open-end loans are credit cards and a home equity line of credit (HELOC).

Close-ended loans decreases with each payment. It means the amount is reduced after an instalment.

In other words, it is a legal term that cannot be modified by the borrower. Personal loans, mortgages, auto payments, EMI and student loans are the most common examples of close-ended loans.

Secured or collateral loan are those loans that are protected by an asset. Houses, Vehicles, Savings accounts are the personal properties used to secure the loan.

2. DATA SET

A collection of data is taken from the banking sector. The Data set is in ARFF (Attribute-Relation File Format) format that is acceptable by Weka. ARFF file is composed of tags that include the name, types of attributes, values and data itself. For this paper we are using 12 attributes like gender, marital status, qualification, income, etc.

The table below represents the data set that we have used:

Table-1: Data set variables along with description and type

Variable Name	Description	Type
Loan_ID	Unique ID	Integer
Gender	Male/Female	Character
Marital_Status	Applicant married(Y/N)	Character
Dependents	Number of Dependents	Integer
Education_Qualification	Graduate/Under Gradute	String
Self_Employed	Self-employed(Y/N)	Character
Applicant_Income	Applicant income	Integer
Co_Applicant_Income	Co-applicant income	Integer
Loan_Amount	Loan amount in thousands	Integer
Loan_Amount_Term	Term of loan in months	Integer
Credit_History	Credit history meets guidelines	Integer
Property_Area	Urban/Semi urban/Rural	String
Loan_Status	Loan Approved(Y/N)	Character

Now in machine learning model, we first apply the training data set, in this data set the model is trained with known examples. The entries of new applicants will act as a test data which are to be filled at the time of submitting the application. After performing such tests, model can determine whether the loan approved to the person is safe or not basically about the loan approval on the basis of the various training data sets.

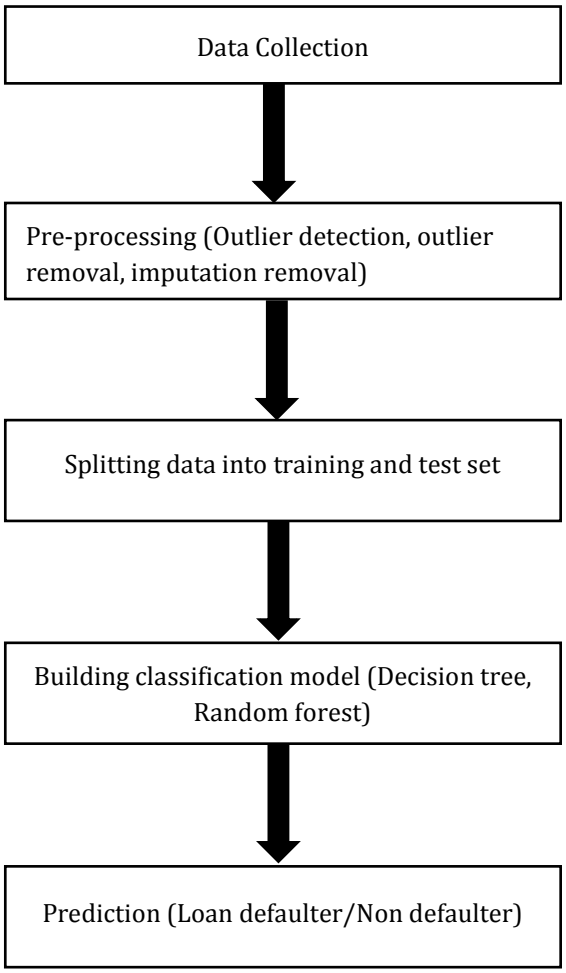


Fig-1: Chronology of Data

The diagram above gives us an outline on how data is used in this machine learning process or model.

Basically, it is divided into four parts in which we use data to predict the outcome of the whole process. First, we use training data set to train our model. After the model is trained, then we test it with unknown examples from the same scenario.

Another process that we use before testing and training data is data pre-processing. In data pre-processing we remove all sorts of values that can cause an error like redundant values, incomplete values, missing data, etc.

3. LOAN PREDICTION METHODOLOGY

The diagram 2 represents the working of our model. It basically gives us a rough idea on how the loan prediction system works. After collecting data, we use feature selection process on data. Feature selection can be defined as a process of reducing number of input variables when we develop a predictive model.

Feature selection is divided into two parts i.e. supervised method and unsupervised method. Supervised method is divided into three parts which are wrapper, filter and intrinsic. In supervised method we use target variable to remove discrepancies in data. While in unsupervised method we do not use target variable to remove discrepancies. Unsupervised method uses the process of correlation.

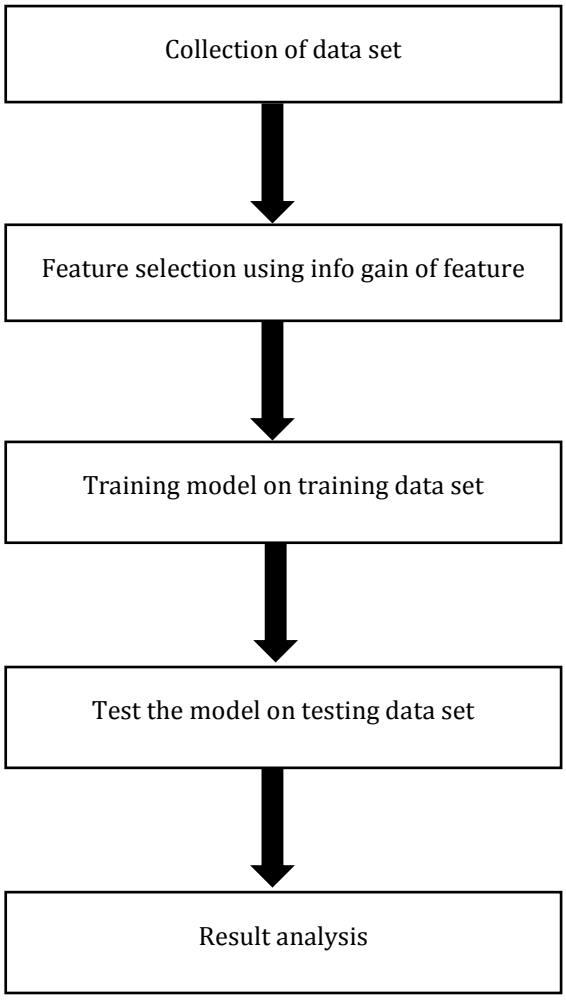


Fig-2: Loan Prediction Methodology

4. WORKING OF THE MODEL

We have represented the working of the model through a use case diagram. The figure below represents the attributes, process of the model that we have built.

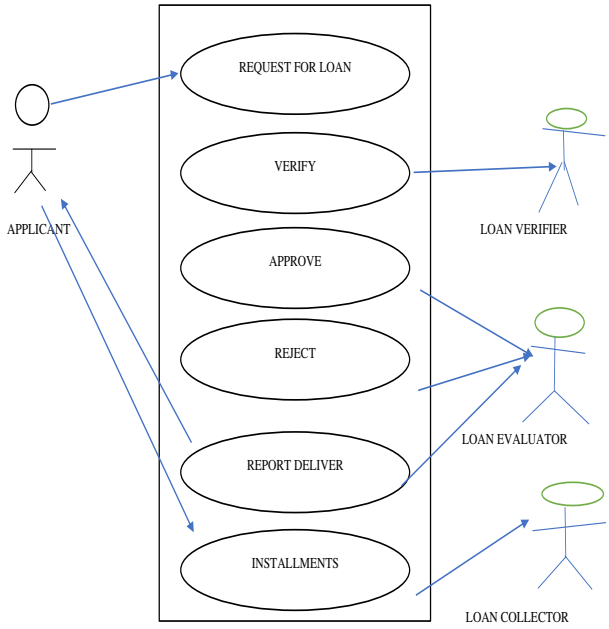


Fig-3: Use case diagram

Table-2: Use case diagram variable and description

Actor	Applicant, Loan Verifier, Loan Evaluator, Loan Collector
Description	An applicant requests for a loan. After request loan verifier verified its document and transfer to loan evaluator may approve or reject the loan.
Data	Applicant personal information and its documents.
Stimulus	User command issue by online loan and application.
Response	Loan may be approved or may be rejected.
Comments	Improve installment policy.

5. EXPLORATORY DATA ANALYSIS

1. The one whose salary is more can have a greater chance of loan approval.
2. The one who is graduate has a better chance of loan approval.
3. Married people would have an upper hand than unmarried people for loan approval.
4. The applicant who has a smaller number of dependents have a high probability for loan approval.
5. The lesser the loan amounts the higher the chance for getting loan.

6. Model used for training and testing

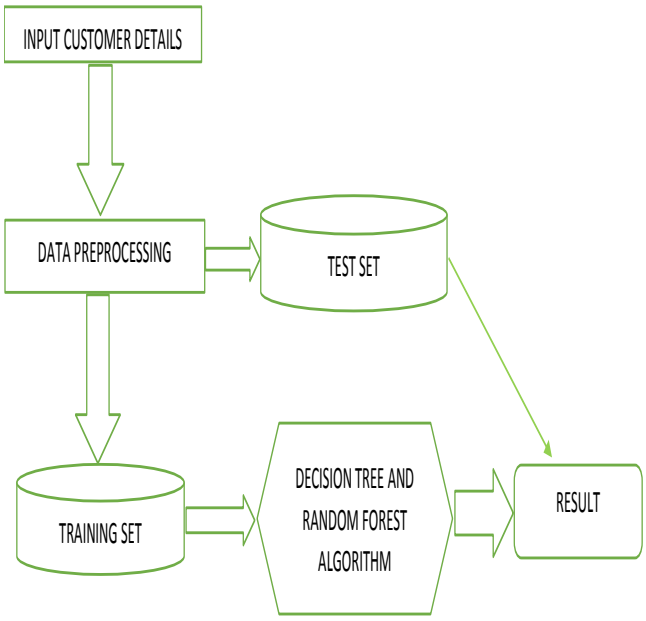


Fig- 4: Training and testing model

7. MACHINE LEARNING METHODS

Two machine learning classification models are used for the prediction of application that can be used in android applications. These models can also be accessed in the open source software R, which is licensed under GNU GPL. The brief description of each model is explained below.

7.1 Decision tree

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin toss comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

This model is an extension of C4.5 classification algorithms. We experimented with J48 Decision Tree classifier which is an implementation of C4.5 Decision Tree. In case of this classifier, the lower the confidence factor, the more pruning is done. For this we have used different confidence factors and analysed them with higher confidence factor and with the increase of confidence factor the accuracy has increased in each case. With the confidence factor of 0.15 the best accuracy is 62.12% and with a confidence factor of 0.25 it is 63.39%. It means that when less pruning is done the accuracy improves.

7.2 Random forest

Random forest or random decision forests are an ensemble learning method used for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

We have done several trials with Random Forest with different parameters: executions with supervised and unsupervised discretization's (equal-frequency and equal-width), with all attributes. In the experiments without attribute selection the best result was 85.75% and it was achieved with unsupervised equal-frequency 5 bins discretization with 450 trees and seed equal to 4.

Table-3: Parameter setting for machine learning models

Model	Parameter Setting
Decision Tree	Min Split=20, Max Depth=30, Min Bucket=7
Random Forest	Number of trees=450, number of variables=8

8. CONCLUSIONS

The main purpose of the paper is to classify and analyze the nature of the loan applicants. From a proper analysis of available data and constraints of the banking sector, it can be concluded that by keeping safety in mind that this product is much effective or highly efficient. This application is operating efficiently and fulfilling all the major requirements of Banker. Although the application is flexible with various systems and can be plugged effectively.

This paper work can be extended to higher level in future so the software could have some better changes to make it more reliable, secure, and accurate. Thus, the system is trained with a present data sets which may be older in future so it can also take part in new testing to be made such as to pass new test cases.

There have been numbers cases of computer glitches, errors in content and most important weight of features is fixed in automated prediction system. So, in the near future the so – called software could be made more secure, reliable and dynamic weight adjustment. In near future this module of prediction can be integrated with the module of automated processing system.

REFERENCES

[1] J. R. Quinlan. Induction of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086.

[2] A. Goyal and R. Kaur, "A survey on Ensemble Model for Loan Prediction", International Journal of Engineering Trends and Applications (IJETA), vol. 3(1), pp. 32-37, 2016.

[3] G. Shaath, "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R".

[4] A. Goyal and R. Kaur, "Accuracy Prediction for Loan Risk Using Machine Learning Models".

[5] Hsieh, N. C., & Hung, L. P. (2010). A data driven ensemble classifier for credit scoring analysis. Expert systems with Applications, 37(1), 534-545.

[6] https://en.wikipedia.org/wiki/Exploratory_data_analysis

[7] https://www.experian.com/blogs/ask-experian/credit-education/score- basics/what-is-a-good-credit-score/

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

MACHINE LEARNING TECHNIQUES FOR RECOGNIZING THE LOAN ELIGIBILITY

Mr. Abhiroop Sarkar*1

*1Bachelors Of Technology, Artificial Intelligence, G H Raisoni College Of Engineering, Maharashtra, India.

ABSTRACT

Loan can be considered to be a debt incurred by a person or an organization. Loans are usually lend by any single candidate/organization to another such party. The person who borrows the money has to agree to certain conditions like interest, extra charges etc. This study aims the prediction of whether a person is approved for being sanctioned a loan or not. There were many parameters like marital status, credit-history, gender etc. that has been considered for processing and analysis of Loan eligibility. Manually, analysis of Loan prediction become time consuming and costly, so this study has been performed to find the best algorithm which can automate the process to facilitate the Banker staff as well as customer to receive the eligibility analysis on immediate basis. The dataset has been splitted into training set and testing set where train used for training the algorithms upon which the test data has been used to make the predictions over the recognized entity.

Keywords: Loan-Eligibility, Machine Learning, Random Forest, Logistic Regression.

I. INTRODUCTION

A bank provides a customer with many services like safety of their money, interest options, quick withdrawals and other such benefits. The bank has various sources of income to provide the afore mentioned functionalities [1], still the main source of income stays on their credit lines. So, the interest gathered on the loan affects their profit the most. Hence whether a customer will repay or not the loan is important for the bank. Hence the loan is only given to the customers who are eligible to repay [2]. A loan is led by an organizations or other entities to people or other organizations. The borrower acts upon a debt for which he or she has to take authority to pay the interest until the loan is completely returned along with the original amount borrowed. Sanctioning a loan is one of the significant functions of a banking sector. Banks apply interests on loans which are then sanctioned to the customers.

Predicting loan eligibility helps both the bank employees as well as the customers. Here the paper is trying to provide quick and easy way to choose the deserving candidates. The system can calculate on its own the weight of each variable, which is a part of the loan process and test it. This checking will be done privately hence there will be no stakeholders who can alter the results so the applicable candidates will have higher priority to be sanctioned a loan [3] [4]. Machine learning algorithms enable the construction of a new model using previously unknown historical data that can be used to train the model to make better predictions not only for credit risks, but also for other risks such as early payment opportunities leading to loss of income from interest, existing withdrawal risks etc. [5].

A Prediction Model works on or operate on data analysis, statistics and probability to predict an outcome. Every model works on few variables which are likely to come in handy for future results. A statistical model is created based on the data collected from various resources. We can use simple machine learning algorithms or even a complex software for our project. If more data is used then the model becomes more better and so the errors are reduced meaning then the model will be able to predict with higher accuracy and even take less time [6]. In this paper we are considering logistic regression, random forest and decision tree machine learning algorithms for comparison. We will split the dataset into train and test classes and then predict the model using the three algorithms and find the best suited algorithm from among them [7].

In machine learning we split our complete dataset into training and testing dataset. We have used the splitting method like; class train_test_split() for achieving better result. There are always some issues that are faced with the random state parameter here we would get different accuracy for different random state thus not giving the perfect accuracy for our model, so we use stratified k fold cross validation for stratified sampling [8][9].

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

II. DATASET

Data collection is the process of collecting and measuring data in relation to targeted changes in an established system, which enables one to answer relevant questions and evaluate results. The purpose of all data collection is to obtain quality evidence that leads to analysis and constructs concrete and misleading answers to the questions presented [10]. The dataset has been divided into two categories the train dataset and the test dataset. The train dataset consists of six hundred and fourteen (614) rows and twenty-two (22) columns while the test dataset consists of three hundred and sixty-seven (367) rows and twenty (20) columns.

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y

Figure 1: (Training data entries)

```
[7] test=pd.read_csv('loan-test.csv')
    test.head()
```

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
0	LP001015	Male	Yes	0	Graduate	No	5720	0	110.0	360.0	1.0	Urban
1	LP001022	Male	Yes	1	Graduate	No	3076	1500	126.0	360.0	1.0	Urban
2	LP001031	Male	Yes	2	Graduate	No	5000	1800	208.0	360.0	1.0	Urban
3	LP001035	Male	Yes	2	Graduate	No	2340	2546	100.0	360.0	NaN	Urban
4	LP001051	Male	No	0	Not Graduate	No	3276	0	78.0	360.0	1.0	Urban

Figure 2: (Testing data entries)

The dataset consists of the parameters: gender, marital status, education, income credit history etc. From the data set we can infer that the applications from the 'male' gender is more than the counterpart and also that most of the applicants are married.

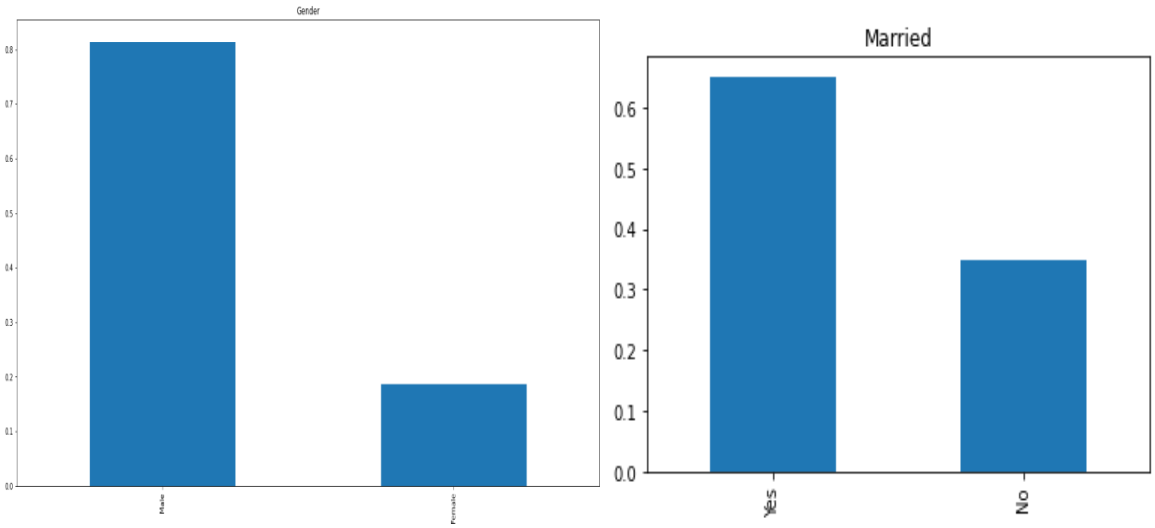


Figure 3: Gender & Marital status

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

The heat map for the following dataset is here it will show which variables are more related for the applicant to receive the loan and thus only those factors would be considered when the final model is being created.

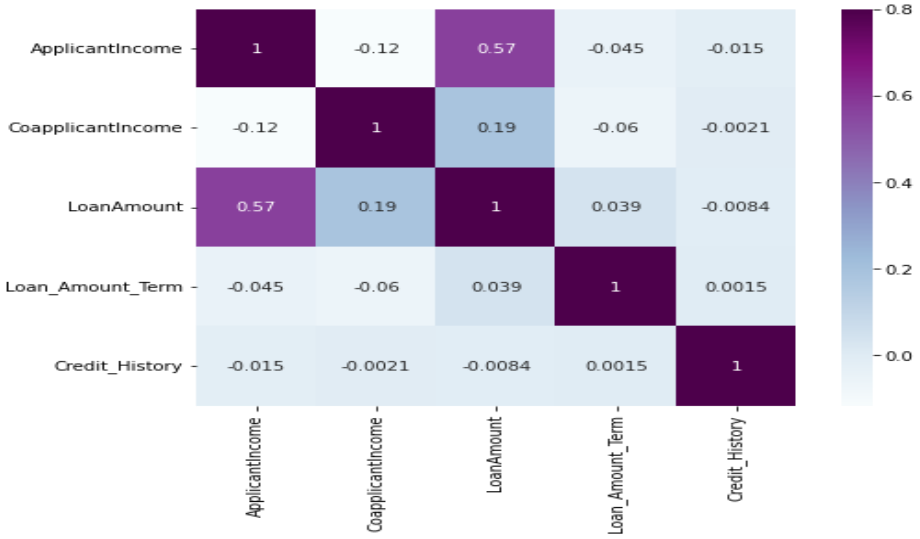


Figure 4: Heatmap between featured metrics

III. METHODOLOGY

For the model training we will compare the mean accuracy score of three machine learning algorithms: logistic regression, decision tree, random forest.

3.1. Logistic regression

It is a widely used Machine Learning algorithms, which belongs to the category of supervised learning. It is used to predict the dependent value based on a set of independent variables. Its primary function is the prediction of the dependent value. This value is usually categorical so the output has been a categorical or a discrete value. It can be either Yes or No, 0 or 1 etc., instead of giving the value as numeric form i.e., either 0 or 1, it gives a probabilistic value that lie between 0 and 1[11]. In this algorithm, instead of making a regression line, we make an 'S' shaped logistic function, which predicts two maximum possible values (0 or 1). The curve from the logistic function shows the likelihood of something like whether the cells are infectious or not, a cat is fat or not based on various different features [12][13]. Logistic Regression is quite an important machine learning technique as it has the power to give probabilities and classify new data from continuous and discrete datasets.

The equation for the straight line can be shown as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

As we know that in logistic regression y can be between 0 and 1 only, so we will divide the above equation by (1-y):

$$\frac{y}{1-y}\text{; 0 for y= 0, and infinity for y=1}$$

But the range has to be between - [infinity] to + [infinity], taking logarithm of the equation we will get:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

This can be considered to be the final mathematical representation for logistics regression. Logistic Regression can be broadly divided into three categories:

Binomial: In binomial logistic regression, there are only two possible types of the dependent variables, i.e., it is either 0 or 1.

Multinomial: In multinomial logistic regression, there are 3 or more possible unordered types of the dependent variable, such as 'horses', 'zebras', or 'dogs'.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

Ordinal: In ordinal logistic regression, there are 3 or more possible ordered types of dependent variables, such as 'low', 'medium' or 'high'.

3.2. Decision tree

Decision tree is one of the supervised learning techniques that is used for classification as well as regression problems though it is preferred to solve classification problems. It is a classifier, where the internal nodes show the features of a dataset, branches are used for decision rules and each leaf node represents outcome. It is a tree structured classifier. In a Decision tree, there are two nodes known as the decision node and the leaf node [14]. Decision nodes are used in making a decision and have multiple branches, on the other hand leaf nodes are the output based on the decision and do not contain any branches. The decisions depends upon the features present in the given dataset.

It is a graphical representation for getting all the possible solutions to a problem depending upon a few valid conditions. It is called a decision tree as, like in a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. The CART algorithm which stands for Classification and Regression Tree algorithm is used for building the tree. The attribute selection is one of the issues that arises when a decision tree is being implemented. Attribute selection measures is a technique used to solve such problems. Using this technique, we can easily select the best attribute for the tree nodes [15] [16].

3.3. Information Gain:

Information gain is the measurement of changes in entropy after the dataset is segmented. It is used to find out how much information a feature can give us about a particular class. We use the value of information gain, to split the node and create the decision tree. A decision tree algorithm tries to reach the maximum possible value of information gain, and a node having the highest information gain is split first.

It can be calculated using the below formula:

Information Gain= Entropy(S) - [(Weighted Avg) *Entropy (each feature)]

Entropy: Entropy is used to measure the impurity in an attribute. It specifies the randomness present in the data. Entropy can be calculated as:

Entropy(S) = -P (yes) log2 P (yes) - P (no) log2 P (no)

Where,

S= Total number of samples

P (yes) = probability of yes

P (no) = probability of no

3.4. Random forest

Random Forest is a machine learning algorithm that is a part of the supervised learning technique. It can be used for classification as well as regression problems. It is derived on the basis of ensemble learning, which is the method of combining multiple classifiers to solve a complex problem and to increase the performance of the model. Random forest contains many individual decision trees that work as an ensemble. Each tree in the forest gives out a prediction and the class that has the maximum number of outcomes becomes the model's prediction [17]. A large number of relatively uncorrelated models operating as a group will give better performance than any of the individual constituent models. There is a low correlation between the models. Uncorrelated models can produce ensemble predictions which are more accurate than any of the individual predictions. The prerequisites for random forest for performing with better results are:

- There has to be actual signal in the features so that the models built using those features do better than random guessing.
- The predictions (and errors) generated by the individual trees should have low correlations with each other.
- The greater the number of trees in the forest, the higher is the accuracy and thus it prevents the situation of overfitting [18].
- Random forest is considered to be a better algorithm because it takes less training time than other algorithms and it increases the accuracy even for large datasets [19].

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

IV. RESULT ANALYSIS

We have tried to predict whether an applicant is applicable for a loan or not using three machine learning algorithms below are the observations for all of them.

4.1. Logistic regression

Logistic Regression is a classification algorithm used to assign observation to a discrete set of class. It is one among the machine learning algorithm which is used for classification problem, it is a predictive analysis algorithm based on the concept of probability.

```
i=1
mean = 0
kf = StratifiedKFold(n_splits=5,random_state=1,shuffle=True)
for train_index,test_index in kf.split(X,y):
 print ('\n{} of kfold {} '.format(i,kf.n_splits))
 xtr,xvl = X.loc[train_index],X.loc[test_index]
 ytr,yvl = y[train_index],y[test_index]
 model = LogisticRegression(random_state=1)
 model.fit(xtr,ytr)
 pred_test=model.predict(xvl)
 score=accuracy_score(yvl,pred_test)
 mean += score
 print ('accuracy_score',score)
 i+=1
 pred_test = model.predict(test)
 pred = model.predict_proba(xvl)[:,1]
print ('\n Mean Validation Accuracy',mean/(i-1))
```

```
1 of kfold 5
accuracy_score 0.8048780487804879

2 of kfold 5
accuracy_score 0.8373983739837398

3 of kfold 5
accuracy_score 0.7967479674796748

4 of kfold 5
accuracy_score 0.7967479674796748

5 of kfold 5
accuracy_score 0.8032786885245902

 Mean Validation Accuracy 0.8078102092496335
```

Figure 5: Algo. for Logistic Regression

This shows the mean validation accuracy for logistic regression algorithm. We can see that it has come around 80.78%

4.2. Decision tree

In Decision Analysis, Decision Tree can be used to visually and explicitly represent decision and decision making. The aim of using this algorithm was to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules inferred from training prior data.

```
from sklearn import tree
i=1
mean = 0
kf = StratifiedKFold(n_splits=5,random_state=1,shuffle=True)
for train_index,test_index in kf.split(X,y):
    print ('\n{} of kfold {} '.format(i,kf.n_splits))
    xtr,xvl = X.loc[train_index],X.loc[test_index]
    ytr,yvl = y[train_index],y[test_index]
    model = tree.DecisionTreeClassifier(random_state=1)
    model.fit(xtr,ytr)
    pred_test=model.predict(xvl)
    score=accuracy_score(yvl,pred_test)
    mean += score
    print ('accuracy_score',score)
    i+=1
    pred_test = model.predict(test)
    pred = model.predict_proba(xvl)[:,1]
print ('\n Mean Validation Accuracy',mean/(i-1))
```

```
1 of kfold 5
accuracy_score 0.7073170731707317

2 of kfold 5
accuracy_score 0.6991869918699187

3 of kfold 5
accuracy_score 0.7154471544715447

4 of kfold 5
accuracy_score 0.7235772357723578

5 of kfold 5
accuracy_score 0.680327868852459

 Mean Validation Accuracy 0.7051712648274024
```

Figure 6: Algo. for Decision Tree

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

This shows the mean validation accuracy for decision tree algorithm. We can see that it has come around 70.51%

4.3. Random forest

Random Forest is an ensemble machine learning algorithm that is used for classification and regression problem. Random Forest applies the technique of bagging (bootstrap aggregating) to decision tree learner. The beginning of the random forest algorithm starts with randomly selected "k" features to find the root node by using the best split approach.

```
from sklearn.ensemble import RandomForestClassifier
i=1
mean = 0
kf = StratifiedKFold(n_splits=5,random_state=1,shuffle=True)
for train_index,test_index in kf.split(X,y):
 print ('\n{} of kfold {} '.format(i,kf.n_splits))
 xtr,xvl = X.loc[train_index],X.loc[test_index]
 ytr,yvl = y[train_index],y[test_index]
 model = RandomForestClassifier(random_state=1, max_depth=10)
 model.fit(xtr,ytr)
 pred_test=model.predict(xvl)
 score=accuracy_score(yvl,pred_test)
 mean += score
 print ('accuracy_score',score)
 i+=1
 pred_test = model.predict(test)
 pred = model.predict_proba(xvl)[:,1]
print ('\n Mean Validation Accuracy',mean/(i-1))
```

```
1 of kfold 5
accuracy_score 0.8048780487804879

2 of kfold 5
accuracy_score 0.8373983739837398

3 of kfold 5
accuracy_score 0.7886178861788617

4 of kfold 5
accuracy_score 0.8130081300813008

5 of kfold 5
accuracy_score 0.7459016393442623

 Mean Validation Accuracy 0.7979608156737305
```

Figure 7: Algo. for Random Forest

This shows the mean validation accuracy for random forest algorithm. We can see that it has come around 79.79%

V. CONCLUSION

We can easily conclude that logistic regression can be considered to be the best among the three machine learning algorithms with an accuracy of 80.78%, closely followed by random forest at 79.79% and finally by decision tree with 70.51%. This paper gives a general idea that we can prefer logistic regression for loan eligibility. Based on the results ideas for including other machine learning algorithms like XGBoost and others can be compared, research has already in action for these algorithms. It is inclusive of all the parameters needed to evaluate the creditworthiness of a client. The model is trained to produce results with satisfactory accuracy, after which it produces accurate results as to whether a borrower should be lent money or not without any tedious manual work.

VI. CONFLICT OF INTEREST

The authors of the article entitled "Machine learning Techniques for Recognizing the Loan Eligibility" declare that there are no conflicts of interest regarding the research manuscript. None of the either Human Being or Animals are affected throughout the research processing.

VII. REFRENCES

[1] Deepak Ishwar Gouda, Ashok Kumar A, Anil Manjunatha Madivala, Dilip Kumar R, Dr.Ravikumar, "LOAN APPROVAL PREDICTION BASED ON MACHINE LEARNING", International Research Journal of Engineering & Technology, Volume-8 Issue-11, January 2021.

[2] Sharayu Dosalwar ,Dr. Vishwanath Karad ,Ketki Kinkar,Rahul Sannat,Nitin Pise, "Analysis of Loan Availability using Machine Learning Techniques", September 2021, DOI: http://dx.doi.org/10.48175/IJARSCT-1895

[3] Kumar, R., Jain, V., Sharma, P. S., Awasthi, S., & Jha, G. (2019). Prediction of Loan Approval using Machine Learning. International Journal of Advanced Science and Technology, 28(7), 455 - 460. Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/460

[4] Ramya S , Priyesh Shekhar Jha , Ilaa Raghupathi Vasishtha , Shashank H , Neha Zafar, "Monetary Loan Eligibility Prediction using Machine Learning", IJESC, Volume:11 Issue-7.

[5] Prateek Dutta, "A Study on Machine Learning Algorithm for Enhancement of Loan Prediction", "International Research Journal of Modernization in Engineering Technology and Science", Volume -3 issue-1, January 2021.

[6] AFRAH KHAN, EAKANSH BHADOLA, ABHISHEK KUMAR and NIDHI SINGH, "LOAN APPROVAL PREDICTION MODEL A COMPARATIVE ANALYSIS", Advances and Application in Mathematical Science, January,2021.

[7] A. Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017, pp. 1-6, doi: 10.1109/ICCCNT.2017.8203946.

[8] Efosa G. Adagbasa, Samuel A. Adelabu, Tom W. Okello, "Application of deep learning with stratified K-fold for vegetation species discrimination in a protected mountainous region using Sentinel-2 image", Geocarto International, 19 December,2019, DOI: https://doi.org/10.1080/10106049.2019.1704070

[9] A. Fernandez-Carrillo, D. de la Fuente, F. W. Rivas-Gonzalez, and A. Franco-Nieto "An automatic Sentinel-2 Forest types classification over the Roncal Valley, Navarre: Spain", Proc. SPIE 11156, Earth Resources and Environmental Remote Sensing/GIS Applications X, 111561N (3 October 2019); https://doi.org/10.1117/12.2533059

[10] Prateek Dutta, "A Study on Machine Learning Approach for Market Segmentation", International Journal of Scientific Research in Engineering and Management", Volume-5 Issue-7, July 2021.

[11] Michael P. LaValley, "Logistic Regression", Circulation, DOI: https://doi.org/10.1161/CIRCULATIONAHA.106.682658

[12] Wright, R. E. (1995). Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), Reading and understanding multivariate statistics (pp. 217–244). American Psychological Association.

[13] Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting", The Journal of Educational Research, April 2010, DOI: https://doi.org/10.1080/00220670209598786

[14] Harsh Patel, Purvi Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms", International Journal of Computer Science and Engineering, October 2018, DOI:

[15] http://dx.doi.org/10.26438/ijcse/v6i10.7478

[16] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, May-June 1991,

[17] Doi: 10.1109/21.97458.

[18] Anthony J. Myles, Robert N. Feudale, Yang Liu, Nathaniel A. Woody, Steven D. Brown, "An introduction to decision tree modeling", Journal of Chemometrics, 2004, DOI: https://doi.org/10.1002/cem.873

[19] M. Pal, "Random Forest classifier for remote sensing classification", International Journal of Remote Sensing, 2007, DOI: https://doi.org/10.1080/01431160412331269698

[20] V.F.Rodriguez-Galiano, B.Ghimire, J.Rogan, M.Chica-Olmo, J.P.Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification", ISPRS Journal of Photogrammetry and Remote Sensing, January 2012,
DOI: https://doi.org/10.1016/j.isprsjprs.2011.11.002

[21] Devetyarov D., Nouretdinov I. (2010) Prediction with Confidence Based on a Random Forest Classifier. In: Papadopoulos H., Andreou A.S., Bramer M. (eds) Artificial Intelligence Applications and Innovations. AIAI 2010. IFIP Advances in Information and Communication Technology, vol 339. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-16239-8_8

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 8, August-2020

Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval

Amruta S. Aphale
Department of Computer Science and Engineering
Savitribai Phule Pune University
Vishwakarma Institute of Technology, Pune

Prof. Dr. Sandeep. R. Shinde
Department of Computer Science and Engineering
Savitribai Phule Pune University
Vishwakarma Institute of Technology, Pune

Abstract - In today's world, taking loans from financial institutions has become a very common phenomenon. Everyday a large number of people make application for loans, for a variety of purposes. But all these applicants are not reliable and everyone cannot be approved. Every year, we read about a number of cases where people do not repay bulk of the loan amount to the banks due to which they suffers huge losses. The risk associated with making a decision on loan approval is immense. So the idea of this project is to gather loan data from multiple data sources and use various machine learning algorithms on this data to extract important information. This model can be used by the organizations in making the right decision to approve or reject the loan request of the customers. In this paper, we examine a real bank credit data and conduct several machine learning algorithms on the data for that determine credit worthiness of customers in order to formulate bank risk automated system.

Keywords— Machine learning, bank credit, classification, confusion matrix, predictive analysis.

I. INTRODUCTION

Bank plays a vital role in market economy. The success or failure of organization largely depends on the industry's ability to evaluate credit risk. Before giving the credit loan to borrowers, bank decides whether the borrower is bad (defaulter) or good (non defaulter).The prediction of borrower status i.e. in future borrower will be defaulter or non defaulter is a challenging task for any organization or bank. Basically the loan defaulter prediction is a binary classification problem Loan amount; costumer's history governs his credit ability for receiving loan. The problem is to classify borrower as defaulter or non defaulter. However developing such a model is a very challenging task due to increasing in demands for loans. Prototypes of the model which can be used by the organizations for making the correct or right decision for approve or reject the request for loan of the customers. This work includes the construction of an ensemble model by combining different machine learning models. Banks struggle a lot to get an upper hand over each other to enhance overall business due to tight competition. Credit Risk assessment is a crucial issue faced by Banks nowadays which helps them to evaluate if a loan applicant can be a defaulter at a later stage so that they can go ahead and grant the loan or not. This helps the banks to minimize the possible losses and can increase the volume of credits.

II. BACKGROUND

The most important background information on machine learning algorithms and their theoretical formulation are out- lined in this section. These algorithms are used in analyzing the bank credit data.

A. Machine Learning Algorithms

Machine learning techniques can be grouped broadly into two main categories. They include:

(i) **Supervised Learning:** The main feature of this algorithm consists of target or outcome variable (or dependent variable). The target variable is used to predict other features from a given set of pre- dictors (independent variables). Furthermore, using the target variable, a function is generated that maps input to desired outputs. The training process then continues until the model achieves the desired level of accuracy on the training data. Supervised learning techniques are achieved using regression and classification algorithms or approaches that range from non-linear regression, generalized linear regression, discriminant analysis, Support Vector Machines (SVMs) to decision trees and ensemble methods.

(ii) **Unsupervised Learning:** In unsupervised learning, there is no target or outcome variable to predict or estimate. This algorithm is used mainly for segmenting or clustering entities in different groups for specific intervention. Examples of unsupervised learning algorithms include Apriori and K-means algorithms.

The various machine learning approaches and the algorithms that describe them are shown in Fig. 1

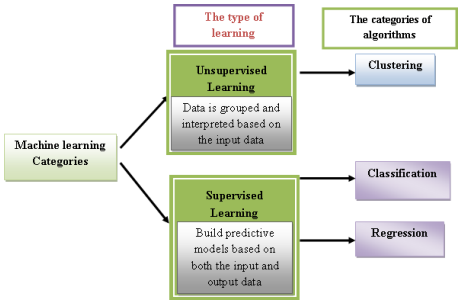


Fig. 1. Machine learning Tasks

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 8, August-2020

Labeled data is known in the literature to be suitable for classification algorithms. The dataset used in this paper is a labeled data and is, therefore, suitable for doing classification analysis. And thus, we employed various classification algorithms described comprehensively in Section II-B. Some of the algorithms are implemented in MATLAB [§] and some taken from the *Python scikit-learn package* to predict the creditworthiness of bank customers with regards to their ability to pay their credit or otherwise within a given time frame.

B. Classification Algorithms

Classification algorithms work by predicting the best group to which a data point belongs to by "learning" from labeled observations. It uses a set of input features for the "learning" process. Classification algorithms are good for grouping data that are never seen before into their various groupings and are therefore extensively used in machine learning tasks. Some of the well-known classification algorithms used in this paper are briefly discussed below:

1) Neural Networks: The neural network supports both classification and regression algorithms and therefore, is very appropriate for studying the classification problem in this paper.
2) Discriminant Analysis: The discriminant analysis is based on the assumption that different classes of data are generated by using different Gaussian distributions. The main types of discriminant algorithms used for classification are the linear and the quadratic discriminant. We used the quadratic discriminant classifier in this paper.
3) Naive Bayes: This classification technique is based on Bayes' theorem that assumes independence between predictors, thus, the presence of a particular feature in a class is independence of another feature in a another class. Naive Bayes classification is therefore, based on estimating $P(X \mid Y)$, the probability or probability density of features X given class Y.
4) K-Nearest Neighbor: The KNN algorithm is used for both classification and regression problems. However, the KNN is more widely used in classification problems in the industry and thus will be used in doing classification and predictive analysis in this paper. The KNN is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common amongst its K nearest neigh- bors measured by a distance function. The common distance functions used are the *Euclidean*, *Manhattan*,*Minkowski* and *Hamming distance*.
5) Linear Regression: It is used to estimate real values based on continuous variable(s). In linear regression, a relationship is established between independent and dependent variables by fitting the best line. This best fit line is known as regression line and represented by a linear equation $Y = a\,X + b$, where Y is the dependent variable, a is the slope, X is the independent variable and b is the intercept. The coefficients a and b are derived based on minimizing the sum of squared difference of distance between data points and regression line.
6) Ensemble Learning/method: An example of ensemble learning method is the TreeBagger, where the bagging stands for bootstrap aggregation. Every tree in the ensemble is grown on an independently drawn sample of input data. To compute the prediction for the ensemble of trees, TreeBagger takes an average of predictions from individual trees (for regression) or takes votes from individual trees (for classification). Ensemble techniques such as bagging combine many weak learners to produce a strong learner.
7) Decision Trees: There are two kinds of decision trees; classification trees and regression trees. A deci- sion tree can be described as a flow-chart like structure in which internal node represents test on an attribute, each branch represents outcome of the test and each leaf node represent decision taken after computing all attributes or a response after computing all given attributes.

III. RELATED WORK

The related work n the application of machine learning and data approaches to study financial data are comprehensively described below. Li *et al.* [6] conducted research on using attributes of customers to assess credit risk by using a weighted-selected attribute bagging method. They benchmarked their result experimentally by using two credit databases and reported outstanding performance both in term of prediction accuracy and stability as compared with another state of the art methods. A data mining approach is also pro- posed by Moro *et al.* [7] to predict the success or otherwise of a Portuguese retail bank in telemarketing. They applied various data mining models on the bank telemarketing data and reported that the neural network data mining method was the best for analyzing the data. The role of machine learning techniques in business data mining is outlined by [8]. Their work described the strengths and weaknesses of various machine learning techniques within the context of business data mining approach. Their analysis revealed that Rule Induction Technique was the best approach in mining business data, followed by that of the neural network approach. C. Tsai and M. Chen [9] used a hybrid machine learning approach to study credit rating by comparing four different types of hybrid machine learning techniques. They showed experimentally that 'classification + classification' hybrid model based on a combination of logistic regression and neural networks provides the highest prediction accuracy and also maximize the profit. Bank default data was used by [10] to model bank failure predictions using neural network approach. They compared their result with other machine learning approaches and concluded that the neural network approach is a promising method in terms of pre- dictive accuracy, adaptability, and robustness. [11] proposed. They experiment with the hybrid recommendation algorithms on two sets of data and reported high scalability and better performance in terms of accuracy and coverage. A hybrid online sequential extreme learning machine with the simplified hidden layer is proposed by [12]. The algorithm is a combination of the Online Sequential Extreme Learning Machine and the Minimal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 8, August-2020

Resource Allocation Network. Their experimental results showed that the algorithm has a comparable performance as that of the original online sequential extreme learning machine but with a reduced number of hidden layers.

IV.METHODOLOGY

The proposed model focuses on predicting the credibility of customers for loan repayment by analyzing their behavior. The input to the model is the customer behavior collected. On the output from the classifier, decision on whether to approve or reject the customer request can be made. Using different data analytics tools loan prediction and there severity can be forecasted. In this process it is required to train the data using different algorithms and then compare user data with trained data to predict the nature of loan. To extract patterns from a common loan approved dataset, and then build a model based on these extracted patterns. The training data set is now supplied to machine learning model; on the basis of this data set the model is trained. Every new applicant details filled at the time of application form acts as a test data set. After the operation of testing, model predict whether the new applicant is a fit case for approval of the loan or not based upon the inference it conclude on the basis of the training data sets. To extract important information and predict if a customer would be able to repay his loan or not.

TABLE I
THE COMPOSITION OF THE DATASET
Bank Credit Data

Value	Count	Percentage
No	23364	77.88%
Yes	6636	22.12%
Training Dataset		
No	14092	78.29%
Yes	3908	21.71%
Test dataset		
No	9272	77.27%
Yes	2728	22.73%

If T and N denotes the number of clients that will not default the credit payment and clients that will default in the payment of their credit respectively, then the total number of the dataset is expressed as $T + N$. Furthermore, *TP* (True Positive) and *TN* (True Negative) represent the total positive and negative cases/instances that are rightly classified, respectively. The *FP* and *FN* also denote the number of predicted/classified instances that are incorrectly predicted *yes* when it is actually *no* and the number of instances that are predicted *no* when it actually *yes*, respectively. These constitute the entries to the confusion matrix shown in Table II

TABLE II
LAYOUT of CONFUSION MATRIX

Actual class		***Predicted class***	
		no	*yes*
	no	**TP**	**FP**
	YES	**FN**	**TN**

V. EXPERIMENT SETUP

The major steps we employed in developing the machine learning tasks/algorithms are further discussed below

- Step 1: *Collect the data*: The dataset used in this paper is from cooperative bank .
- Step 2: *Prepare the input data*: This step was done by the original owners of the dataset. And the composition of the dataset is shown in Table I.
- Step 3: *Analyze the input data*: understand the relationship among different features. A plot of the core features and the entire dataset.The dataset is further split into 2/3 for training and 1/3 for testing the algorithms. Furthermore, in order to obtain a representative sample, each class in the full dataset is represented in about the right proportion in both the training and testing datasets. The various proportions of the training and testing datasets used in the paper are shown in Table I.
- Step 4:*Train the algorithm*: The various classification algorithms are trained using a different set of data. The training dataset is shown in Table I
- Step 5: *Test the algorithm*: The various algorithms are used to predict the effectiveness of the algorithm on the test dataset. In evaluating the performance of the classification algorithms, It include accuracy, precision, recall, specificity and F-measure (F1-measure). These values are calculated using the Python scikit-learn tool with input values as the entities of the confusion matrix. The formula for the various evaluating metrics is shown in III, with their definitions. In this paper, a 'positive' instance refers to *no*(signifying there will not be a default in the payment of the loan) whereas the 'negative' instance refers to *yes* (signifying there will be a default in the payment of the loan).

A. *Extracting the Importance Features for Predicting Credit Defaulters*

The total number of features within the bank credit Defaulters dataset. However, not all have significant influence in determining the ability of a given customer in paying his/her loan or not. The designed system is tested with test set and the performance is assured. Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time. Common metrics calculated from the confusion matrix are Precision; Accuracy

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 8, August-2020

A. The Predictive Model

The most important features since these features are to develop a predictive model using ordinary linear regression model. This can serve as a tool in determining the credit worthiness of bank clients because these are among the main features taken into consideration by most banks in advancing loans to customers.

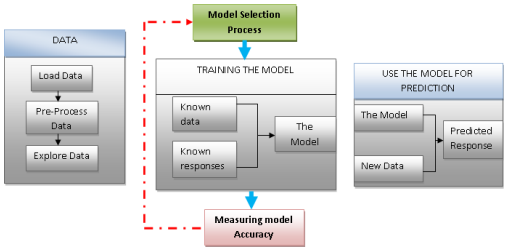


Fig 2. Work Flow in Machine learning

Evaluation Metrics

1 Accuracy:
It measures how often the classifier is correct for both true positives and true negative cases. Mathematically, it is defined as:
Accuracy = (True Positive + True Negative)/Total Predictions
2 Sensitivity or Recall:
measures how many times did the classifier get the true positives correct. Mathematically, it is defined as:
Recall = True Positive/ (True Positive + False Negative)
3 Specificity:
It measure how many times did the classifier get the true negatives correct. Mathematically, it is defined as:
Specificity = (True Negative)/ (True Negative + False Positive)
4 Precision:
Precision measures off the total predicted to be positive how many were actually positive. Mathematically, it is defined as:
Precision = (True Positive)/(True Positive + False Positive)

model prediction

actual loan status		no default (0)	default (1)
	no default (0)	TN	FP
	default (1)	FN	TP

Fig 3:Confusion Matrix

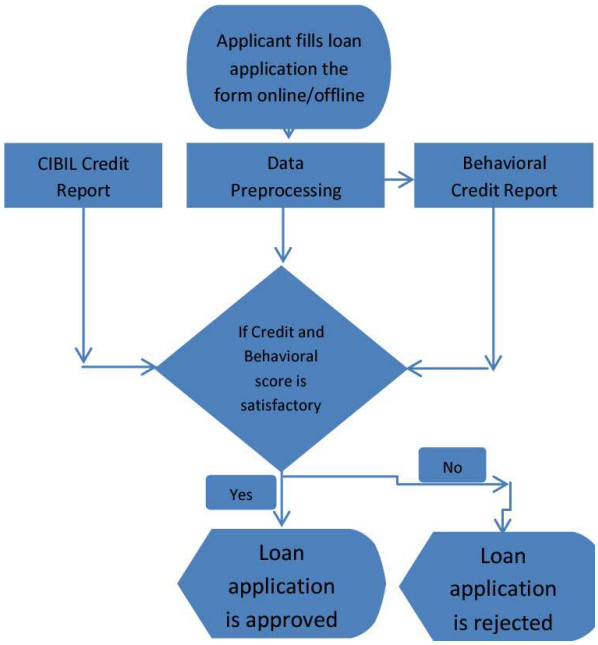


Fig. 4. Predict loan approval flow model

VI.CONCLUSION

In this paper, machine learning approach to study bank credit dataset in order to predict customers' credit worthiness their ability to pay their loan. We employed different machine learning algorithms on the dataset in order to determine which algorithms are the best fit for studying bank credit dataset. The experiment revealed that, apart from the Nearest Centroid and Gaussian Naive Bayes, the rest of the algorithms perform credibly well in term of their accuracy and other performance evaluation metrics. Each of these algorithms achieved an accuracy rate between 76% to over 80%. We also determined the most important features that influence the credit worthiness of customers. These most important features are then used on some selected algorithms and their performance accuracy compared with the instance of using all features. The experimental results showed no significance difference in their predictive accuracy and other metrics. We further formulated a predictive model using linear regression, that composed of the most important features, for predicting customers credit worthiness. Predict loan approval in Banking system that will incorporate the most important features that determine credit worthiness of customers in order to formulate bank risk automated system.

REFERENCES

[1] G. McLachlan, K.-A. Do, and C. Ambroise, *Analyzing microarray gene expression data*, vol. 422. John Wiley & Sons, 2005.
[2] E. Ngai, Y. Hu, Y. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems*, vol. 50, no. 3, pp. 559–569, 2011.
[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.
[4] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 8, August-2020

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Van- derplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] J. Li, H. Wei, and W. Hao, "Weight-selected attribute bagging for credit scoring," *Mathematical Problems in Engineering*, vol. 2013, 2013.

[7] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, vol. 62, pp. 22–31, 2014.

[8] I. Bose and R. K. Mahapatra, "Business data mininga machine learning perspective," *Information & management*, vol. 39, no. 3, pp. 211–225, 2001.

[9] C.-F. Tsai and M.-L. Chen, "Credit rating by hybrid machine learning techniques," *Applied soft computing*, vol. 10, no. 2, pp. 374–380, 2010.

[10] K. Y. Tam and M. Y. Kiang, "Managerial applications of neural networks: the case of bank failure predictions," *Management science*, vol. 38, no. 7, pp. 926–947, 1992.

[11] M. Ghazanfar and A. Prugel-Bennett, "Building switching hybrid rec- ommender system using machine learning classifiers and collaborative filtering," *IAENG International Journal of Computer Science*, vol. 37, no. 3, 2010.

[12] M. Er, L. Zhai, X. Li, and L. San, "A hybrid online sequential extreme learning machine with simplified hidden network," *IAENG International Journal of Computer Science*, vol. 39, no. 1, pp. 1–9, 2012.

[13] M. Lichman, "UCI machine learning repository," 2013.

LOGISTIC REGRESSION BASED LOAN APPROVAL PREDICTION

Sai Aparna Vangaveeti[1], Naga Likitha Venna[2], Prasanna Naga Sri RamyaYajamanam Kidambi[3], Harika Marneni[4], Naga Satish Kumar Maganti[5].

[1,2,3,4] Department of CSE, Gudlavalleru Engineering College, India.

[5]Assistant Professor, Department of CSE, Gudlavalleru Engineering College, India.

E-mail: maganti.nagasatishkumar@gmail.com[5].

ABSTRACT

As we know that now-a-days there is a rapid growth in banking sector, resulting lots of people are applying for bank loans. Finding out the applicant to whom the loan will be approved is a difficult process. In this paper, we proposed a model which predicts loan approval/rejection of an applicant using machine learning techniques. This can be done by training the model with the data of the previous records of the people applied for loan.

Keywords: Banking Sector, loan, predict, machine learning.

INTRODUCTION

Distribution of the loans is the main business part of almost every bank. The main portion of the bank's asset is directly from the profit earned from the loans distributed by the banks.

The prime goal in banking domain is to invest their assets in safe hands. Lending money to unsuitable loan applicants results in the credit risk. Today many banks approve loans after a long procedure of verification, yet there is no guarantee whether the picked candidate is the right candidate or not.

Estimating the risk, which is involved in a loan application, is one of the most significant concerns of the banks in order to survive in the highly competitive market. Through our proposed model we can predict whether that specific customer is safe or not and the entire procedure of approval of features validation is automated by machine learning technique.

Data mining algorithms are used to study the loan-approved data and exact patterns, which would help in predicting the reasonable defaulters, thereby helping the banks for making better choices in the future. Loan Prediction is extremely useful for employee of banks and for the applicant also. The main aim of this model is to provide a speedy, immediate and simple approach to pick the deserving applicants.

The Loan Prediction System automatically calculates the weight of each feature involved in loan processing and on new test data same features are processed with respect to their associated weight. A period breaking point can be set for the applicant to check whether his/her loan can be approved or not.

This model is solely for the managing authority of Bank/finance companies, entire procedure of prediction is done secretly that is, no stakeholders would able to alter the processing. Result against specific Loan Id can be send to different departments of banks in order to take an appropriate action on application.

RELATED WORK

In [1] the author acquaints a structure to successfully recognize the Probability of Default of a Loan applicant. The metrics got from the predictions reveal the high accuracy of the built model.

In [2] an effective model was proposed for predicting the right customers who have applied for loan. Decision Tree is applied to foresee the traits significant for believability.

The model proposed in [3] has been built using data from banks to predict the status of loans. This model uses three classification algorithms namely j48, bayes Net and naive Bayes. The model was implemented using Weka.

In [4] a decision tree model was utilized as a classifier and for feature selection genetic algorithm is utilized. The model was tried utilizing Weka.

In [5] two data mining models were created for credit scoring that helps in decision making of giving loans for the banks in Jordan. With the consideration of accuracy rate, the regression model is found to perform better than radial function model.

The work in [6] analyses support vector machine based models for credit-scoring created using the different default definitions. The work inferred that the expansive definition models are better than the narrow definition models in their performance.

In [7] financial data analysis was done by figuring out techniques like Decision Tree, Random forest, Boosting, Bayes classification, Bagging algorithm etc. Techniques like Support Vector Machine, Decision Tree, Logistic Regression, Neural Network, Perception model are combined in this model. The accuracy rate of each of these techniques is studied. The analysis results show the performance is extraordinary based on accuracy.

PROPOSED METHODOLOGY

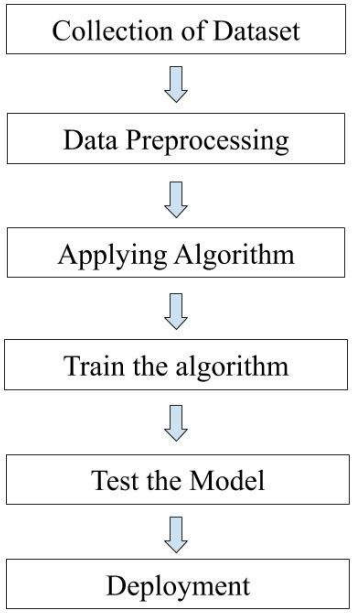


Fig 1: Flow of implementation

Logistic Regression

In our proposed model we had used Logistic regression which is one of the popular Machine Learning algorithms that comes under the Supervised Learning technique. It is applicable for categorical dependent variable using a given set of independent variables. Thus, the outcome must be a categorical or discrete value. The output can be either Yes or No, 0 or 1, true or false, etc. but instead of giving the exact value as 0 or 1, it gives some probabilistic values which lies between 0 and 1. Logistic regression is much similar to linear regression except that how they are used. It is used for solving regression problems, whereas Logistic regression is used for solving the classification problems.

In Logistic regression, rather than fitting a regression line, we fit an "S" shaped logistic function, which predicts two greatest values (0 or 1). The curve from the logistic function demonstrates the probability of something, for example, regardless of whether the cells are destructive or not, a mouse is corpulent or not founded on its weight, and so on. It is a significant algorithm because it can provide probabilities and classify the use of different types of data and easily determines the most effective variables that are used for classification.

Logistic Function (Sigmoid Function)

The sigmoid function is a numerical function used to outline predicted values to probabilities. It maps any real value to another value that is in between 0 and 1.The value must be in between 0 and 1 which means it can exceed the limit, then it forms a curve like the "S" form.

The S-structure curve is also known as the sigmoid function or the logistic function. In logistic regression, we utilize the concept of threshold value, which characterizes the probability of either 0 or 1 and the value below the threshold values tends to 0.

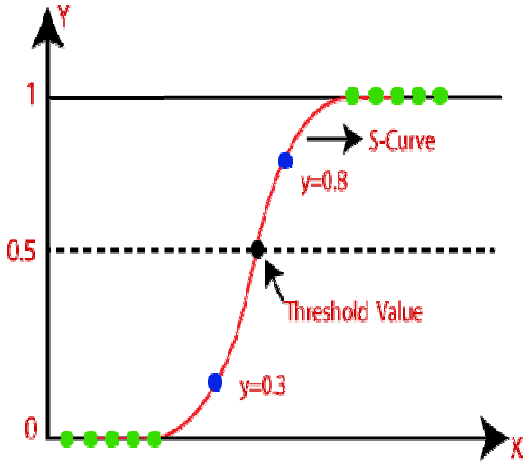


Fig 2: Logistic Function

Logistic Regression Equation

The equation can be obtained from Linear Regression equation. The mathematical steps to obtain the equations are given below:

The equation of the straight line can be written as:

$$y = b0+b1x1+b2x2+b3x3+\ldots\ldots+bnxn$$

In Logistic Regression y can be between 0 and 1 only, so let's divide the above equation by (1-y):

y/1-y;0 for y = 0 and infinity for y = 1

But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$\log(1/1-y) = b0+b1x1+b2x2+b3x3+\ldots\ldots+bnxn$$

Above one is the final equation of Logistic Regression.

Steps in Logistic Regression

For implementing the Logistic Regression using Python, we had done the following steps:

1. Data Pre-processing
2. Fitting Logistic Regression to the Training set
3. Predicting the test result
4. Test accuracy of the result
5. Visualizing the test set result.

EXPERIMENTAL RESULTS

Based on the data given by the loan applicant, we can predict whether the loan of particular applicant is approved or not using a User Interface. User interface contains input variables with their corresponding fields and a field to display the output. Input variables are Gender, Marital status, Dependents, Education, Applicant income, Loan Amount, Loan amount term, Credit History, Property Area. The applicant need to give these values and based on these, the model will predict whether the loan will be approved or not.

User Interface

Fig 3: User Interface

CONCLUSION AND FUTURE SCOPE

Finally, in our model by using logistic regression model we predict whether the loan is approved or not. In order to implement this various input variables were used to get the output. Whenever program takes the input data it gives the output in the form of binary i.e., either 0 or 1. If the output is 1 then '1' will be displayed and it indicates that loan is approved. If the output is 0 then '0' will be displayed and it indicates that loan is not approved.

Here, we had implemented loan credibility prediction system that helps the organizations in making the right decision to approve or reject the loan request of the customers. This will definitely help the banking industry to open up efficient delivery channels. In this model, Logistic Regression algorithm is used for the prediction. Incorporation of other techniques that outperform the performance of popular data mining models has to be implemented and tested for the domain.

REFERENCES

[1]. Sudhamathy G and Jothi Venkateswaran "Analytics Using R for Predicting Credit Defaulters", IEEE international conference on advances in computer applications (ICACA), 978-1-5090-3770-4, 2016.

[2]. M. Sudhakar, and C.V.K. Reddy, "Two Step Credit Risk Assessment Model For Retail Bank Loan Applications Using Decision Tree Data Mining Technique", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 5, no.3, pp. 705-718, 2016.

[3]. J.H. Aboobyda, and M.A. Tarig, "Developing Prediction Model Of Loan Risk In Banks Using Data Mining", Machine Learning and Applications: An International Journal (MLAIJ), vol. 3, no.1, pp. 1–9, 2016.

[4]. Z. Somayyeh, and M. Abdolkarim,"Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining A Case Study: Branches of Mellat Bank of Iran", Jurnal UMP Social Sciences and Technology Management, vol. 3, no. 2, pp. 307–316, 2015.

[5]. A.B. Hussain, and F.K.E. Shorouq, "Credit risk assessment model for Jordanian commercial banks: Neuralscoring approach", Review of Development Finance, Elsevier, vol. 4, pp. 20–28, 2014.

[6]. T. Harris, "Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions", Expert Systems with Applications, vol. 40, pp. 4404– 4413, 2013.

[7]. Dileep B. Desai, Dr. R.V.Kulkarni "A Review: Application of Data Mining Tools in CRM for Selected Banks", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2), 2013, 199 – 201.

IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 3, March 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 3, March 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1103095 |

ML Based Loan Approval Prediction System A Novel Approach

Dr. MK Jayanthi Kannan[1], A.R Nithej[2], Padi Akhil[3], Pavan Gowda[4], Prashant Pareek [5]

[1]Professor, School of Engineering and Technology, Jain University Bangalore, India

UG Student , Department of Information Science and Engineering School of Engineering and Technology, Jain University Bangalore, India [2,3,4,5]

ABSTRACT: Loan approval system is important in banks in order to reduce the loss and approve loans only for eligible customers, who are able to repay the loan amount. Various studies can see in this area and many studies are still focusing on this problem as give assets on safe hand in very important for any bank. However, the performance of previous studies is good, but the accuracy can be still increased. The main objective of this study is to increase the performance of loan prediction system. This study is focusing on different machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, K nearest neighbors, Artificial neural network, Naive Bayes, Adaboost, and Voting classifier to predict the loan approval.

KEYWORDS: Decision tree, random forest, support vector machine, linear models, neural networks, adaboost, loan approval

I. INTRODUCTION

Loan is very essential for the current world. Various types of platform are available in the world to take loan on low rate, where some are private, and few are govt. platform. Normally, to give loan, banks are focus on the CIBIL (Credit Information Bureau (India) Limited) score, credit score along with background of applicant to give a loan. However, it has been seen in the history that applicant took loan but didn't repay it. So, to save the bank assets from the defaulter become a necessity of the world as directly or indirectly, banks are important in the growth of any country (Arun, Ishan, & Sanmeet, 2016). Various research research has been done in this area to safe the bank as well as the growth of the country. However, the performance of previous studies is good, but the accuracy can be still increased (Vangaveeti, Venna, Kidambi, Marneni, & Maganti, 2020). Many machine learning algorithms has been used to predict the loan and some of them shown very good performance. Such as: logistic regression (LR) , decision tree (DT), Artificial neural network (ANN), K-Nearest Neighbour (KNN), Naïve Bayes (NB), Support vector machine (SVM), and few algorithms under Ensemble techniques like Random Forest (RF), Adaboost,Voting classifier (VC). Apart from predicting, few works are completed to do analysis of the background details of the applicant by using P2P and Exploratory Data Analysis (EDA).To build the loan prediction system, dataset always play a very important role. Apart from the dataset, pre-processing, and classification is also important. Various types of features are presented in the dataset, to filter the dataset and convert into proper manner, is a part of pre-processing. In the last, model will use the dataset to train the system where many machine learning algorithms will use to complete the training part.This study is divided into six parts: introduction, literature, dataset, methodology, result and conclusion. Details of each step such as: literature, dataset, methodology, result are discussed properly in the perspective to loan prediction. In the last, the performance of this study is compared with other studies.

II. RELATED WORK

A study was conducted to predict whether a borrower will default on a loan is of significant concern to platforms and investors in online P2P lending (Jiang, Wang, Wang, & Ding, 2017[1]. A two-stage method designed to select an effective feature set containing both soft and hard information along with P2P was used to complete the study. An empirical analysis using realword data from a major P2P lending platform in China shows that the proposed method can improve loan default prediction performance compared with existing methods based only on hard information. This study introduced a topic model to extract valuable features from the descriptive text concerning loans and construct four default prediction models to demonstrate the performance of these features for default prediction.

Another study was also completed to create a credit scoring model for credit data and loan approval status Arutjothi and Senthamarai in the year 2017 [2]. Authors used MinMax Normalization and K-Nearest Neighbour (KNN) to conduct

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 3, March 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1103095 |

this research and proposed a system that showed good performance along with useful information. The performance of this paper using logistic regression as a tool, this paper specifically delineates about whether or not loan for a set of records of an applicant will be approved.

Similarly, Vaidya in 2017 [3] and Xiaojun Ma et al. [4]in 2018 used logistic regression, and P2P, Data cleaning, Default rate, LightGBM algorithm, XGboost algorithm, respectively, to predict the loan approval status. The performance of this study was good, and the system was able to predict that applicants will be approved for loan or not. One study focused on various parameters that should be approved or not by Jency, Sumathi, & Sri in 2019[5]. To do analysis, Exploratory DataAnalysis (EDA) was used. After the analysis, it can say that short term loans are chosen mostly by the clients. Another analysis was also conducted to find out the relationship between the Italian bank that, within a bank, approves a loan and the subsequent performance of the loan by Calcagnini, Cole, Giombini, & Grandicelli, [6] in 2018. P2P was used to complete this study.

Similarly, in 2020, a study by Tejaswini, Kavya, Ramya, Triveni, & Maddumala [7] in 2020 focussed on the prediction whether the loan in terms of banking investment is in safe hand or not, by using Logistic Regression (LR), Decision Tree (DT), Random Forest (RF) algorithm. As compared to LR and RF, DT has performed well in terms of accuracy Furthermore in 2020, there was another study done by Vangaveeti et al [8] in 2020, using Logistic Regression algorithm for the prediction to provide a speedy, immediate and simple approach to pick the deserving applicants and the performance was good based on accuracy.

Another study titled 'Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines' by Min Li, Amy Mickel, Stanley Taylor [9] in the year 2018 Logistic Regression was used and In this article, a large and rich dataset from the U.S. Small Business Administration (SBA) and an accompanying assignment designed to teach statistics as an investigative process of decision making are presented.

Entropy method of constructing a combined model for improving loan default prediction in 2019 [10] by Yihen Li used Random forest, logistic regression, artificial neural network the experimental results reveal that the proposed combined model outperforms the two base models on four evaluation metrics.

Again, EDA was used to do analysis that the person is getting loan is reliable or not and after that Logistic Regression, Decision Tree, SVM, Naïve Bayes also used to predict the status of the loan by Blessie & Rekha, [11]in 2019. The performance of the Logistic Regression, Decision Tree, SVM, and Naïve Bayes was 78.91%, 71.92%, 65.27%, and 80.42%, respectively.According to this literature, various algorithms of machine learning has been applied to predict the loan approval. Moreover, several algorithms such as supervised, and ensemble algorithm can be also used to increase the performance to loan approval system.

III. PROPOSED ALGORITHM

A. *Algorithms used:*

- Logistic regression
- Decision Tree
- Random forest classifier
- K-Nearest Neighbor
- Naïve Bayes
- AdaBoost algorithm
- Linear SVM
- Polynomial SVM
- Wavelet SVM

B. *Description of the Proposed Algorithm:*

Logistic regression is a supervised learning classification to predict the probabilty of a target variable

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. The decisions or the test are performed on the basis of features of the given dataset.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 3, March 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1103095 |

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

K-Nearest Neighbor is one of the simplest Machine Learning algorithmsbasedonSupervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

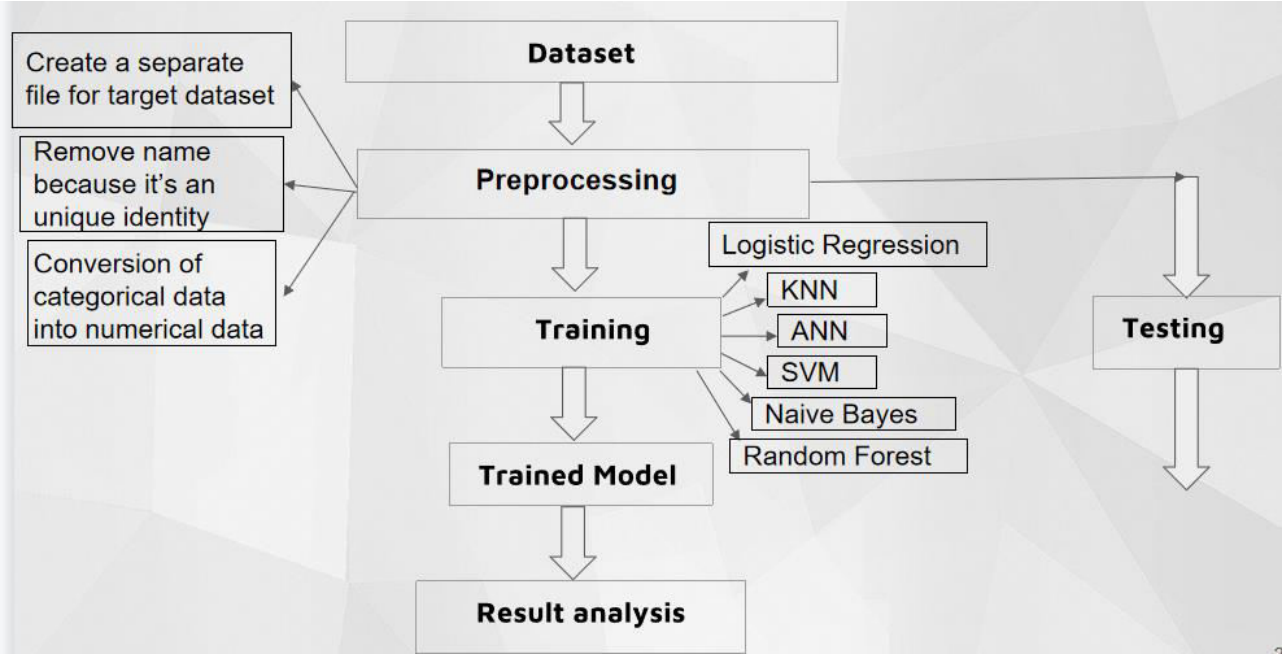
AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique that is used as an Ensemble Method in Machine Learning. Voting is one of the simplest way of combining the predictions from multiple machine learning algorithms. We can train data set using different algorithms and ensemble then to predict the final output. The final output on a prediction is taken by majority vote accordingly.

Linear SVM, Linear SVM is Super classifier. Generally linear SVM is used in biclassification problems, for example, the problem setting. Where there are two classes coming into consideration

Polynomial SVM, polynomail kernel function is a normal kernel function that is commonly used with SVM in machine learning. It represents the relation of vectors in a feature space over polynomials of actual variables by allowing learning on non linear varible linear models.

Wavelet SVM Wavelet variance measures variablitiy in the form of equivalentlly or signal by scale, variability trends, in a signal oer frequency

C Methodology:



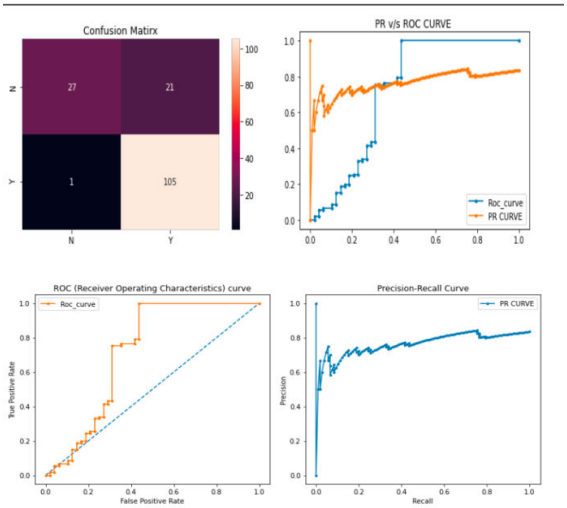
International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

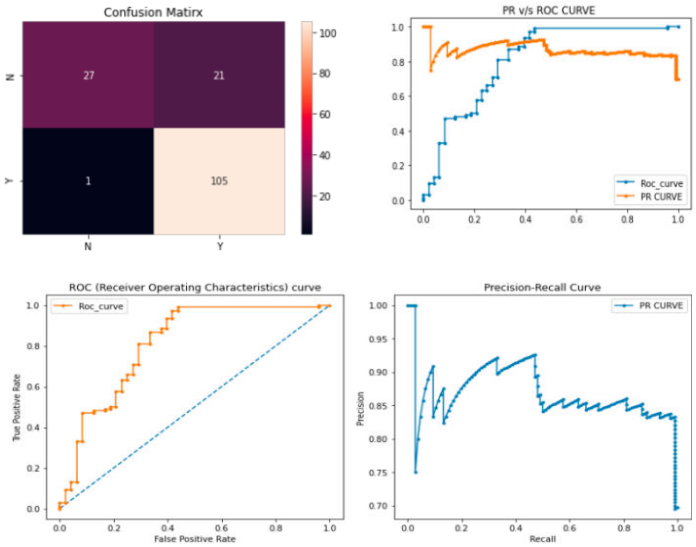
|| Volume 11, Issue 3, March 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1103095 |

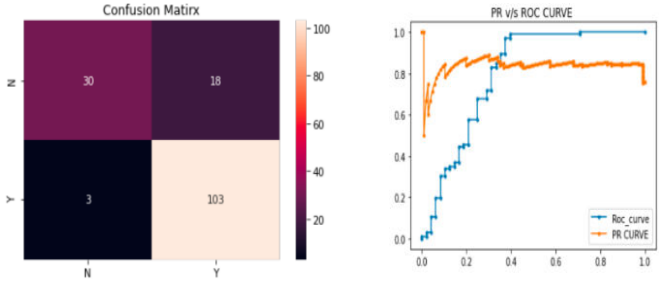
IV. SIMULATION RESULTS

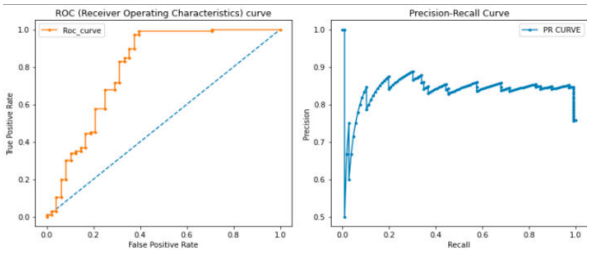


linear SVM

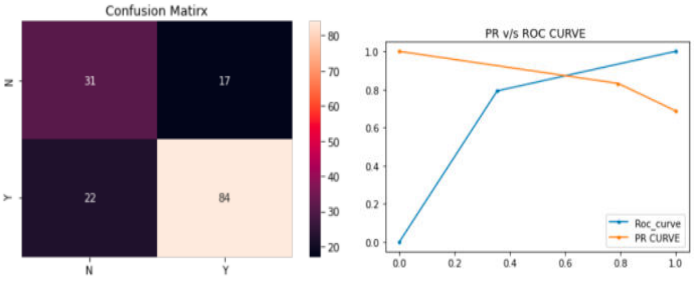


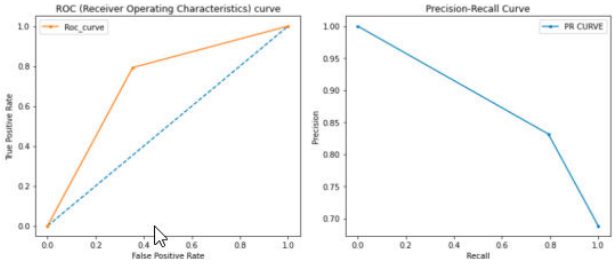
logistic regression





Naive Bayes





Decision Tree

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 3, March 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1103095 |

ALGORITHMS	PERFORMANCE OF THE PREVIOUS STUDY (%)	PERFORMANCE OF THE PROPOSED STUDY (%)
Logistic Regression	78.91	86
Decision Tree	71.92	74
SVM (RBF)	65.27	86
Naïve Bayes	80.42	86

In the previous study, Logistic regression, Decision tree, Support Vector Machine (RBF), and Naive Bayes got a performance of 78.91%, 71.92%, 65.27%, and 80.42%, respectively

It can be seen in the comparison of results, the proposed study performed well and provided a higher performance of 86% ,74%,86%,86% by using Logistic regression, Decision tree, Support Vector Machine (RBF), and Naive Bayes.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented a loan approval prediction system that uses machine learning algorithms to predict the likelihood of loan approval based on the borrower's credit history, income, and other relevant factors. We reviewed several related studies on loan approval prediction using machine learning algorithms and presented a use case and flow diagrams for our loan approval prediction system. Our system has the potential to improve the efficiency and accuracy of the loan approval process, reduce the risk of default, and ultimately benefit both the financial institution and the borrower.

REFERENCES

1. XiaojunMa, JinglanSha, DehuaWang, YuanboYu, QianYang, XueqiNiu , 'Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending', Annals of Operation research 266, 511-529(2018).
2. G Arutjothi , C.Senthamarai, 'Prediction of loan status in commercial bank using machine learning classifier', IEEE conference December 2017, DOI: 10.1109/ISS1.2017.8389442 .
3. Ashlesha Vaidya 'Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval' IEEE July 2017, DOI: 10.1109/ICCCNT.2017.8203946.
4. XiaojunMa, JinglanSha, DehuaWang, YuanboYu, QianYang, XueqiNiu , 'Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning', Electronic Commerce Research and Application volume 31, september-october 2018, page 24-39.
5. X Francis Jency, V.P. Sumathi 'An Exploratory Data Analysis for Loan Prediction', International Journal of recent technology and engineering, Volume-7, issue - 4s, November 2018.
6. Giorgio Calcagnini, Rebel Cole, Germana Giombini & Gloria Grandicelli 'Hierarchy of bank loan approval and loan performance' Economia Politica, 35, 935-954, 2018.
7. J.Tejaswini, T. Mohana Kavya, R. Devi Naga Ramya, P. Sai Triveni Venkata Rao Maddumala 'Accurate loan approval prediction based on machine learning approach', Journal of engineering sciences, vol -11 issue – 4, 2020.
8. Sai Aparna Vangaveeti, Naga Likitha Venna, Prasanna Naga Sri RamyaYajamanam Kidambi, Harika Marneni, Naga Satish Kumar Maganti 'Logistic regression based on loan approval prediction', Journal of Composition Theory, vol – 13, issue – 5, 2020.
9. Min Li, Amy Mickel, Stanley Taylor 'Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines', Journal of statistics education, vol- 26, issue – 1, page 55-66 , 2018.
10. Yiheng Li, Weidong Chen 'Entropy method of constructing a combined model for improving loan default prediction', Journal of the operation research society, volume 72, issue- 5, 2021.
11. E. Chandra Blessie, R. Rekha ,'Exploring the Machine Learning Algorithm for prediction of loan sanctioning process', International Journal of Innovative Technology and Exploring Engineering, , Volume-9 Issue-1 , 2019.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details

PREDICTION OF CUSTOMER LOAN ELIGIBILITY USING RANDOM FOREST ALGORITHM

SRIKANTH EADARA1 , I.PHANI KUMAR2

#1Student, Dept of CSE, VelagaNageswaraRao College Of Engineering, Ponnur(Post),Ponnur(Md)Guntur(D.T)A. Andhra Pradesh.

#2Assoc. Professor, Dept of CSE, VelagaNageswaraRao College Of Engineering, Ponnur(Post),Ponnur(Md)Guntur(D.T)A. Andhra Pradesh

ABSTRACT_A veritably important approach in prophetic analytics is used to study the problem of prognosticating loan defaulters The data is collected from the Kaggle for studying and vaticination. Random timber models have been performed and the different measures of performances are reckoned. The models are compared on the base of the performance measures similar as perceptivity and particularity. The final results have shown that the model produce different results. thus, by using a Random timber approach, the right guests to be targeted for granting loan can be fluently detected by assessing their liability of dereliction on loan. The model concludes that a bank shouldn't only target the rich guests for granting loan but it should assess the other attributes of a client as well which play a veritably important part in credit granting opinions and prognosticating the loan defaulters

1.INTRODUCTION

Finance companies deals with all kinds of loans such as house loans, vehicle loans, educational loans, personal loans etc... And has a presence across areas such as cities, towns and village areas. A Customer-first requests for a loan and after that Finance Company validates the customer eligibility for the loan ap-provel. Details like marital status, gender, education, and number of dependents, Income, Loan Amount, credit history, and others are given in the form to fill up by the applicants. Therefore, a robust model is built taking those details as input to verify whether an applicant is eligible to apply for loan or not. The target variable here is Applicants "Loan Status" and the other variables are predictors. After building the Machine Learning model a Web Application is to be developed for a user interface that allows

the user to see instantly if he/she is eligible to get a loan by entering the given details.

2.LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company Traffic Redundancy Elimination, once these things are satisfied, then next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support.

This support can be obtained from senior programmers, from book or from websites. Before building the system we have to knownthe below concepts for developing the proposed system.

[1] S. S. Sannakki and V. S. Rajpurohit, proposed a “Classification of Pomegranate Diseases Based on Back Propagation Neural Network” which mainly works on the method of Segment the defected area and color and texture are used as the features. Here they used neural network classifier for the classification. The main advantage is it Converts to L*a*b to extract chromaticity layers of the image and Categorisation is found to be 97.30% accurate. The main disadvantage is that it is used only for the limited crops.

[2] P. R. Rothe and R. V. Kshirsagar introduced a” Cotton Leaf Disease Identification using Pattern Recognition Techniques” which Uses snake segmentation, here Hu’s moments are used as distinctive attribute. Active contour model used to limit the vitality inside the infection spot, BPNN classifier tackles the numerous class problems. The average classification is found to be 85.52%.

[3] Aakanksha Rastogi, Ritika Arora and Shanu Sharma,” Leaf Disease Detection and Grading using Computer Vision Technology &Fuzzy Logic”. K-means clustering used to segment the defected area; GLCM is used for the extraction of texture features, Fuzzy logic is used for disease grading. They used artificial neural network (ANN) as a classifier which mainly helps to check the severity of the diseased leaf.

[4] Godliver Owomugisha, John A. Quinn, Ernest Mwebaze and James Lwasa, proposed" Automated Vision-Based Diagnosis of Banana Bacterial Wilt Disease and Black Sigatoka Disease "Color histograms are extracted and transformed from RGB to HSV, RGB to L*a*b.Peak components are used to create max tree, five shape attributes are used and area under the curve analysis is used for classification. They used nearest neighbors, Decision tree, random forest, extremely randomized tree, Naïve bayes and SV classifier. In seven classifiers extremely, randomized trees yield a very high score, provide real time information provide flexibility to the application.

3.PROPOSED SYSTEM

In this project we are using machine learning algorithm called Random Forest to predict loan eligibility and to train this random forest we are using below dataset

Prediction of granting the loan to the customers by the bank is the proposed model. Classification is the target for developing the model and hence using Random Forest with sigmoid function is used for developing the model. Preprocessing is the major area of the model where it consumes more time and then Exploratory Data Analysis which is followed by Feature Engineering and then Model Selection. Feeding the two separate datasets to the model, and then preceding the model.

3.1 IMPLEMENTATION

3.1.1 RANDOM FOREST ALGORITHM

First, Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach.

The author gives 4 links to help people who are working with decision trees for the first time to learn it, and understand it well. The decision tree is a decision support tool. It uses a tree-like graph to show the possible consequences. If you input a training dataset with targets and features into the decision tree, it will formulate some

set of rules. These rules can be used to perform predictions. The author uses one example to illustrate this point: suppose you want to predict whether your daughter will like an animated movie, you should collect the past animated movies she likes, and take some features as the input. Then, through the decision tree algorithm, you can generate the rules. You can then input the features of this movie and see whether it will be liked by your daughter. The process of calculating these nodes and forming the rules is using information gain and Gini index calculations.

The difference between Random Forest algorithm and the decision tree algorithm is that in Random Forest, the process es of finding the root node and splitting the feature nodes will run randomly.

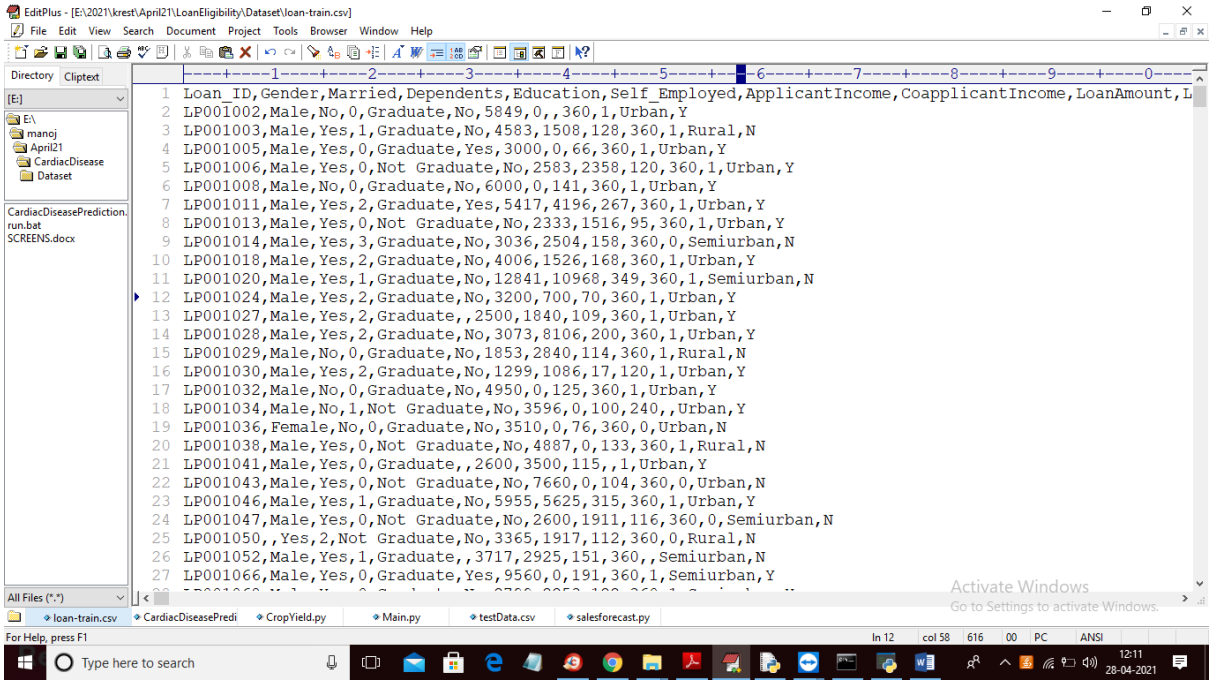
3.1.2 Why Random Forest algorithm?

The author gives four advantages to illustrate why we use Random Forest algorithm. The one mentioned repeatedly by the author is that it can be used for both classification and regression tasks. Overfitting is one critical problem that may make the results worse, but for Random Forest algorithm, if there are enough trees in the forest, the classifier won't overfit the model. The third advantage is the classifier of Random Forest can handle missing values, and the last advantage is that the Random Forest classifier can be modeled for categorical values.

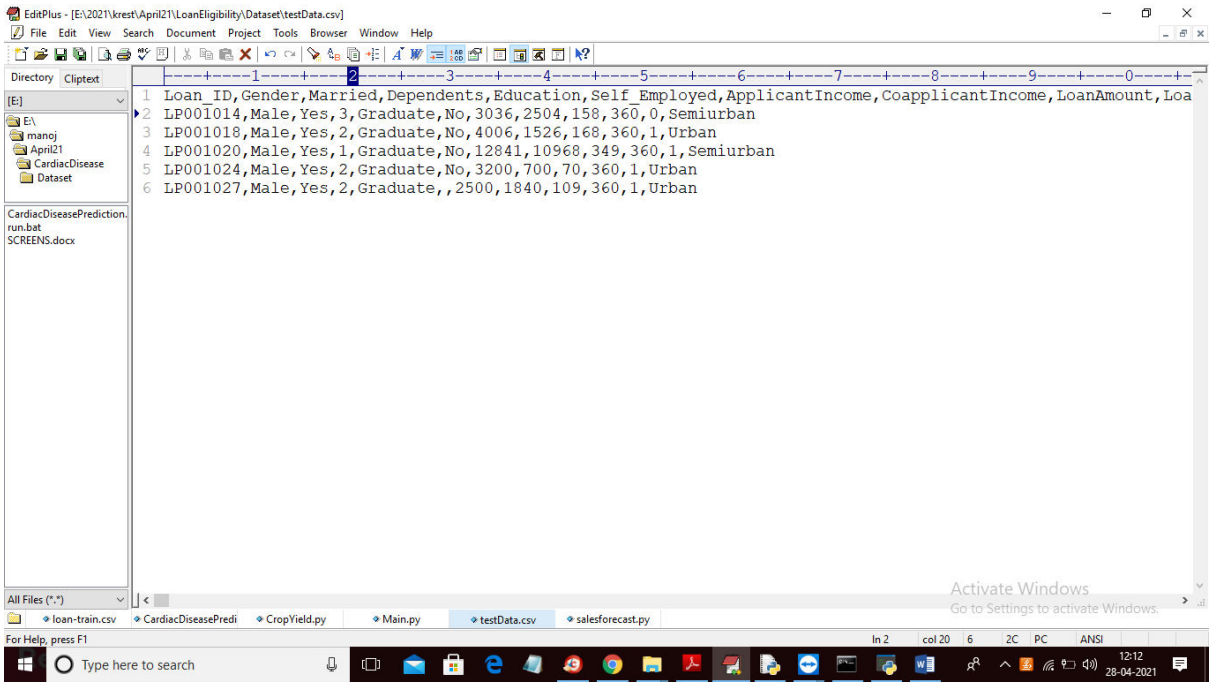
4.RESULTS AND DISCUSSION

4.1 DATASET

In this project we are using machine learning algorithm called Random Forest to predict loan eligibility and to train this random forest we are using below dataset

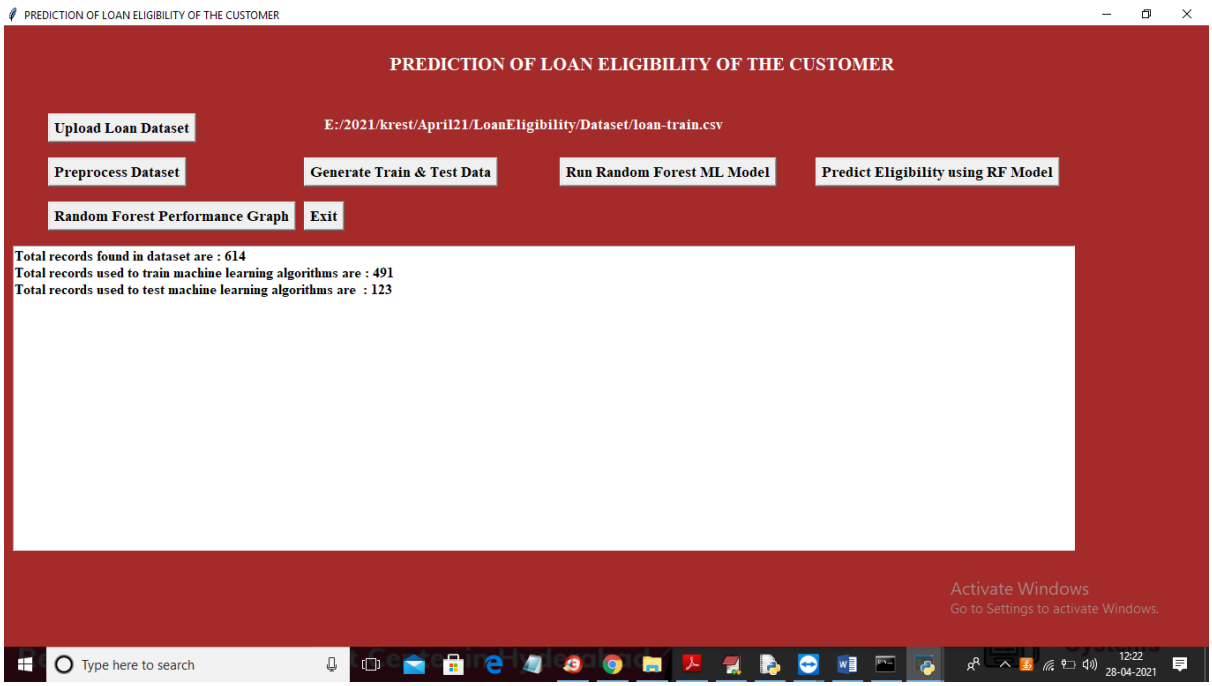


In above dataset in first row we can see dataset column names and in other rows we have dataset values and in last column we have class label as Y or N where Y means eligible and N means not eligible and now we used above dataset to train machine learning model and after training we will upload test dataset and then application will predict class label Y or N and below is test dataset screen shots



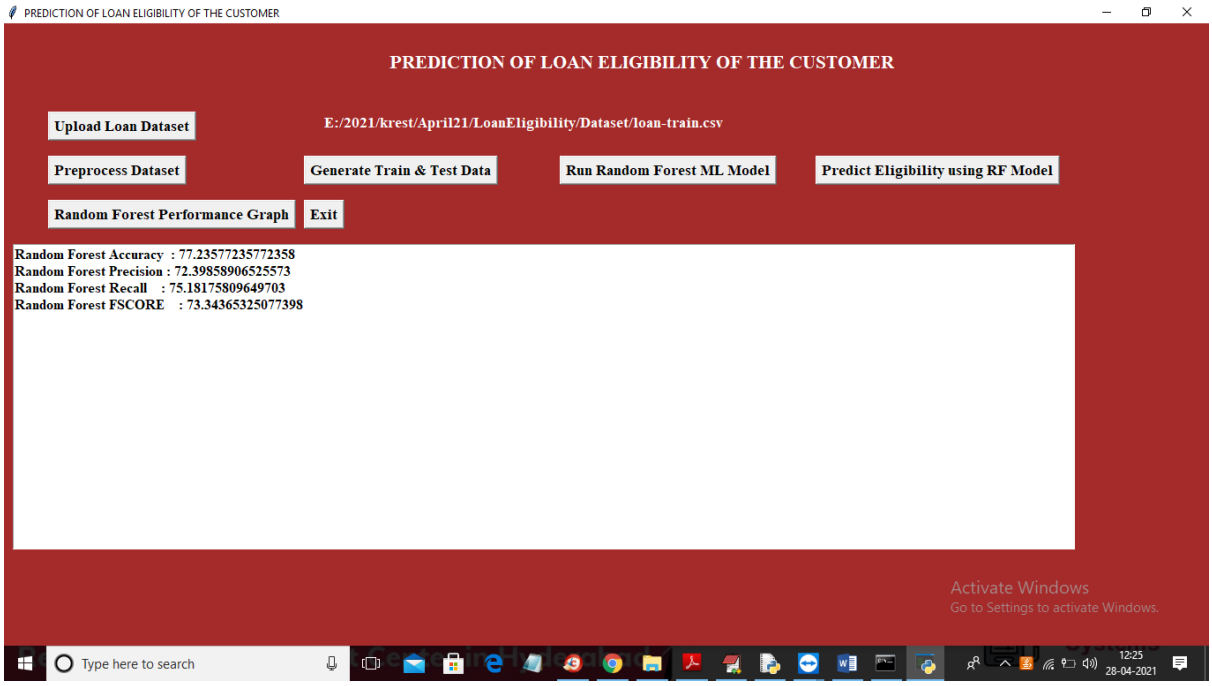
In above test data we don't have any N or Y class label and by analysing above records machine learning will predict eligibility.

4.2 RESULTS

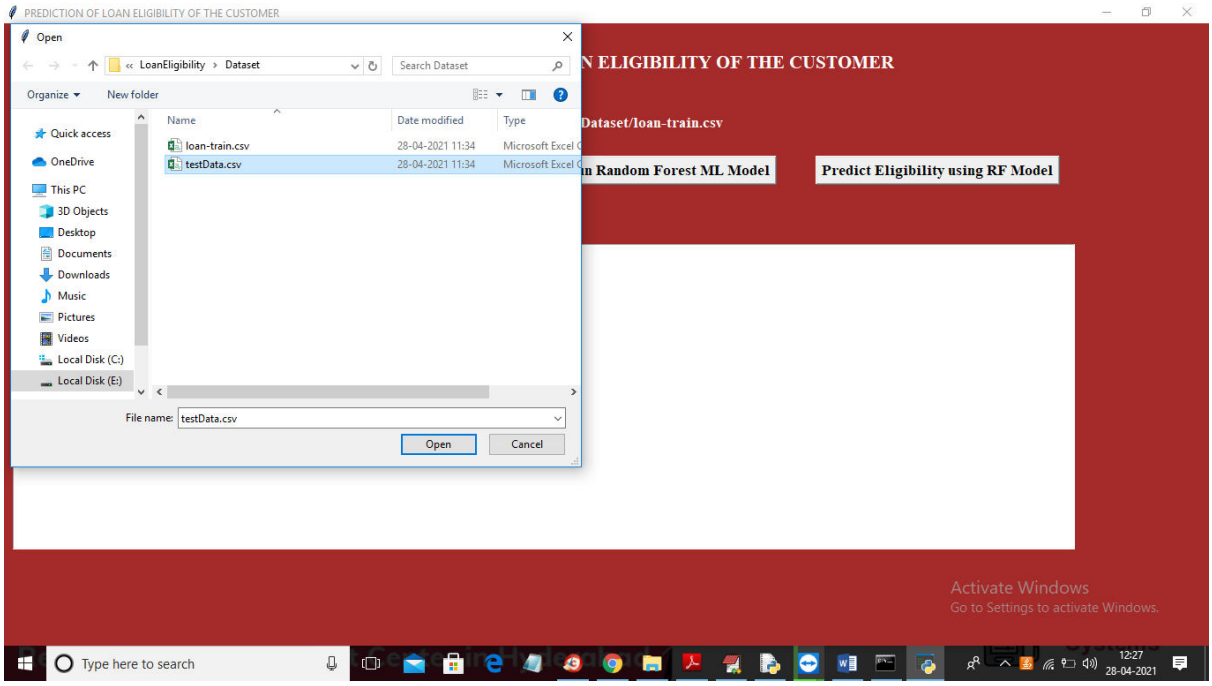


In above screen dataset contains 614 records and using 491 records to train ML and 123 records to test ML accuracy. In below graph we can see importance of each attribute with other attribute by using graph correlation metric

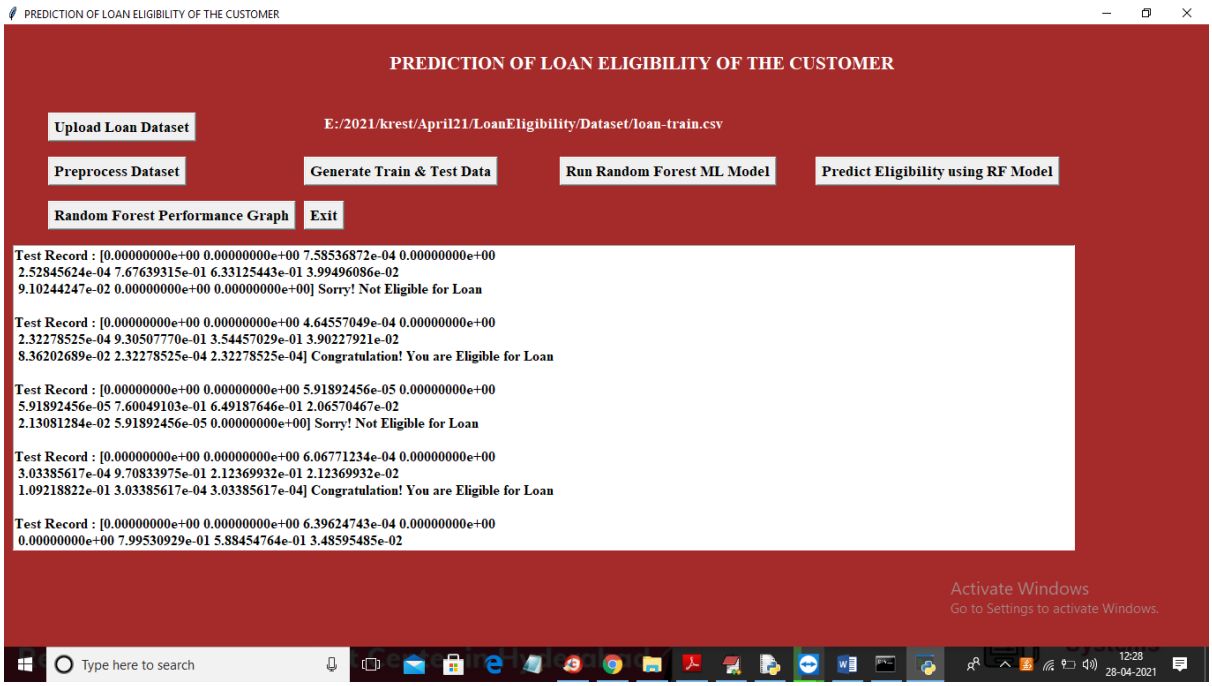
In above graph whatever column in x-axis and y-axis having value >0 will be consider as important features or column. Now click on 'Run Random Forest Ml Model' to build random forest model on above dataset



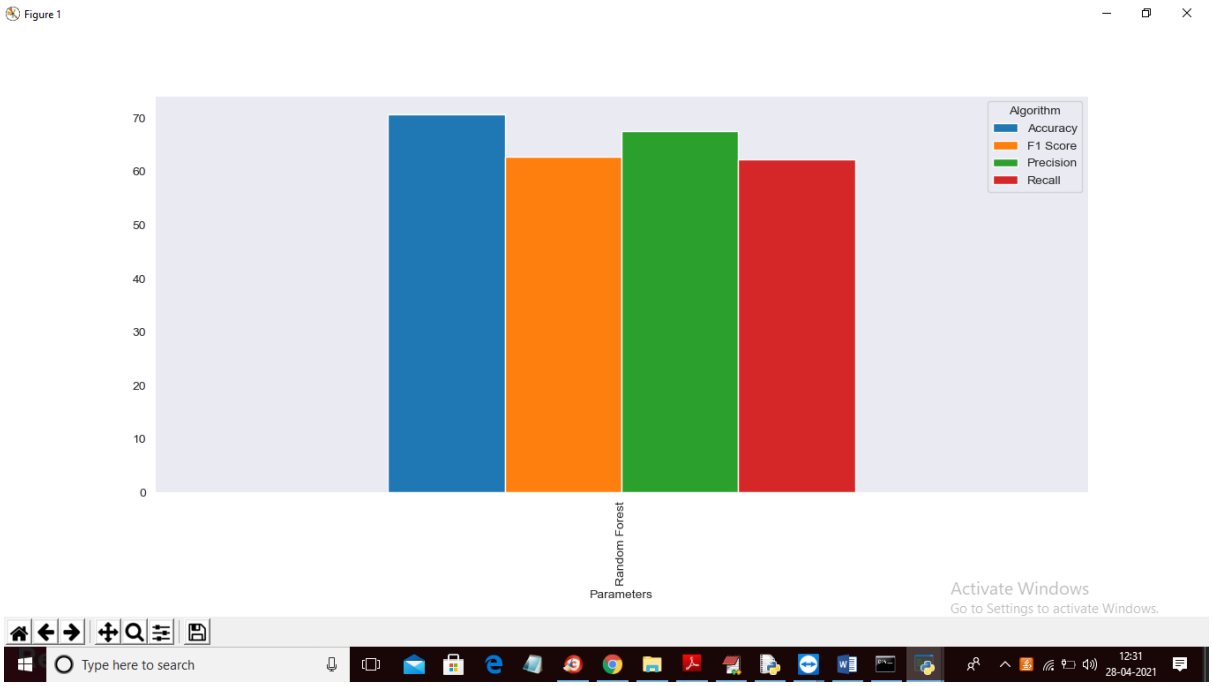
In above screen random forest model generated with 77% accuracy and we can see its precision, recall and FSCORE value and now click on 'Predict Eligibility using RF Model' button to upload test data and perform eligibility prediction



In above screen selecting and uploading 'testData.csv' file and then click on 'Open' button to load test data and then will get below prediction result



In above screen in square bracket we can see normalized test values and after square bracket we can see the prediction result as eligible or not eligible. You can scroll down above text area to view all predicted records and now click on 'Random Forest Performance Graph' button to get below graph



In above graph we can see accuracy, precision, recall and FSCORE values of random forest and graph y-axis represents %value where accuracy got 80% and Precision got 65%. Each metric bar colour name you can see from top right side

5.CONCLUSION

As a result, the proposed model automates the procedure of determining the creditworthiness of the applicant. It focuses on data comprising the essential points of loan applicants. The random forest model is employed in this system. Random forest analysis is a supervised learning method in Machine Learning. As a consequence, it is useful for forecasting the proper result in the current world scenario and also helps the bank to put the money in the right hands and also helps the people in receiving loan in a lot faster pace. The key advantage of this approach is that it provides greater accuracy.

REFERENCE

[1] Toby Segaran, “Programming Collective Intelligence: Building Smart Web 2.0 Applications.” O’Reilly Media.

[2] Drew Conway and John Myles White,” Machine Learning for Hackers: Case Studies and Algorithms to Get you Started,” O’Reilly Media.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman,"The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer ,Kindle

[4] PhilHyo Jin Do ,Ho-Jin Choi, "Sentiment analysis of real-life situations using loca- tion, people and time as contextual features," International Conference on Big Data and Smart Computing (BIGCOMP), pp. 39–42. IEEE, 2015.

[5] Bing Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, May 2012.

[6] Bing Liu, "Sentiment Analysis: Mining Opinions, Sentiments, and Emotions," Cambridge University Press, ISBN:978-1-107-01789-4.

[7] Shiyang Liao, Junbo Wang, Ruiyun Yu, Koichi Sato, and Zixue Cheng, "CNN for situations understanding based on sentiment analysis of twitter data," Procedia computer science, 111:376–381, 2017.CrossRef.

[8] K I Rahmani, M.A. Ansari, Amit Kumar Goel, "An Efficient Indexing Algorithm for CBIR,"IEEE- International Conference on Computational Intelligence & Communication Technology ,13-14 Feb 2015.

[9] Gurlove Singh, Amit Kumar Goel ,"Face Detection and Recognition System using Digital Image Processing" , 2nd International conference on Innovative Mechanism for Industry Application ICMIA 2020, 5-7 March 2020, IEEE Publisher.

[10] Amit Kumar Goel, Kalpana Batra, Poonam Phogat," Manage big data using optical networks", Journal of Statistics and Management Systems "Volume 23, 2020, Issue 2, Taylors & Francis.

AUTHOR'S PROFILE

SRIKANTH EADARA pursing M. Tech in Computer Science and Engineering from Velaga Nageswara Rao College Of Engineering, Ponnur. Affiliated to JNTUK, KAKINADA

I.Phani Kumar ,Qualifications: Ph.D ,M.Tech, having 13 years of *teaching experience,present he is working as Assoc.Prof in Velaga Nageswara Rao College of Engineering,Ponnur,Guntur(D.t),A.P,maild:phanikumar148@gmail.com,

International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1952198

Loan Default Identification and its Effect

Gopal Choudhary[1], Yash Garud[1], Akshil Shetty[1], Rumit Kadakia[1], Sonali Borase[2]

[1]Department of Computer Science and Engineering, MPSTME, NMIMS, Shirpur, District: Dhule, Maharashtra, India

[2]Assistant Professor, Department of Computer Science and Engineering, MPSTME, NMIMS, Shirpur, District: Dhule, Maharashtra, India

ABSTRACT

Now a days banking sector is on boom everyone is applying for loan but banks have limitation that they have limited assets so they can provide loan to limited loan applications but when they provide loan, they must assure that loan is being granted to only genuine customers. So, this paper focuses on we will try to lessen the uncertainty factor and assure the loan approval to genuine customers only and save the bank assets. That is performed by way of mining the massive data of the earlier data of the human beings to whom the loan become acknowledged earlier than and on the idea of those records/reviews the machine was skilled the use of the system mastering version which provide the maximum correct result. The main focus of the paper will be on the loan to be approved of those customers only who will be able to pay it back.

Keywords : Loan, Machine Learning, Training, Testing, Prediction.

I. INTRODUCTION

When a client takes loan from a bank and financers, he needs to pay it on time but he not paid the loan amount to that bank or financer and default that loan amount. Now that same customer willing is to take loan from more financers and banks and default that loan amount. So, these needs to be stopped somewhere.

Loan portfolio are the largest source of revenue of most of the banks and financers but it's also leads to NPA (Non-Performing Assets) or end up with loan default. And this affect the financial performance of the financers and banks. The loan defaulters are trying to fool multiple financers and banks at a time. So, they needed to be stopped at right time.

The very obvious effect of loan default is that it causes the monitory growths of financers and banks. This result in the increase in the rate of interest on the loan, and this leads to overall failure to the economic growth.

II. METHODS AND MATERIAL

The implementation of the project can be divided into two parts.

1) Data analysis and data cleaning

Selecting relevant features

In machine learning and information, feature selection, additionally known as variable choice, attribute selection or variable subset choice, is the procedure of choosing a subset of applicable capabilities (variables, predictors) to be used in model creation. function choice techniques are used for four motives:

- simplification of models to cause them to simpler to interpret with the aid of researchers/customers
- reduces training time
- To avoid the curse of dimensionality

- enhanced generalization by reducing overfitting (formally, reduction of variance)

Null value imputation

Missing data can occur while no records is supplied for one or greater items or for a complete unit. missing records is a totally big problem in actual existence scenario. missing information can also talk to as NA (now not available) values. In Data Frame now and again many datasets sincerely arrive with lacking information, either as it exists and turned into no longer amassed or it in no way existed. for instance, suppose unique user being surveyed may also pick out now not to share their income, a few users may select no longer to proportion the cope with in this manner many datasets went lacking.

Handling Outliers

An outlier can be referred as a data point this is remote from different comparable factors. they'll be due to variability inside the measurement or might also suggest experimental errors. If feasible, outliers must be excluded from the data set. but, detecting that anomalous times might be very difficult, and isn't usually viable.

Different methods of dealing with outliers:

- Univariate method: This technique looks for data points with extreme values on one variable.
- Multivariate method: right here we search for unusual combinations on all the variables.
- Minkowski error: This approach reduces the contribution of ability outliers within the training method.

Training a machine learning model:

Logistic Regression

Rather than predicting precisely 0 or 1, logistic regression generates a possibility—a value between zero and 1, exclusive. for instance, don't forget a logistic regression version for unsolicited mail detection. If the version infers a cost of 0.932 on a particular electronic mail message, it implies a 93.2% chance that the email message is junk mail. extra precisely, it way that inside the limit of infinite schooling examples, the set of examples for which the model predicts 0.932 will certainly be spam ninety 3.2% of the time and the last 6. 8% will nolonger.

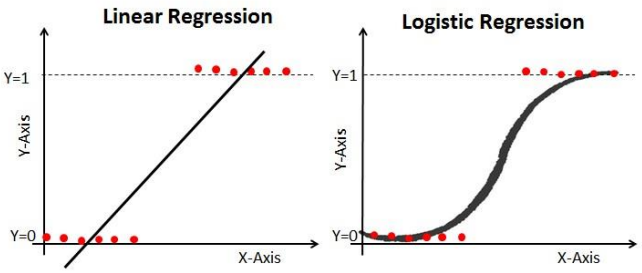


Fig 1. Logistic Regression

Random Forest

Random forest is very easy and flexible, easy to apply ML algorithm that produces, even without hyper-parameter tuning, a top-notch result maximum of the time. it is also one of the maximum used algorithms, as its simplicity and the truth that it can be used for both category and regression duties. in this put up, you'll analyze, how the random forest algorithm works and numerous different vital things approximately it.

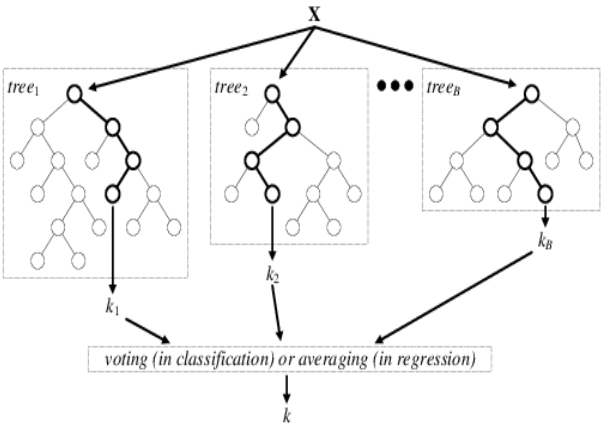


Fig 2. Random Forest

Random Forest is a supervised studying set of rules. Like you can already see from its call, it creates a forest and creates it somehow random. The forest it builds, is an ensemble of decision tree, maximum of the time skilled with the "bagging" technique. the overall concept of the bagging technique is that a combination of learning models increases the overall result.

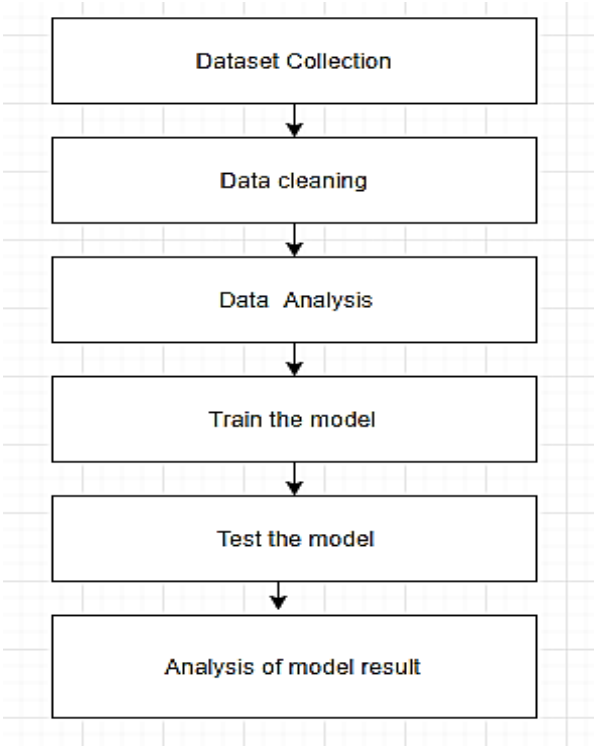


Fig 3. Flow Diagram of the System

III. EXPERIMENTAL RESULT

Data Set

The data which has been trained is being applied to machine learning model, whenever the new customer enters the detail in application form that act as test data set. So, after testing operation is performed, then afterwards the model will predict that will the new customer will be able to pay the loan on time or not on the process of loan approval on the basis of the training dataset

Table 1

Attribute Name	Description	Type
Cust_id	It's a Unique id	Integer
loan_amnt	The Loan amount that has been sanctioned	Integer
term	Duration time of Loan	Character
int_rate	The annual interest rate on the loan amount	Integer
installment	The emi that are monthly paid by the borrower	Integer
grade	LC assigned loan grade	Character
annual_inc	annual_inc: The self-claimed annual income given by the borrower during registration.	Integer
Credit_History	credit history meets guidelines	Integer
Property_Area	Type of area whether it is urban or rural	String
loan_status_coded	Loan Approved(Y/N)	Integer

Accuracy Measure

The model is able to predict 61% of defaulters this can be observed by ROI (Return on Investment). To calculate ROI here we have used fully paid, charged off and loan in grace period So how to calculate ROI = 'total_payment' / 'funded_amount' ROI calculated without using this model=-4.57 and after using this model ROI calculated is 2.22

Grade wise analysis of ROI

As you can see in below table the ROI of C, D, E, G become positive using this model.

	ROI	% Picked	% Default	ROI_w/o_model	% Picked_w/o_model	% Default_w/o_model
A	0.0250712	99.9774	8.64002	0.025071	100	8.63806
B	0.0200738	96.1866	16.1057	0.0152713	100	16.7564
C	0.017765	53.9985	18.7983	-0.0447373	100	25.9779
D	0.0349775	24.597	20.8825	-0.096151	100	34.7948
E	0.0703003	7.08117	22.2561	-0.143831	100	42.5518
F	-0.0234227	3.26508	22.0339	-0.16447	100	47.316
G	0.0961287	2.6694	38.4615	-0.163535	100	50.1027

Table 2. ROI Index Table

IV.CONCLUSION

From a proper analysis of tremendous factors and constraints at the aspect, it is able to be appropriately concluded that the product is a distinctly green aspect. This utility is running nicely and meeting to all Banker necessities. This factor can be without difficulty plugged in lots of different systems. There have been numbers instances of computer system faults, mistakes in content and most essential weight of features is fixed in automatic prediction machine, so in the close to future the so –known as software program may be made extra at ease, dependable and dynamic weight adjustment. In close to future this module of prediction can be combine with the module of automatic processing device. the device is trained on antique schooling dataset in destiny software program may be made such that new testing date have to additionally take part in education records after a few fix times.

V. REFERENCES

[1]. A. Malali, "Predicting Loan Outcomes using Machine Learning," Aug 25, 2016.

[2]. A. J. Hamid, "DEVELOPING PREDICTION MODEL OF LOAN," An International Journal (MLAIJ) Vol.3, No.1, March 2016.

[3]. A. Goyal, "Loan Prediction Using Ensemble Technique," International Journal of Advanced Research in Computer and Communication Engineering, March 2016.

[4]. A. Goyal, "A survey on Ensemble Model for Loan Prediction," International Journal of Engineering Trends and Applications (IJETA) – Volume 3, Feb 2016.

[5]. R. S. T. Sivasree M S, "Loan Credibility Prediction System," International Journal of Engineering Research & Technology (IJERT), Sept,2015.

[6]. S. G, "Credit Risk Analysis and Prediction," International Journal of Engineering and Technology (IJET) , 2014.

[7]. E. I.Altman, "Predicting performance in the savings and loan association industry," Journal of Monetary Economics, oct,1977.

[8]. M. R. Islam, "A DATA MINING APPROACH TO PREDICT," International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, March,2015.

[9]. D. K. Kavitha, "Classifying Data and Predicting Risk towards Multi -," International Journal of Advance Research in, Feb,2016.

Cite this article as :

Gopal Choudhary, Yash Garud, Akshil Shetty, Rumit Kadakia, Sonali Borase, "Loan Default Identification and its Effect", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 865-868, March-April 2019. Available at doi :
https://doi.org/10.32628/CSEIT1952198
Journal URL : http://ijsrcseit.com/CSEIT1952198

2021 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER) | 978-1-6654-1244-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/DISCOVER52564.2021.9663662

Comparative Analysis of Bank Loan Defaulter Prediction Using Machine Learning Techniques

Spoorthi B, Shwetha S. Kumar, Anisha P Rodrigues, Roshan Fernandes, Balaji N
NMAM Institute of Technology, Nitte
Udupi, India
spoorthisapalya1998@gmail.com, shwetaskumar24@gmail.com, anishapr@nitte.edu.in,
roshan_nmamit@nitte.edu.in, balaji.hiriyur@gmail.com

***Abstract*—Nowadays, there are numerous risks identified with the banking sector regarding giving loans to the clients and for the individuals who get the loan. The examination of risk in bank credits needs to understand what is the reason for this risk. Likewise, the quantity of exchanges in the financial a rea is quickly developing and information volumes are accessible which address the client's conduct, and the risk of giving loans are expanded. The objective of this paper is to discover the nature or details of the clients who are applying for the loan. This paper proposes a comparative study of three machine learning models, namely, Random Forest, Naive Bayes (Gaussian model, Multinomial model, and Bernoulli Model), and Support Vector Machine (Linear kernel, Gaussian RBF kernel, and Polynomial kernel), to predict whether a customer may get a loan or not. In this paper, we analyze the evaluation parameters, namely, classification a ccuracy, p recision, r ecall, a nd F 1-Score f or these machine learning models to foresee which model is best suitable for predicting a loan.**

Keywords: Bank Loan Defaulter Prediction, Random Forest, Naive Bayes, Support Vector Machine.

I. Introduction

Today, in the world most people depend on a bank to lend them a loan for a different purpose may be for building a home, buying land, education purposes, and for some personal goals. A loan is a significant k ind o f r evenue f or t he fi nancial area just as well as a risk for banks. Huge parts of a bank's resources straightforwardly come from the interests acquired on advances given. Loan Prediction is useful for representatives of banks just as for the candidate. It can give extraordinary benefits t o t he bank.

While before, banks used to enlist profoundly proficient people whose sole reason for existing was to assess candidates and after close survey, chooses and tell whether an applicant was qualified f or a l oan o r n ot. The g reat g oal i n t he financial sector is to contribute their resources in safe hands. Through this framework, we can foresee if that specific c andidate for loan sanction is a non-defaulter and the entire interaction of approval from the bank is predicted by using the machine learning models, thereby making the entire process easy and effective.

The primary objective of the proposed work is to understand the various general parameters used by the banks to evaluate a candidate's profile to grant him the loan amount. These parameters are carefully analyzed in the proposed work and the necessary parameters required by the machine learning models are retained.

The entire process involves training the data set and predicting if the loan can be given to the individual based on different attributes / features. The expectation model not just aids the candidate yet, also, helps the bank by limiting the risk and decreasing the number of defaulters. In the current situation, a loan should be affirmed physically by an agent of the bank or manager, which implies that the individual will be answerable if the client is qualified for the loan and computing the risk related to it. It is time-consuming if it is done manually by a human and may lead to a lot of errors. So, a loan prediction model is required that can foresee rapidly whether the credit can be passed or not with minimal measure of the possibility of risk and errors.

The paper is organized as follows: Section 2 discusses the existing work, section 3 describes the methodology, section 4 gives the results and analysis of the proposed work, section 5 concludes the proposed work and discusses the future scope.

II. Related Work

Aboobyda Jafar Hamid and Tarig Mohamed Ahmed [1] utilized data mining techniques such as decision tree (J48) integrated with Naive Bayes and Baysenet to show how predictive approaches may be applied in real-world loan approval. To adopt this strategy, they employed the Weka tool and examined a dataset containing eight attributes: gender, employment, age, credit amount, credit history, housing, and income. After applying these models to the dataset, they concluded that the decision tree resulted in better accuracy for loan prediction.

S. Vimala and K.C. Sharmila [2] applied an NB technique combined with K-nearest neighbor and the Binning method in their study. They used this method by analyzing a dataset with seven factors, including age, income, occupation, a current loan, including the duration, number, and approval status. Experimentation demonstrated that combining the kNN and Binning algorithms with the NB produced better loan sanctioning mechanism prediction.

The evaluation and work done by Arora, Nisha and Pankaj Deep Kaur [3] use Bolasso to choose the largest credits based on their strength and to examine how accurately the results

can be foreseen in group computations such as Random Forest, SVM, Naive Bayes, and KN Nearest Neighbors (KNN). Bolasso enabled Random Forest algorithm (BS-RF) is used to provide the best results in the risk assessment of credit and enhances accuracy by employing sophisticated component choices.

The research and work done by Kumar Arun, Garg Ishaan, Kaur sanmet [4] used various approaches in their paper for loan prediction. For the observation and loan prediction processes, a dataset with ten characteristics was used. The subprocesses comprised data gathering, feature selection, training, testing, and performance analysis.

Lin Zhu et al. and Nazeeh Ghatasheh [5] worked on several models to construct a loan default prediction model. In their study, they determined that Random Forest outperforms other algorithms such as logistic regression, decision-making bodies, and precision vector machines. According to the conclusions of this paper, the random forest algorithm is among the best options to estimate credit risk.

A novel ensemble methodology has been proposed by Maher Alaraj, Maysam Abbod, and Ziad Hunaiti for the unique type of client mortgages [6]. This ensemble methodology focuses on the neural network. They suggest that the proposed approach provides better results and accuracy than a single classification and another alternative.

Aditi Kacheria and colleagues used the Naive Bayesian method [7]. They also used the k-NN and binning algorithms to enhance the quality of the data and classification accuracy. KNN was used to cope with missing values, and the binning technique was used to remove abnormalities.

Addo P M, Guegan D, and Hassani B [8] selected four different models in their paper, they are Logistic Regression, Random Forest, Gradient Boosting model, and Multilayer Neural Network models. They also demonstrated the importance of data quality assurance, such as data analysis and cleansing before modeling to eliminate redundant variables, using these models. The study also found that the algorithm and feature selection are two important factors to consider when selecting whether or not to grant a loan to a person.

In 2017, Goyal and Kaur [9] proposed the concept of loan prediction using certain methods for machine learning. There are 13 characteristics in the dataset. The results of this article reveal that TGA was superior to other models. Several researchers have contributed to predict the home loan defaulters using specific machine learning techniques [10][11]12][13]. Machine learning approaches are widely used in sentiment analysis, chatbot designs, and many more [14][15][16].

The existing works have applied different machine learning techniques on the loan prediction data set and analyzed the various evaluation parameters. None of the researchers have used the variants of Naive Bayesian classifier and Support Vector machine techniques and compared the results. So, in the proposed work, we have analyzed the bank loan data set using the variants of Naive Bayesian classifier techniques, namely, Gaussian NB, Multinomial NB, and Bernoulli NB, and the variants of Support Vector Machine, namely, linear kernel, Gaussian RBF kernel, and Polynomial Kernel. The results obtained after performing the analysis are compared to each other.

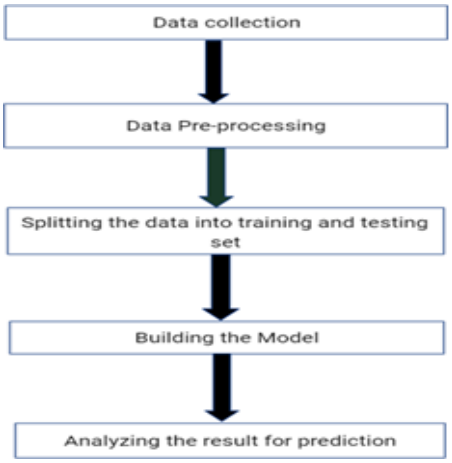


Fig. 1. Bank Loan Defaulter Prediction Methodology

III. Methodology

In the proposed work we have considered 983 bank loan sample users' data set. It consists of 13 parameters shown in Table 1. The description of the parameters is given below. The source of the Data set is taken from Kaggle.

Loan Approval Status: Approximately two-thirds of applicants have been approved for a loan. Sex: There are almost three times as many men as women. Marital Status: Married applicants are more likely to be given loans; married applicants make up two-thirds of the population in the sample. Dependents: The bulk of the population has no dependents and is, therefore, more willing to take a loan.

Education: Approximately $\frac{5}{6}^{th}$ of the population is a graduate, and graduates have a greater rate of loan approvals.

Employment: $\frac{5}{6}^{th}$ of the population is not self-employed. Property Area: Semi-urban candidates are more likely to be approved for loans. Total Income: It is the combination of both applicant income and co-applicant income. The probability of loan getting approved will be high if the total income is high for that particular client. Loan Amount: 40% of the candidate have applied for minimum loan amount and chances of being approved is high. Credit History: An applicant with a credit history is considerably more likely to be accepted. Loan Amount Term: The preponderance of loans are taken out for 360 months (30 years).

The figure 1 depicts the working of the proposed model.

The steps involved in the data processing are:

1) Importing the data-set.
2) The data pre-processing stage involves, identifying the Null/missing values in the data set and replacing them with the mean value of the existing attribute values.
3) The data pre-processing stage also involves converting all the categorical values into continuous values.

TABLE I
DATA-SET VARIABLES ALONG WITH DESCRIPTION AND TYPE

Sl. No.	Variables	Description
1	Loan ID	Unique Loan ID
2	Gender	Male / Female
3	Married	Applicant Married (Yes / No)
4	Dependents	Number of Dependents
5	Education	Applicant Education (Graduate / Under-graduate)
6	Self Employed	Self Employed (Yes / No)
7	Applicant Income	Applicant Income
8	Co-applicant Income	Co-applicant Income
9	Loan Amount	Loan amount in thousands
10	Loan Amount Term	Term of loan in months
11	Credit History	Credit history meets guidelines
12	Property Area	Urban / Semi Urban / Rural
13	Loan Status	Loan approved (Yes / No)

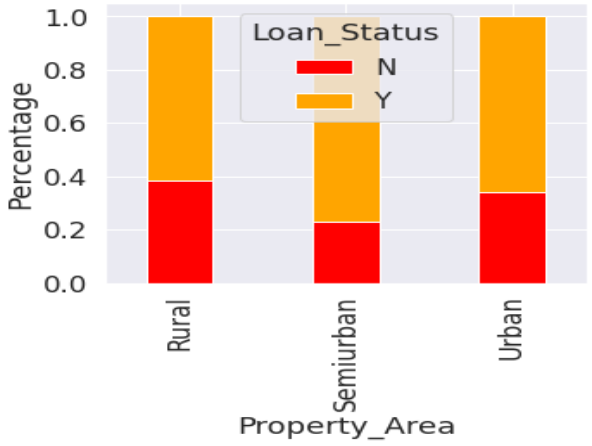


Fig. 2. Property-Area.

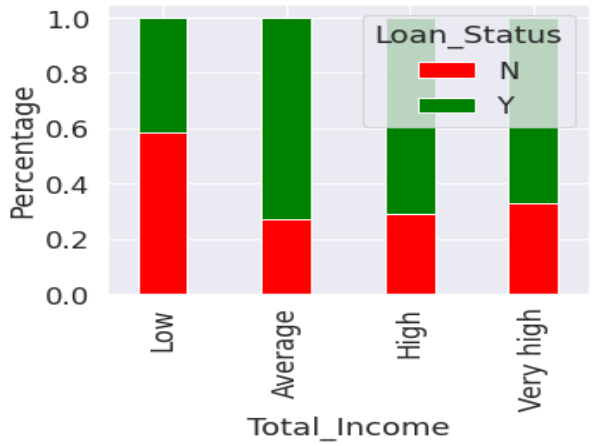


Fig. 3. Total Income.

Fig. 4. Loan Status

4) Splitting the data-set into 70% of training set and 30% of testing set.
5) Build the model and analyze the evaluation parameters of the models being used.

1) Data Set Analysis: The data set used for predicting the bank loan defaulter has been analyzed. Figure 2 gives the analysis of the Property Area parameter against the target parameter - loan status, having values Yes or No. The possible values for this parameter include rural, semi-urban, and urban. The analysis shows that the applicant living in semi-urban has a high chance of loan approval compared to the rural and urban.

The applicant and co-applicant income might have a great impact on the loan status as few of the applicant may not have co-applicant. Figure 3 gives the analysis of the total income parameter against the target parameter - loan status, having values Yes or No. The possible values for this parameter include low, average, high and very high. To solve this, we combine both the attributes. The combined effect of Applicant income and co-applicant income is total income. The chances of a loan getting approved for the applicant having a high total income is high compared to the applicant having a low and average total income.

The chances of a loan getting approved for a particular application depends on different attributes like salary, property of the applicant, the applicant history and many more.

Figure 4 shows the loan approval status based on loan amount term for the considered data-set whether the loan might be approved and unapproved based on the regular payment.

A. Machine Learning Models

For the prediction of application that can be used in android applications, two machine learning classification models are used. These models are also available in the GNU GPL - licensed open-source software R. There are three machine learning models are used in this paper they are:

1) **Random Forest** - Random forest is a monitored learning algorithm that produces a huge number of decision-making trees in training and produces a class that is the class mode or mean of the individual trees' predictors. It's used for things like classification and regression, among other things. A Decision Tree is a decision-

making tree frame- work. Each of its non-leaf nodes constitutes a function test where every branch stores a category and each leaf node shows the feature characteristic contribution across a range of values. The decision tree begins with a root node, then tests the appropriate function characteristics in the group to be categorized, then picks output branches based on their values before entering the leaf node, where the categorization is termed the decision result.

2) **Naive Bayes** - The Bayes theorem supports this classification strategy, which ensures predictor independence because the presence of one feature in one class does not imply the presence of another character in another class. As a result, the likelihood or probability density of characteristics X (feature matrix) was given class Y (response vector) forms the foundation of Naive Bayes classification. Naive Bayes models are classified into:

   a) **Gaussian Model** - Gaussian Naive Bayes is based on Bayes' Theorem and proposes that classifiers must be unbiased of one another. It accepts continuous values and implies that each class is evenly distributed.
   b) **Multinomial Model** - This variant is an event-based model with features represented as vectors, with sample representing the frequency with which specific events have occurred.
   c) **Bernoulli Model** - This category expects data to be expressed as binary-valued feature vectors; that is, numerous features may exist, but all are considered to be a binary-valued variable.

3) **Support Vector Machine** - The data points closest to the hyperplanes or data points that are termed support vectors if eliminated, are the direction of the hyperplane. As a result, they are referred to as vital components of data collection. The Kernel Function is a way of taking data as input and converting it into the kind of data processing needed. "Kernel" is utilized since the Support vector Machines employs a collection of arithmetic functions to alter data by providing a window. In this paperwork, we have used linear kernel, Gaussian RBF kernel and Polynomial Kernel.

   a) **Linear Kernel** - Linear kernel is defined by the dot product of two vectors x and y where x is the input parameter and y is the output parameter.
   b) **Gaussian RBF Kernel** - It is used to perform the transformation when there is no prior knowledge about data. The radial basis method is added to improve the transformation.
   c) **Polynomial Kernel** - It represents the similarity of vectors in training set of data in a feature space over polynomials of the original variables used in kernel. The output of the polynomial kernel function depends on the area of the vector field in low dimensional space.

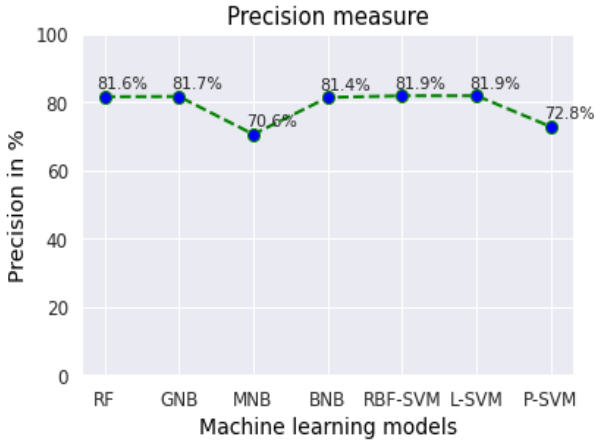


Fig. 5. Precision values of classification models

IV. Results and Discussion

Cleansing the information, processing the information, input the missing values, exploratory evaluation on statistics, version production and its evaluation are the basic steps accompanied in the analytical technique. Evaluation are done for these samples using different classification method to find out which model produces an accurate result for the data samples. It is based on its accuracy that the model can reliably identify the beneficial effects of all positive predictions it produces. When the classes are extremely uneven, the precision score is a good indicator of prediction success.

The precision values of linear SVM kernel show that out of 134 actual positive values,132 is predicted accurately and out of 51 actual negative value, 29 are predicted incorrectly. The Figure 5 shows the comparison of precision values for the different classification model.

The model's ability to correctly estimate positives from real positives is measured by its recall score. This is in contrast to precision, which evaluates how many positive predictions are made by models out of all positive assertions. The recall values of SVM RBF kernel shows that out of 134 actual positive values, all 134 is predicted correctly and out of 51 actual negative value,0 is falsely predicted as negative. Figure 6 shows the comparison of recall values for different models.

The F1 score is a function of precision and recalls score that represents the model's score. the precision value of the SVM linear kernel is 0.819 and the recall value is 0.985. Using these values the F1 score of the SVM linear kernel is evaluated. The figure 7 shows that the F1 score values of the SVM linear kernel are better compared to other models.

The model's ability to properly identify both positives and negatives out of all the predictions is represented by the model's accuracy score. The SVM linear kernel performs better compared to the random forest, Gaussian NB, Multinomial NB, Bernoulli NB, SVM RBF Kernel and SVM Poly Kernel. SVM Classifier has a faster convergence than other classifiers, and it can be utilized even if the number of positive and negative instances are not equal, because it can normalize

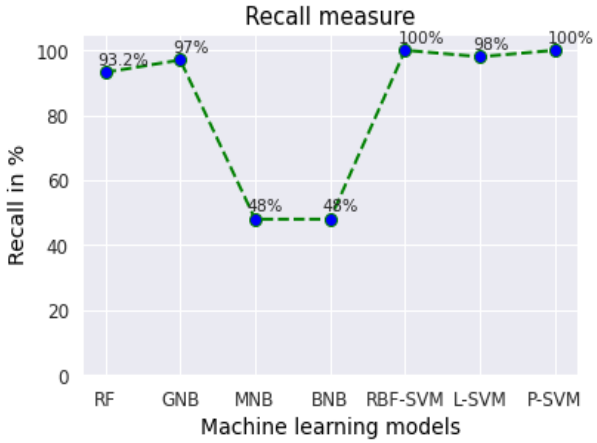


Fig. 6. Recall values of classification models

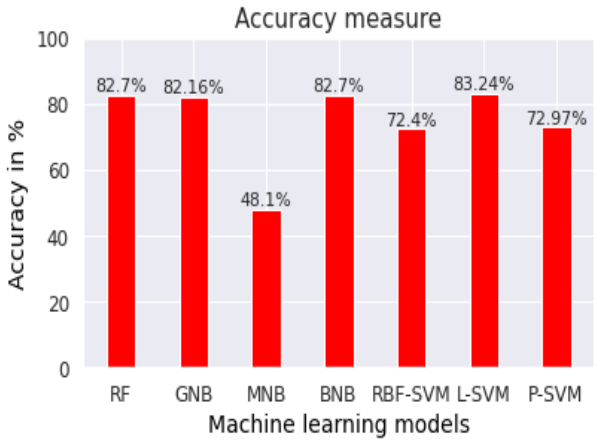


Fig. 8. Accuracy of classification models

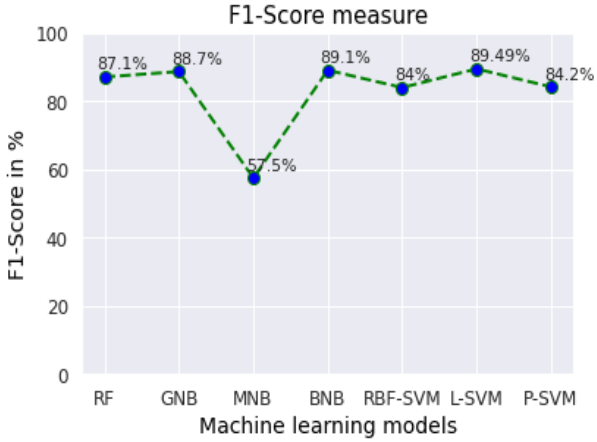


Fig. 7. F1 score values of classification models

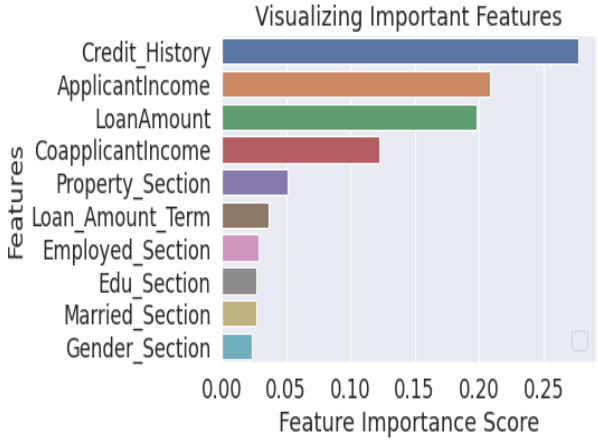


Fig. 9. Visualizing Feature Importance

the data. 83.24 is the greatest possible accuracy through using the SVM model with the given data-set. The prediction accuracy of MNB algorithm is lower than the other algorithms. The main characteristic of Naive Bayes is the assumption of independent predictors. If a categorical variable contains a category (in the test data set) that was not observed in the training data set, the model will assign a 0 (zero) probability and thus unable to generate a prediction. It is also known as a bad estimator. 83.24 is the greatest possible accuracy through using the SVM model with the given data-set. Figure 8 shows the accuracy of each classification models.

A number of the following insights are advanced about loan approval through inspecting the consequences. Candidates with a '0' credit score records did not get their loans accepted, like 'zero' credit history very much less legal responsibility of the applicant which means that he's/she's having fewer chances of repaying. Moreover, candidate with few dependents and greater earnings followed by using much less quantity of loan requested got their loan sanctioned. Information such as the gender of the applicant and the education of the applicant has been considered as the least crucial attributes.

The relevance of each field on each prediction provided by classification or regression analysis is shown by the feature importance values. The data-set consists of 13 features. The feature importance is analyzed using Random Forest algorithm and are calculated based on the training data. Table II gives the feature importance score obtained. After being fit, the model provides a feature importance's property that can be accessed to retrieve the relative importance scores for each input feature. Figure 9 shows which feature have more weight age while evaluating the model. Figure 9 shows which feature have more weight age while evaluating the model. Credit history score is above 0.25 which is the highest score compared to the gender, education, employed, Loan amount term whose values ranges between 0.00 to 0.05.

V. Conclusion and Future Scope

The paper's key aim is to study the dataset using different machine learning models in order to predict the chances of loan getting approved for that particular candidate. Different machine learning models have been used in order to determine which algorithm performs the best for the dataset. Apart

TABLE II
FEATURE IMPORTANCE SCORE

Sl. No.	Important Features	Scores
1	Credit History	0.279579
2	Applicant Income	0.206035
3	Loan Amount	0.205250
4	Co-applicant Income	0.123400
5	Property Section	0.050221
6	Loan Amount Term	0.031859
7	Married Section	0.028636
8	Employed Section	0.026682
9	Education Section	Education Section
10	Gender	0.023776

from the Multinomial Naive Bayes, the rest of the algorithm achieved an accuracy range between 70% to 85% in which SVM linear performed the best compared to the other models. The different parameter namely precision, recall and F1-Score are compared for the different classification models. The important features are taken into consideration while evaluating the model. In the future, the software could be developed so that new research data could be used after a certain period. An application with proper UI can be developed, as a way to take numerous inputs from the consumer like cope with loan quantity, loan duration, and many others.

REFERENCES

[1] Hamid, A. J., & Ahmed, T. M., "Developing prediction model of loan risk in banks using data mining", Machine Learning and Applications: An International Journal (MLAIJ), vol. 3, no. 1, 2016.

[2] Vimala, S., & Sharmili, K. C., "Prediction of loan risk using naive bayes and support vector machine", International Conference on Advanced Computing Technologies (ICACT), vol. 4, no. 2, pp. 110-113, 2018.

[3] Arora, N., & Kaur, P. D., "A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment", Applied Soft Computing, vol. 86, 2020.

[4] Arun, K., Ishan, G., & Sanmeet, K., "Loan approval prediction based on machine learning approach", IOSR J. Comput. Eng, vol. 18, no. 3, pp. 18-21, 2016.

[5] Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K., "A study on predicting loan default based on the random forest algorithm", Procedia Computer Science, vol. 162, pp. 503-513, 2019.

[6] Alaraj, M., Abbod, M., & Hunaiti, Z., "Evaluating Consumer Loans Using Neural Networks Ensembles", International Conference on Machine Learning, Electrical and Mechanical Engineering, January 2014.

[7] Aditi Kacheria, Nidhi Shivakumar, Shreya Sawkar, Archana Gupta, "Loan Sanctioning Prediction System", International Journal of Soft Computing and Engineering (IJSCE), vol. 6, no. 4, pp. 50-53, 2016.

[8] Addo, P. M., Guegan, D., & Hassani, B., "Credit risk analysis using machine and deep learning models", Risks, vol. 6, no. 2, 2018.

[9] Goyal, A., & Kaur, R., "Loan prediction using ensemble technique", International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no. 3, pp. 523-526, 2016.

[10] Jency, X. F., Sumathi, V. P., & Sri, J. S., "An exploratory data analysis for loan prediction based on nature of the clients", International Journal of Recent Technology and Engineering (IJRTE), vol. 7, no. 4, 2018.

[11] Supriya, P., Pavani, M., Saisushma, N., Kumari, N. V., & Vikas, K., "Loan prediction by using machine learning models", International Journal of Engineering and Techniques, vol. 5, no. 22, 144-148, 2019.

[12] Nikhil Madane, Siddharth Nanda, "Loan Prediction using Decision tree", Journal of the Gujrat Research History, vol. 21, no. 14, 2019.

[13] Arutjothi, G., & Senthamarai, C., "Prediction of loan status in commercial bank using machine learning classifier", 2017 International Conference on Intelligent Sustainable Systems (ICISS) pp. 416-419, IEEE, December 2017.

[14] Rodrigues, A. P., Chiplunkar, N. N., & Fernandes, R., "Aspect-based classification of product reviews using Hadoop framework", Cogent Engineering, vol. 7, no. 1, 2020.

[15] Rodrigues, A. P., Chiplunkar, N. N., & Fernandes, R., "Social Big Data Mining: A Survey Focused on Sentiment Analysis", Handbook of Research on Emerging Trends and Applications of Machine Learning, pp. 528-549, IGI Global, 2020.

[16] Fernandes, R., & Rodrigues, A. P., "An Approach Toward Stateless Chatbots with the Benefit of Tensorflow Over Spacy Pipeline", Advances in Artificial Intelligence and Data Engineering, pp. 483-496, Springer, Singapore, 2021.

Machine Learning Algorithm to Predict Fraudulent Loan Requests

Nazmul Hasan
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
nazmul15-7914@diu.edu.bd

Tanvir Anzum
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
tanvir15-7890@diu.edu.bd

Tareq Hasan
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
tareq15-9213@diu.edu.bd

Nusrat Jahan
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
nusratjahan.cse@diu.edu.bd

2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 978-1-7281-8595-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCNT51525.2021.9579517

***Abstract*— Machine learning is a strategy that enable computers to automatize information-driven model building and programming through a scientific discovery of statistically important patterns within the obtainable data. The learning capability of a machine and the ability to do predictive analysis is very obligatory in this age of vast information. In this study, we focused on banking sector where too many individuals are applying for bank credits. Though, it is really troublesome to determine whom loan should be granted or whom should be rejected. For banking organizations acceptance of loan is a main task. The prediction model that we formed in this paper for predicting fraudulent loan requests. In this paper, we were working with six algorithms – Decision tree, Support vector machine, Random forest, K nearest neighbors, Ada-Boost, and Logistic regression to predict the fraudulent loan request from customers. We got 83.75% accuracy from K-Nearest Neighbors algorithm which was better than other five machine learning approaches.**

Keywords*— Machine learning, Fraud detection, Data mining, Classification model, KNN

I. INTRODUCTION

A loan may be a type of debt incurred by an individual or various entity. The lender that is normally a corporation, establishment, or government—advances an amount of cash to the recipient. In instead, the recipient had to concur to an individual set of rules as well as any finance charges, interest, reimbursement date, and different conditions. People borrowed loan that is a good amount of cash for a period of time and they have to be paid the loan back with given rate of interest. The purpose of the loan is often anything supported the customer requirements. A. J. Hamid and T. M. Ahmed [1] discussed different loan requests such as open-ended loan, close ended loan, secured loan, unsecured loan, and Mortgage loan. An open-ended loan is basically an addition of credit where anyone can borrow money whenever they needed and they have to pay back the money on an ongoing basis as example credit card. The features of the credit cards are popular sorts of open ended loan. There is a limitation of credit that is available for both of these sorts of loans. At the time of purchasing anything the remaining credit will decrease automatically. At the time of making too much expenditure, you can spend your credit until you run out of it. Close ended loans are such kind of loan that you must pay an amount at a certain date. At the time of expenditure, the balance is decreasing automatically in the closed ended loan and when do not have existing credit you are allowed to use it is close ended loan. You will have to make an application if you want to lend money. Closed ended loans have different types. Secured loans are unit loans that believe associate quality. In secured loans if anyone fails to repay the loan, then the lender has the rights to possess the asset and use of it. Before taking a secured loan you must calculate all your assets. There might be a complicated situation at the time of unsecured loans lend and at the same time it has a higher concern rate. To fulfil the requirements of the unsecured loans the credit history and revenue must be checked. The lender has to choose an alternative way to retrieve the loan if anyone failed to retard back the unsecured loan. Conventional Loans are a kind of loans that are not assured by the government organization or Veterans management.

According to Fannie Mae and Freddie Mac Conventional loans might be conforming. Nonconforming loans do not match the qualifications of Fannie and Freddie. The main motive of the banks to earn interest income by lending money to the companies. Commercial banks offer's maximum amount of cash because they are given a short time to repay. They also charge different interest rates for various customers to equal several risk factors. Personal loans are not larger than the business loans and in business loan lenders remarkably expand the loan profit and lower the interest rate. In addition, with interest charges through fees and late penalties banks earn their profit. The income and the stability of the banks strongly affects the economic strength of an area. The more loan facilities the entrepreneurs and corporations will get, the newer business will develop and grow. Commercial banks think lending money to the companies can improve the local economic condition and companies also can grow and generate their profit in business. A recent study found that 50 percent of people spend less on the basis of their income. But most of them are trying to figure out how to get rid of this uncertain condition. They can get rid of this uncertainty by taking a loan. The loans can be classified into two parts that is business and private loans. The reason for people taking loans are given below.

- Establish their business: For starting a business capital is a must. But most of the people have a problem with cash, which they can't afford. This is the main reason that's why people take loans for starting their business. Both individuals and corporate needs money to expand their business. Taking a loan, they can either establish

or expand their business and that's the first reason on why people are taking loans.

- Debt consolidation: When someone has more than one debt, then he/she can consolidate the debt and pay them together. It might be a good idea for everyone if the bank considers a lower interest rate. By consolidating various debt one can make his payment easier. If anyone wants to consolidate all his loans, then he must have enough cash to pay the debt, and that's why people are planning to consolidate their existing loan.
- Paying for school fees: Another popular reason for taking loans is to assist children by paying their tuition fees. The value of higher studies in many countries is high and many parents can't afford the tuition fees for their children. As education is a primary right for everyone that's why parents are ready to take loan for their children so that they can study well and could help in the progress of the country.
- Build a house: If anyone wants to build a house, then he/she must have enough money for this. Only a few can make their dream true by using their monthly salary. Otherwise, who wants to build a house have to apply for a loan because building a house costs a huge amount of money. Therefore, if anyone wants to build a house and if has not enough money, then a loan can be applied to build a house.

We organized our whole paper as follows – In section II, literature review is listed to know about previous work, section III for describing proposed method, section IV to shoe experiment results and finally, for conclusion and future work we introduce section V.

II. Related Work

In 2016 A. J. Hamid and T. M. Ahmed proposed a method that can easily classify loan risk using data mining techniques which can help the banking sector. Using the data from banking sector a model has been built which can predict the state of the loan. J48, bayesNet and naive bayes model was used to build the proposed model and accuracy for j48 was 78.3784 %, bayesNet was 77.4775 attempt to naive bayes was 73.8739 %. [1]. In 2017 [2], Girija V. Attigeri and et al., used an empirical approach for building a model which can easily identify credit risk assessment by using supervised learning algorithms. Logistic regression (LR) and Neural network (NN) algorithm were used to evaluate the test. Training accuracy of LR was 78.60%. The dataset has been collected from UCI Repository. Using this machine learning model bad customer could be easily identified. Mohammad Ahmad Sheikh and et al., [3] also developed a model using Logistic Regression (LR) with a sigmoid function for predicting loan defaulters. They proposed model deal with whether the loan will be accepted or not. The dataset was collected from Kaggle. Based on the data set they got 81% accuracy from the model.

In 2020 J. Tejaswini and et al., proposed a model that can predict if a customer will pay his loan or not. Decision Tree machine learning algorithm showed the best result as compared to Logistic Regression and Random Forest machine learning approaches [4]. Lin Zhu and et al., [5] used Random Forest (RF) algorithm for building a model that can identify the loan defaulters and helps to lessen the risk of user loan default. Because of remaining imbalance class in the dataset author had to use SMOTE method to get rid of the problem. Author collected the data set from Lending club for the primary quarter in 2019 which had total 115,000 real loan request from users with 102 features and experimental results show that, RF algorithm achieved 98%, which is the best accuracy as compared to logistic regression (LR), decision tree (DT) and SVM algorithms. In 2016 at paper [6], Kumar Arun and et al., proposed a model that will define which person is safe or not for assigning loans. This model is done by using the previous records of the people who have already gotten the loan. Taking these records, the machine will be trained using a machine learning algorithm that will help to fetch the most accurate result. So in this paper author tried to find out the risk factor which will help to find out the safe person and save many bank assets. S. F. Eletter and et al., at paper [7] formed a model using Artificial Neural Network (ANN) which helps to make a decision about a loan evaluating the credit applications. The dataset was collected from the Jordanian Commercial bank, which contains loan applications of 140 people where 94 cases used as training data and 46 used as testing data. From the testing data, the model classifies 95% accuracy. In 2013 M. Nazari and et al., used Artificial Neural Network (ANN) algorithm to check whether a customer is suitable for assigning loans or not for reducing the risk of loan defaulters. SPSS and MATLAB software was used for doing the mathematical calculations. The dataset was collected from Iranian commercial banks. The dataset contains 497 samples which have 18 variables that help to find out the suitable customers. For training purposes, 70 percent of data was taken and for the testing purpose, 20 percent of data was taken to train the model [8].

S. T. Li and et al., [9] proposed a model using Support Vector Machine (SVM) algorithm to check whether the applicants are strong enough for assigning loans. The dataset was collected from a neighborhood bank in Taiwan. The dataset contains 600 applications and 17 variables. The accuracy achieved by using the SVM algorithm was 84.83%. In 2019 I. O. Eweoya and et al., proposed a model using Support Vector Machine (SVM) algorithm to check the fraud in the loan application using a variety of data that hold the credit history and past occurrence of the consumer. The dataset has 5000 instances and 9 attributes. To reduce the risk of financial fraud machine learning algorithm plays a vital role here. The model achieved 81.3% accuracy, using the SVM kernel. They also proposed a model using Decision Tree Algorithm for identifying fraud in bank loans using previous records of the client. The proposed model also helps to reduce the risk of bank loans. The model achieved 75.9% accuracy using testing and training datasets in MATLAB [10, 11]. In this paper [12], the author used a supervised machine learning approach to find out the fraud in the loan administration. The Naive Bayes algorithm was used in this model to find out the loan defaulters. The dataset contained 5000 samples data and the algorithm achieved 78 % accuracy.

In 2017 A. Gahlaut and et al., proposed a model that will help the customer in identifying their credit score, which is good or bad by evaluating their past records to reduce the risk. The dataset was collected from UCI data repository. The author used six supervised algorithms where Random Forest algorithm provided the best accuracy [13]. Rashmi Malhotra and et al., at paper [14] used the comparison of multiple discriminant analysis (MDA) along with neural networks for predicting the problems in loan defaulters. Total 1078 samples with 12 different credit unions. SPSS program used for

analyzing the discriminant analysis and MATLAB used for performing the Neural network. The Neural Network model achieves the best accuracy as compared to the Discriminant model. In 2013 A. K. I. Hassan and et al., [15] proposed a model comparing three supervised neural network algorithm to find out the loan defaulters. The dataset was collected from a German bank which has 1000 samples for training and testing purpose. Levenberg-Marquardt achieves the best accuracy as compared to others.

In paper [16] E. Turkson and et al., proposed a model to check the loan defaulters by using real bank data. Several machine learning algorithms were used here to find out the best accuracy and reduce the risk of the loan default. The most important features of this paper are to find out whether a customer will default or not. The dataset was collected from UCI machine learning data repository which has 23 features. The data set was submitted by I-Cheng Yeh. The accuracy rate achieved using several algorithms was between 76% to over 98%. Y. B. Wah and et al., at paper [17] proposed a model that used data containing demographic characteristics, payments to find out the credit risk in the banking sector. The author used this data for reducing the risk of the banking sector and also helps to find out the proper customer. The dataset contains 4305 credit card applicants where 31% were rejected and 69% were accepted applicants. Three machine learning algorithms was used here in this model that was Logistic regression, Classification and Regression Tree and Neural Network and Neural Network achieved a slightly higher accuracy as compared to other algorithms (NN = 76.46%, LR = 74.56%, CART = 73.66%) [18].

From above statistics, we can say that in recent years bank scams are increasing alarmingly. Now it's important to predict that kind of incident to avoid loan fraud occurrence. That's why in this paper, we attempted to build a predicting model by using machine learning algorithm that can be able to predict fraud in loan default.

III. Proposed Methodology

We developed a model to predict loan request from customers. In this recent years, bank organizations worried about the approval of a loan request from customers. In this section, we discussed about data collection and data visualization. In Fig. 1 we presented our whole work through a flow chart.

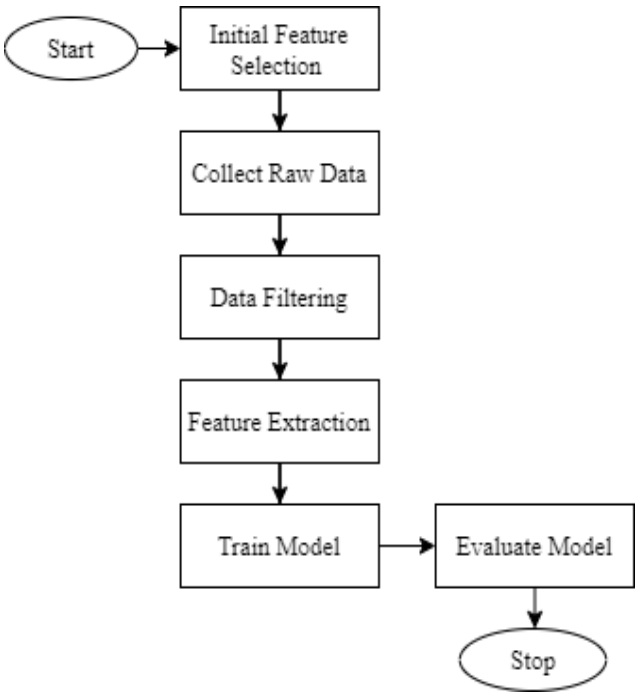


Fig. 1. Flow chart diagram to predict fraud in loan default.

A. Data Collection

Data is most crucial part of data science. To get better prediction we have to collect data from real source. Data should be error free and accurate. In this study, we collected data from an online platform “Analytics Vidhya”, which also provide different problem to get better solution [19]. Here, Table I to present the sample dataset.

TABLE I. Sample dataset for prediction Loan.

	Loan ID	Gender	Married	Dependents	Education	Self-employed	Applicant Income	Co-Applicant Income	Loan Amount	Loan Amount Term	Credit History	Property area	Loan Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	N/A	360.0	1.0	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y

B. *Dataset Visualation*

Visual presentation is best to understand any information. Now, we analysis our dataset in different angles. Plot pairwise relationships in a dataset showed in Fig. 3.

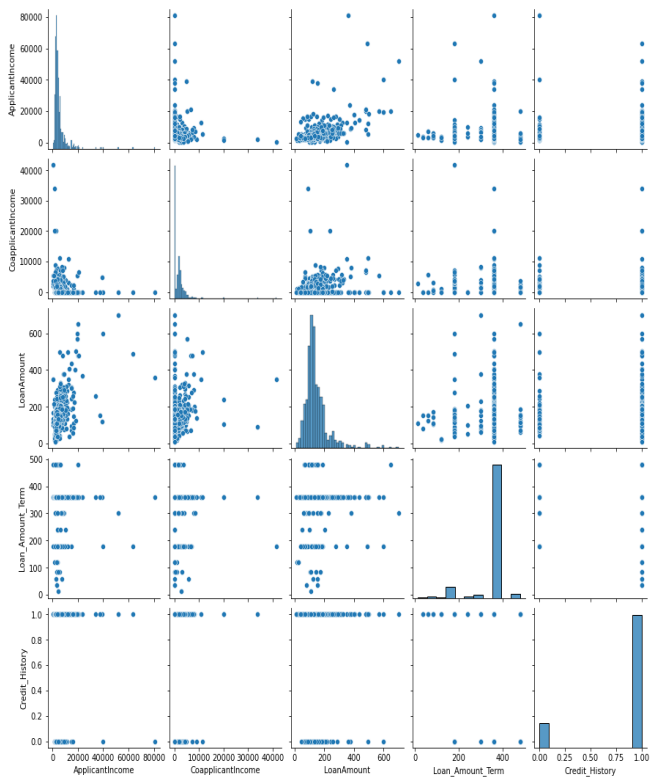


Fig. 3. Plotted pair wise dataset.

We are working with total 614 sample data with 13 attributes. Here, some attribute are continuous variable and some attributes are categorical variable. We also checked missing value in dataset to pre-process the whole data. In Fig. 4, we counted missing value.

```
Loan_ID              0
Gender              13
Married              3
Dependents          15
Education            0
Self_Employed       32
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount          22
Loan_Amount_Term    14
Credit_History      50
Property_Area        0
Loan_Status          0
dtype: int64
```

Fig. 4. Missing value count.

From Fig. 4, following we can see the list of missing value. Six attributes have no missing data. Dataset should be well organized with real data. However, finding missing value and filling missing value with their reprehensive mode value can increase the quality of a dataset. After that, filling missing value counting in Fig. 5.

```
Loan_ID              0
Gender               0
Married              0
Dependents           0
Education            0
Self_Employed        0
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount           0
Loan_Amount_Term     0
Credit_History       0
Property_Area        0
Loan_Status          0
dtype: int64
```

Fig. 5. Fix all missing value issues.

Data should be relative to get better prediction results. Here, heat map in Fig. 6 to find out relative importance of the dataset.

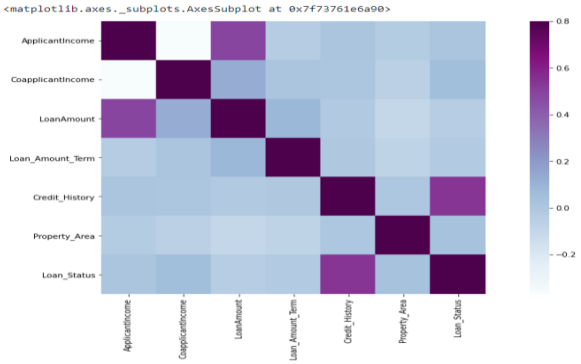


Fig. 6. Heat map of relative importance of dataset.

It is obvious that, Application Income and Loan Amount is correlated, Co-applicant Income correlated with Loan Amount and Credit history is correlated with Loan Status.

We can identify the distribution of entire data with help of Mean and Standard Deviation. When the data is normally distributing maximum data is centralized near the mean value of the data. To understanding of distribution, we can simply plot distribution plot. Fig. 7 and 8 are an example data distribution.

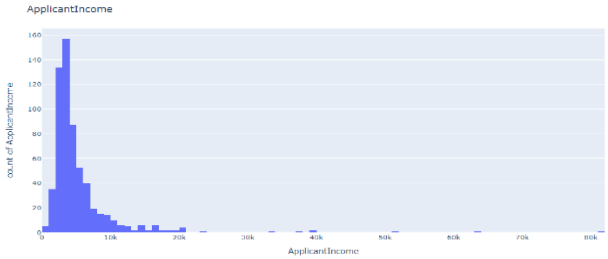


Fig. 7. Before Data Distribution of Application Income.

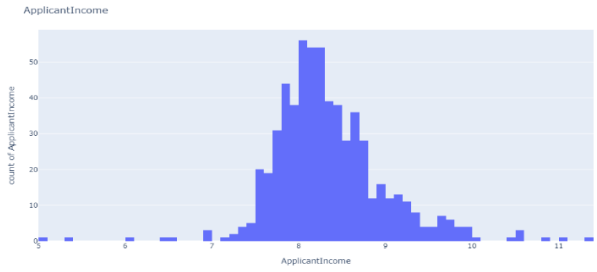


Fig. 8. After converting Normal Distribution of Application Income.

IV. Result Analysis

A. SVM

A supervised machine learning formula which might be used for classification and regression challenges. Basically, SVM vastly using for classification problem. A linear kernel is often used as traditional multiplication of any two given observations. The product between two vectors is that the summation of the multiplication of every try of input value. Mathematically it express with equation (1)

$$k(x, x_i) = SUM(x * x_i) \quad (1)$$

B. Random Forest

Random forest applied general approach of bootstrap that aggregating with tree learners. Here, given a training set $P = p_1, \dots, p_n$ with responses $Q = q_1, \dots, q_n$ bagging approach repeatedly select a sample randomly with replacement of the training set as well as make tree with these samples. After training, prediction results for test samples p' can be made by averaging the prediction results from all the separate classification tree on p':

$$\hat{f} = \frac{1}{A}\sum_{a-1}^{A} f_a(p') \quad (2)$$

C. Decision tree

Decision tree is a common supervised algorithm. This algorithm creates tree structure. Whole tree present different layer based on training data. Decision is presented in branch and final result is presented through leaf node. In decision tree algorithm it spilt data by helping Gini index.

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2 \quad (3)$$

D. Adaboost classification

Ada-boost classifier forms a strong classifier by combining multiple weak classifier. A single algorithmic rule might classify the object unwell. However, if we have a tendency to mix more than one classifiers with choice of training set at each iteration and distribution correct quantity of weight in final vote, we are able to have sensible accuracy for general classifier. Mathematically it will express with the following equations 4,

$$H(y) = sign\left(\sum_{t-1}^{T} \alpha_t h_t(y)\right) \quad (4)$$

E. K-nearest neighbors

KNN method is another well-known machine learning method, that could be used for classifying and regression problem. There is no supposition for underlie data distribution, that's why it's called non-parametric. In K-NN algorithm, K is representing the number of nearest neighbors. All nearest neighbors are the core composing factor. Normally K is an odd number if the number of class is two. When k = 1, then the method is aware of the nearest neighbor algorithm. KNN working process is: Calculate distance, Find the closest neighbors, Vote for labels.

For distance metrics, we will use the Euclidean metric equation 5.

$$d(a, a') = \sqrt{(a_1 - a'_1)^2 + \cdots + (a_n - a'_n)^2} \quad (5)$$

Finally, the input x gets imposed to the class with the largest probability.

$$Q(B = j|A = a) = \frac{1}{K}\sum_{i \in C} I(b^{(i)} = j) \quad (6)$$

F. Logistic Regression:

Generally logistic regression used for predicting binary classes. The target variable or result is dichotomous(binary) in nature. That means there are accepts only two possible classes. Either TRUE=1 or FALSE=0. It is a bit similar to the linear regression method. We can say it as a generalized of linear model. Linear regression equation is:

$$y = \beta 0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + +\beta_n X_n \quad (7)$$

Where, y is representing dependent attribute and X_1, X_2… and X_n are representing explanatory attributes.
The required sigmoid function is in equation 8:

$$Q = \frac{1}{1 + e^{-y}} \quad (8)$$

We are working with six different classification method in our paper. We are trying to find their accuracy level difference and find out which one is best. This will be shown in Table II.

TABLE II. Presenting Six Classification Techniques

Classifier	Accuracy (%)	Precision		F1-Score		Recall		Support	
		0	1	0	1	0	1	0	1
KNN	83.7%	0.44	0.98	0.56	0.89	0.90	0.82	21	133
SVM	83.1%	0.44	0.98	0.59	0.89	0.90	0.82	21	133
Decision Tree	83.1%	0.44	0.94	0.55	0.87	0.73	0.81	26	128
Logistic Regression	83.1%	0.44	0.98	0.59	0.89	0.90	0.82	21	133
Ada-boost	82.4%	0.47	0.98	0.59	0.89	0.90	0.82	21	133
Radom Forest	79.8%	0.49	0.95	0.51	0.89	0.81	0.83	26	128

Here, F1-score is for controlling the balance between recall and precision. Where precision is mention the ration of true positive prediction value and the total true positive class values predicted. Another name is F-score. Recall is also important to define a performance of an algorithm. It is presented the number of Positives value predicted divided by the number of True Positives and the number of False Negatives.

From Table II, we can see that KNN gives us better accuracy (83.7%) than all listed classification methods. If we able to build a system to predict fraud in loan default using KNN classifier, it will be decrease fraud in loan default.

V. Conclusion and Future Work

In this paper study, our aim was to develop a loan credibility predication system. We select this field to help the bank organization from a fraud. Recent years, the crime rate is increasing day by day. Loan prediction is not an easy task. Only real data can provide more accurate result. For this paper work, we considered data from an online source. In this paper, KNN classifier with 83.7% accuracy is used for predicting fraud in loan default and the outcome were compared with five different classifier methods – SVM, Decision tree, LR, Ada-boost and Random Forest. The experiment shows that KNN classifier performs outstanding than the all other classifiers.

Creation of dataset and pre-processing were one of our main challenges and limitations. Some user creates common data and in this case need to apply proper data cleaning approach. In further study, our target is to gather more data and collect local data for getting more accurate results. Data volume and local data will be helpful to perform a crucial experiment.

References

[1] Hamid, A. J., Ahmed, T. M., "Developing Prediction Model of Loan Risk in Banks Using Data Mining". Machine Learning and Applications: An International Journal, 3(1), pp. 1-9, 2016.

[2] Attigeri, G. V., Pai, M. M., Pai, R. M., "Credit Risk Assessment Using Machine Learning Algorithms". Advanced Science Letters, 23(4), pp. 3649-3653, 2017.

[3] M. A. Sheikh, A. K. Goel, and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, pp. 490-494, 2020.

[4] J. Tejaswini, T. Mohana Kavya, R. Devi Naga Ramya, P. Sai Triveni, Venkata Rao Maddumala, "Accurate Loan Approval Prediction Based On Machine Learning Approach", Journal of Engineering Science, Vol 11, Issue 4, pp. 523-532, 2020.

[5] Zhu, L., Qiu, D., Ergu, D., Ying, C., Liu, K., "A study on predicting loan default based on the random forest algorithm". Procedia Computer Science, 162, pp. 503-513, 2019.

[6] Arun, K., Ishan, G., and Sanmeet, K., "Loan approval prediction based on machine learning approach". IOSR Journal of Computer Engineering (IOSR-JCE), 5, pp.18–21, 2016.

[7] Eletter, Shorouq Fathi, Saad Ghaleb Yaseen, and Ghaleb Awad Elrefae, "Neuro-based artificial intelligence model for loan decisions." American Journal of Economics and Business Administration, vol. 2, no. 1, 2010.

[8] Nazari, M., and Alidadi, M., "Measuring Credit Risk of Bank Customers Using Artificial Neural Network". Journal of Management Research, vol. 5(2), 2013.

[9] Li, S., Shiue, W., Huang, M., "The evaluation of consumer loans using support vector machines". Expert Systems with Applications, Vol. 30(4), pp. 772-782, 2006.

[10] Eweoya, I. O., Adebiyi, A. A., Azeta, A. A., Amosu, O., "Fraud prediction in loan default using support vector machine". Journal of Physics: Conference Series, 1299, 012039, 2019.

[11] Eweoya, I. O., Adebiyi, A. A., Azeta, A. A., Azeta, A. E., "Fraud prediction in bank loan administration using decision tree". Journal of Physics: Conference Series, 1299, 012037, 2019.

[12] Eweoya, I. O., Adebiyi, A. A., Azeta, A. A., Chidozie, F., Agono, F. O., Guembe, B., "A Naive Bayes approach to fraud prediction in loan default," Journal of Physics: Conference Series, 1299, 012038, 2019.

[13] Gahlaut, A., Tushar, Singh, P. K., "Prediction analysis of risky credit using Data mining classification models". 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017.

[14] Malhotra, R., Malhotra, D., "Evaluating consumer loans using neural networks". Omega, 31(2), pp. 83-96, 2003.

[15] Hassan, A. K., & Abraham, A., "Modeling consumer loan default prediction using ensemble neural networks". International Conference On Computing, Electrical and Electronic Engineering (Icceee), 2013.

[16] R. E. Turkson, E. Y. Baagyere and G. E. Wenya, "A machine learning approach for predicting bank credit worthiness," 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR), Lodz, pp. 1-7, 2016.

[17] Yap Bee Wah and Irma Rohaiza Ibrahim, "Using data mining predictive models to classify credit card applicants," 2010 6th International Conference on Advanced Information Management and Service (IMS), pp. 394-398, 2010.

[18] Krishnan, Varun B., "Bank Frauds up by 45% in 10 Years, Show Data," The Hindu. The Hindu, October 9, 2019.

[19] "Loan Prediction" Accessed on: January 25, 2021. [Online]. Available: https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 04 | Apr 2022 www.irjet.net p-ISSN: 2395-0072

Loan Analysis Predicting Defaulters

Mudit Manish Agarwal [1], Harshal Mahendra Shirke [2], Vivek Prafullbhai Vadhiya [3], Manya Gidwani [4]

[1,2,3]*Student, Department of Information Technology, Shah & Anchor Kutchhi Engineering College, Mumbai, India*
[4] *Professor, Department of Information Technology, Shah & Anchor Kutchhi Engineering College, Mumbai, India*

Abstract - ***Due to the advancements in the domain of Artificial Intelligence and Data Science, its utilization is becoming more common in every possible domain. Nowadays, the majority of the industries make use of AI and its applications in some or the other way. Taking the advantage of the field of Data Science results in creating effective and modern applications, products irrespective of the domain. One of the industries where the application of AI and Data Science is proving to be effective is the Finance Industry commonly known as the Banking Sector. Banks face severe losses due to the loan defaults made by the client and hence to overcome this problem, there lies a need to create a credit risk scoring model which can analyze and predict the loan defaults. Hence, with the help of Machine Learning, we aim to create a Loan Default Analysis model which could predict the loan defaults and integrate the model into a web application for the user for easy usability.***

Key Words*: *Loan Default, Machine Learning, data mining, prediction, web application.

1. INTRODUCTION

Due to the Covid-19 Pandemic, there is a huge loss of capital caused to the banking sector, financial institutions, small scale finance companies, etc. Nearly a year and a half, everything has been shut down thereby leaving people with no source of income. Due to this reason, there has been a significant increase in the loan defaults made by the client. Now, in this new normal, there is a need especially for the banks to strengthen their loan sanctioning system. Since banks may face huge losses due to the defaults made by their clients which increases the rejection rate of the loan applicants. This affects the bank's overall reputation and also at the same time, due to the rejection of new loan applicants it causes huge financial loss to the bank since, the most common type of unsecured loans are debt consolidation, credit-card loans, student loans, and personal loans [1].

The Loan Analysis Predicting Defaulters (LAPD) is an attempt at creating a better credit risk scoring model which can correctly identify which applicant will be a defaulter in the future. This is done by analyzing the historic data and identifying the patterns. Such a model would minimize credit risk and prevent the clients who are capable of repayment from getting rejected. Various classification algorithms such as Logistic Regression, Decision Tree, Random Forest, etc. have been applied to build various models and compare them to find out the best accuracy [4]. The source of the dataset is Kaggle which is a highly imbalanced dataset. Data Balancing techniques such as SMOTE, SMOTE ENN have been implemented to balance the dataset. Important features such as loan_amount, term, home_ownership, issue_date_year, etc have been extracted on which the target attribute depends for prediction. Once the model creation and comparison are done, the model which gives better accuracy is integrated into the website for the end-user

2. BUSINESS PROBLEM

Banks may suffer significant losses as a result of customer defaults, which raises the rejection rate of loan applications. To address this issue, we require a more accurate credit risk assessment algorithm that can predict which applicants will default in the future. This will be accomplished by analyzing previous data and discovering trends. This methodology would reduce credit risk and prevent clients who are capable of repayment from being turned down [3].

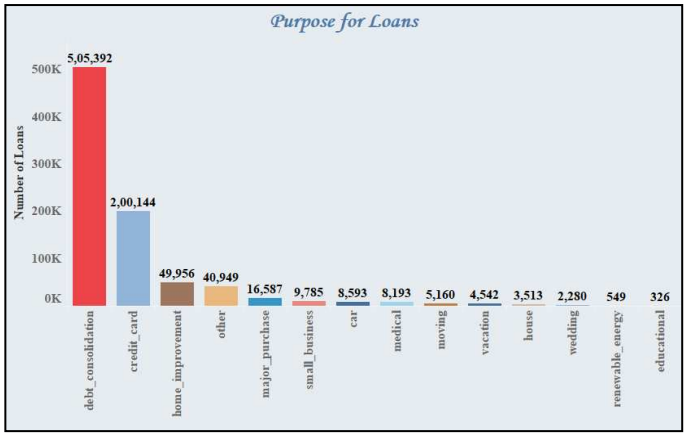


Fig. 2.1 Purpose of loan

3. DATA DESCRIPTION

The Data Set was taken from Kaggle The dataset includes entire loan data for all loans granted between 2007 and 2015, including current loan status (Current, Late, Fully Paid,

etc.) and most recent payment information. There are 8,55,969 items in our dataset, with 73 attributes including the target variable. Furthermore, the dataset is extremely imbalanced, with 46467 entries of failed loans. This dataset contains a variety of attributes, including category, numeric, and date data. The number of defaulters increased significantly between 2012 and 2014, with the LIBOR Scandal, Hurricane Sandy, and Hostess Files for Bankruptcy being some of the primary factors. The major reason for a borrower's request for a loan is the loan purpose. Debt consolidation is the most common reason for taking out a loan, followed by credit card debt.

Some important features from the dataset

- loan_amnt - Amount of money requested by the borrower.
- int_rate - The interest rate on the loan.
- grade - Loan grade with categories A, B, C, D, E, F, G.
- annual_inc - Borrowers annual income.
- purpose - The primary purpose of borrowing.
- installments – Monthly amount payments for the opted loan.
- term – duration of the loan until it's paid off
- to – the ratio of your gross monthly income that goes to paying your monthly debt payments [6].

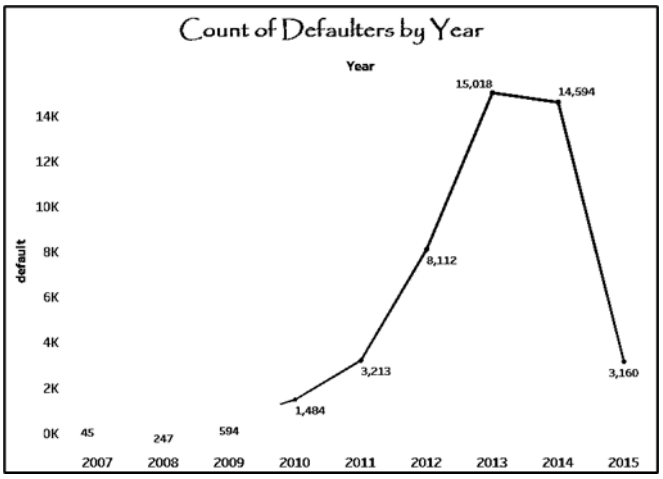


Fig. 3.1 Understanding the dataset

4. DATA PRE-PROCESSING

The data were inspected, cleaned, and prepared as follows before data mining model analysis: Changing the names of the variables — We changed the names of the variables to suit our needs and domain expertise. Retrieved the values of variables such as emp length and term, which had values in a string from which we extracted the numeric component.

E.g.: -'10+ years' -> 10;

'2 years' -> 2;

'1 year' -> 1;

'<1 year' -> 0;

We did the same thing for the term with string values.

For example: - 36 months -> 36

60 months -> 60

Values from one variable are replaced with values from another variable. Based on the application type, the income value of the borrower is replaced with the income value of the co-borrower. If the application type is INDIVIDUAL, the borrower's income will be preserved, but if the application type is JOINT, the current value of the borrower's income will be substituted with joint income. while joint income is the sum of the borrower's and co-income. borrower's Dti/ratio inc exp borrower and dti joint/ratio inc exp joint follow the same technique.

4.1 Label Encoders

The Label Encoder is a method for converting category data to numeric variables. We used the Label encoder approach to categorize the aforementioned variables, which are categorical. We have two functions in this technique: fit and transform.

Label Encoder's Steps:

1. Obtaining the one-of-a-kind values
2. Putting them in ascending order is a good idea.
3. Starting with 0,1,2, and so on, mapping the values.
4. All three processes will be completed by the Fit function.
5. The data values in the data frame will be replaced by the transform function.

4.2 Train Test Split

Splitting the data into train and test in the ratio 7:3 where the training part contains 70 % of the data and Testing part contains 30% of the data Train Size – 598978 records with 8 features Test Size – 256991 records with 8 features. The Dataset was highly Imbalanced with variable 0 (Non – Defaulters) = 809502 and 1 (Defaulters) = 46467. In order, the balance this day out we have used two techniques which are Smote and Smote ENN [3].

4.3 Balancing the dataset

4.3.1 SMOTE

Imbalanced classification entails creating prediction models for datasets with a significant class imbalance. When working with unbalanced datasets, the difficulty is that most machine learning approaches will overlook the minority class, resulting in poor performance, even though performance on the minority class is often the most

significant. Oversampling the minority class is one way to deal with unbalanced datasets. Duplicating instances in the minority class is the easiest way, but these examples don't provide any new information to the model [7]. Instead, new instances may be created by synthesizing old ones. The Synthetic Minority Oversampling Technique, or SMOTE for short, is a kind of data augmentation for the minority population. Imbalanced classification entails creating prediction models for datasets with a significant class imbalance. When working with unbalanced datasets, the difficulty is that most machine learning approaches will overlook the minority class, resulting in poor performance, even though performance on the minority class is often the most significant. Oversampling the minority class is one way to deal with unbalanced datasets. Duplicating instances in the minority class is the easiest way, but these examples don't provide any new information to the model. Instead, new instances may be created by synthesizing old ones. The Synthetic Minority Oversampling Technique, or SMOTE for short, is a kind of data augmentation for the minority population.

4.3.2 SMOTE ENN

This method, developed by Batista et al. (2004), combines the ability of SMOTE to generate synthetic examples for minority classes with the ability of ENN to delete some observations from both classes that are identified as having different classes between the observation's class and its K-nearest neighbor majority class. The following is a description of the SMOTE-ENN procedure.

(SMOTE begins) Select data at random from the minority class. Calculate the distance between the randomly generated data and the k closest neighbors. Add the result to the minority class as a synthetic sample by multiplying the difference by a random value between 0 and 1. Repeat steps 2–3 until the required minority class proportion is achieved. (SMOTE comes to an end) (ENN begins) Calculate K to be the number of closest neighbors. If K can't be determined, it'll be 3. Find the observation's K-nearest neighbor from the dataset's other observations, then return the majority class from the K-nearest neighbor. If the observation's class and its K-nearest neighbor's majority class are not the same, the observation and its K-nearest neighbor are removed from the dataset. Repeat steps 2 and 3 until each class has the required proportion of students. (The ENN comes to a close.) [3].

After using the techniques the data was balanced with equal parts using smote and in the ratio 6:4 by using smote ENN.

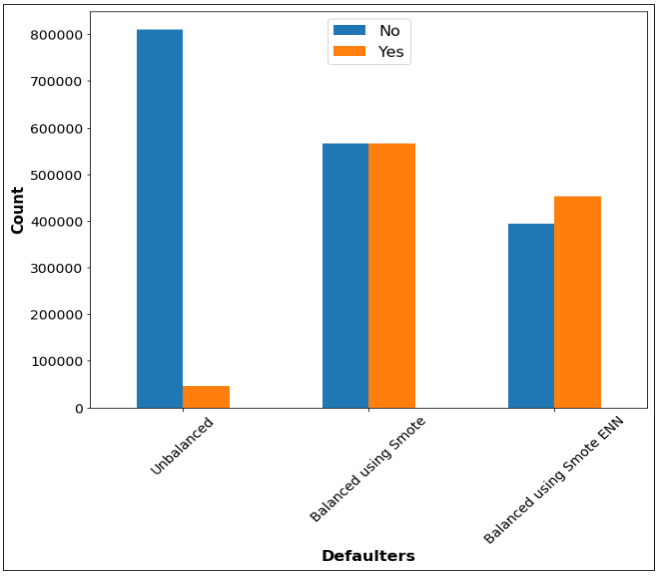


Fig. 4.1 Dataset after balancing

4.4. Algorithms

4.4.1 Logistic Regression

The method of modeling the likelihood of a discrete result given an input variable is known as logistic regression. The most frequent logistic regression models have a binary result, which might be true or false, yes or no, and so forth. Multinomial logistic regression can be used to model situations with more than two discrete outcomes. Logistic regression is a handy analytical tool for determining if a fresh sample fits best into a category in classification tasks. Because components of cyber security, such as threat detection, are classification issues, logistic regression is a valuable analytic tool [2].

4.4.2 Decision Tree

For classification and regression, Decision Trees (DTs) are a non-parametric supervised learning approach. The objective is to learn basic decision rules from data attributes to develop a model that predicts the value of a target variable. A tree is an approximation of a piecewise constant. Decision Trees are a form of supervised machine learning in which the data is continually split according to a parameter (you describe what the input is and what the related output is in the training data). Two entities, decision nodes, and leaves can be used to explain the tree. The decisions or ultimate outcomes are represented by the leaves. And the data is separated at the decision nodes. At first, we consider the entire training set to be the root. Categorical feature values are desired. If the values are continuous, they must be discretized before the model can be built. Records are distributed recursively based on attribute values. As the root of the internal node, we apply statistical approaches to rank characteristics.

4.4.3 Ada Boost

The AdaBoost algorithm, short for Adaptive Boosting, is a Boosting approach used in Machine Learning as an Ensemble Method. The weights are re-allocated to each instance, with larger weights applied to improperly identified instances. This is termed Adaptive Boosting. In supervised learning, boost is used to decrease bias and variation. It is based on the notion of successive learning. Each succeeding student, except the first, is produced from previously grown learners. In other words, weak students are transformed into strong students [6]. With a little modification, the AdaBoost method operates on the same idea as boosting. Adaptive Boosting is an effective ensemble approach for both classification and regression issues. It is most commonly used to solve categorization difficulties. It outperforms all other models in terms of model correctness, which can be verified by following the steps in order. To apply the boost and implement AdaBoost, one can start with decision trees and then move on to random forests. As we progress through the steps outlined above, accuracy improves. The weight-assigning approach used after each iteration distinguishes the AdaBoost algorithm from all other boosting algorithms, which is its strongest feature.

4.4.4 Random Forest

Random forest is a supervised machine learning algorithm that is commonly used to solve classification and regression issues. It creates decision trees from several samples, using the majority vote for classification and the average for regression. One of the most essential characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous and categorical variables, as in regression and classification. For classification difficulties, it produces superior results. Random Forest Algorithm Characteristics-

1) It outperforms the decision tree algorithm in terms of accuracy.

2) It is a useful tool for dealing with missing data.

3) Without hyper-parameter adjustment, it can provide a fair forecast.

4) It overcomes the problem of decision tree overfitting.

5) At the node's splitting point in every random forest tree, a subset of characteristics is chosen at random.

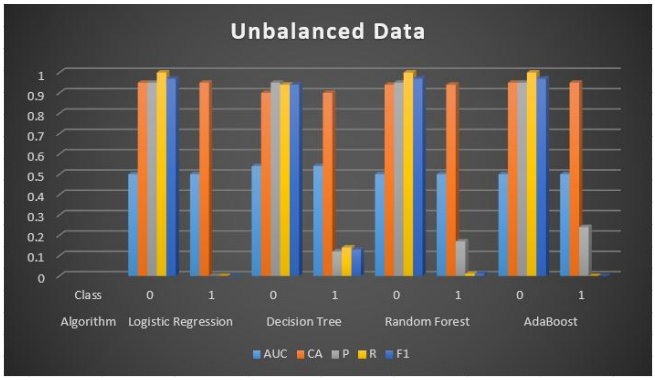


Fig. 4.2 Comparison of Accuracy of Unbalanced Data

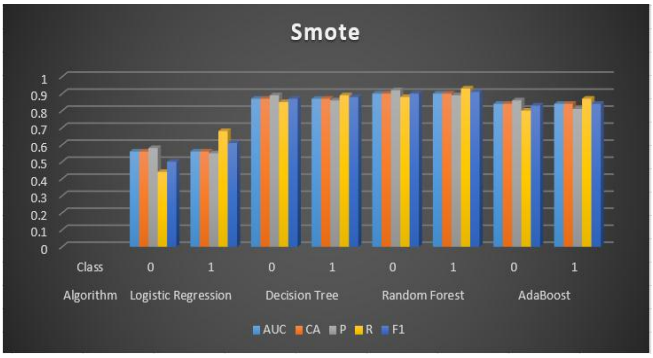


Fig. 4.3 Comparison of Accuracy of Balanced Data using Smote

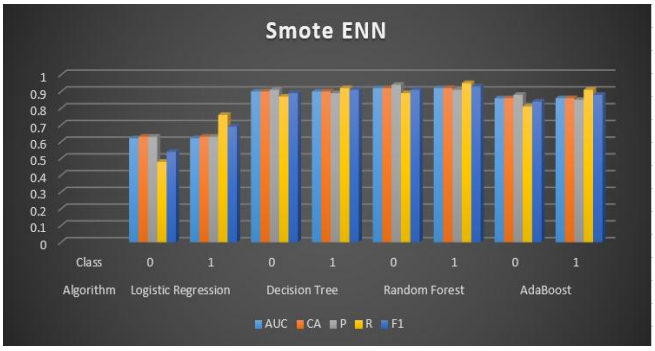


Fig. 4.4 Comparison of Accuracy of Balanced Data using Smote ENN

5.5 Pickle

For serializing and de-serializing a Python object structure, the Python pickle package is utilized. Pickling an object in Python allows it to be stored on a disc [3]. Pickle works by first "serializing" the item before writing it to file. Pickling is a Python function that converts a list, dict, or other Python object into a character stream. The assumption is that this character stream provides all of the data required to recreate the object in another Python function.

Parameters		AUC	CA	P	R	F1
Algorithm	Class					
Logistic Regression	0	0.62	0.63	0.63	0.48	0.54
	1	0.62	0.63	0.63	0.76	0.69
Decision Tree	0	0.9	0.9	0.91	0.87	0.89
	1	0.9	0.9	0.89	0.92	0.91
Random Forest	0	0.92	0.92	0.94	0.89	0.91
	1	0.92	0.92	0.91	0.95	0.93
AdaBoost	0	0.86	0.86	0.88	0.81	0.84
	1	0.86	0.86	0.85	0.91	0.88

Fig. 4.5 Parameters for Evaluation

After applying these algorithms, we were able an achieve an accuracy of 92 % which was given by Random Forest using Smote ENN method after which we used the pickle library to store the model and flask to integrate it with our webpage.

5. WEB TECHNOLOGIES

For our front-end part, we have used a Python framework called Flask which is used to make a webpage. For our project, we integrated our ML model into a website which is used to get rich output and visualization.

5.1 Flask:

Flask is a Python-based web application framework. The Werkzeug WSGI toolkit and the Jinja2 template engine are the foundations of Flask. Both are Pocco initiatives.

And for the database part, we have used MySQL and also we used the session to authenticate our login system with bank users.

5.2 MYSQL:

MySQL is a quick, easy-to-use relational database management system (RDBMS) that is utilized by many small and large enterprises. MySQL AB, a Swedish business, is responsible for its development, marketing, and support.

For our ML learning model, we used the Pickle python library for integration with the flask webpage also we show the graphs generated from the ML model into our webpage for Easy to understand our users.

Normally Flask is used HTML templates to render the webpage and for CSS we have used bootstrap and tailwind CSS for validation JS function. Also, in the footer, we mention our contributors' details. Total 4 webpages are there on our website which is Home, Login, Register, and Predict.

Screens on our website:

Home Page:

On this page, we make a basic webpage which is consist of Navbar, Body, and Footer.

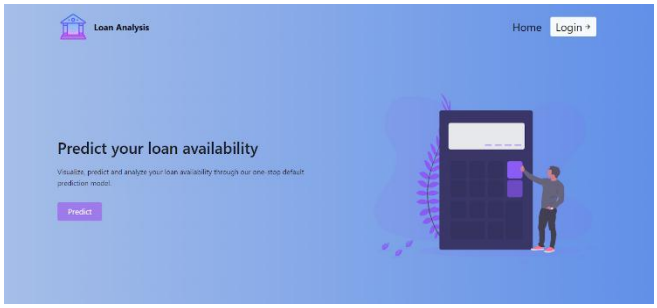


Fig. 5.1 Home Page

Login Page:

On this page for successful login, users have to give their perfect credentials as per the registration. After successful login bank users can use predict calculator or ML model for their client/customers.

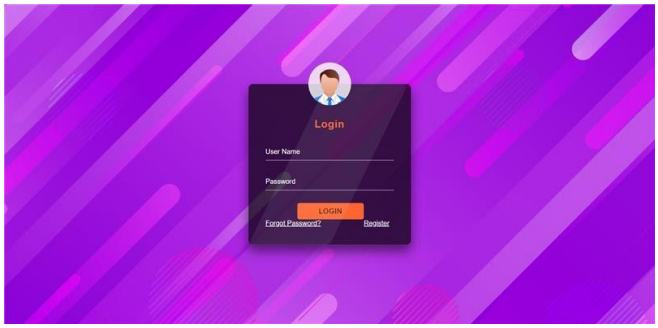


Fig. 5.2 Login Page

Register Page:

On this page, we ask for Basic details of users for registration with the database.

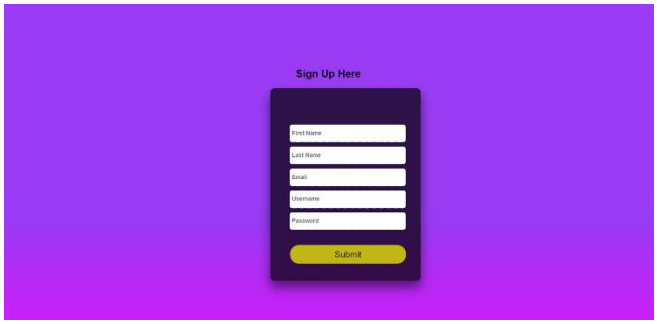


Fig. 5.3 Registration Page

Predict Page:

On this page we made one predicting calculator field for our ML model.

After filling all input fields we give the output as per the ML model results and Accuracy.

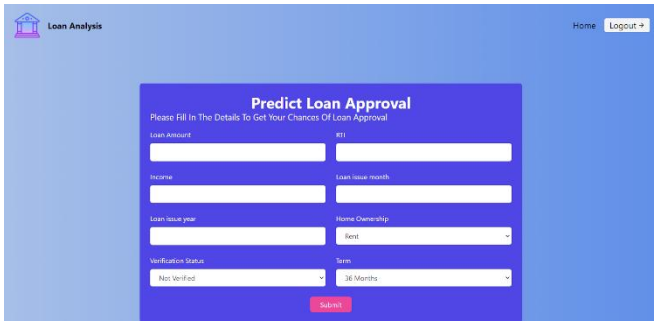


Fig. 5.4 Predict Page

6. FUTURE SCOPE

Since the domain of the project (LAPD) revolves around data science and machine learning, it is possible to add Natural Language Processing (NLP) based chat-bot which would make the project more interactive and would help the user in navigating the website or would recommend or resolve queries regarding the usage of the application. The chat-bot can be integrated either by using python language or by using the Dialogflow platform powered by Google. In our project (LAPD) we have developed a full-fledged website for the end-user to predict the loan default. An alternative to the website can be a mobile application which would prove much handier for the user. Cross-Platform technology frameworks such as Flutter which is also backed by Google can be brought into consideration for developing the application [4]. Apart from this, currently, the dataset which we have used is restricted to a particular organization. If it is possible to collect a wider range of data, it would not only act as a knowledge base for the model but also help the model in understanding and predicting the loan default correctly indirectly resulting in getting better accuracy

7. CONCLUSION

Due to a sudden halt caused by the COVID-19 pandemic, it has affected almost every possible business of industries thus leading to a financial dip. Many financial and banking firms both large scale and small scale are the prime sectors where the pandemic has largely impacted. The loan defaulters rate rose high because of the lockdown and not many were able to pay back. Hence, this invoked the need of developing a system for predicting the loan defaults thereby strengthening the loan sanctioning process.

This LAPD is thus an attempt to design and develop a credit risk scoring model which could analyze and predict the possibility of the loan default and thus reduce the rejection rate of the new loan applicants and capital loss faced by the bank which they would have made by sanctioning the loan to the clients

REFERENCES

[1] Ahmad Al-qerem, Ghazi Al-Naymat, Mays Alhasan, “Loan Default Prediction Model Improvement through Comprehensive Preprocessing and Features Selection”, published.

[2] P. Maheswari, CH. V. Narayana, “Predictions of Loan Defaulter - A Data Science Perspective”, published.

[3] Lin Zhua, Dafeng Qiua , Daji Ergua, Cai Yinga, Kuiyi Liub, “A study on predicting loan default based on the random forest algorithm”, published.

[4] Hafiz Ilyas Tariq Aziz, Asim Sohail, Uzair Aslam, “Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess (SEMMA)”, published.

[5] Mehul Madaan, Aniket Kumar, Chirag Keshri, Rachna Jain and Preeti Nagrath, “Loan default prediction using decision trees and random forest: A comparative study”, published.

[6] Harish Puvvada, Vamsi Mohan Ramineedi, “Loan Default Prediction”, published.

[7] Haotian Chen, Ziyuan Chen, Tianyu Xiang, Yang Zhou, “Data Mining on Loan Default Prediction”, published.

RESEARCH ARTICLE OPEN ACCESS

Accuracy Prediction for Loan Risk Using Machine Learning Models

Anchal Goyal [1], Ranpreet Kaur [2]
Research Scolar [1], Assistant Proffesor [2]
Department of Computer Science
RIMT –IET (PTU),Mandi Gobindgarh
Punjab - India

ABSTRACT
Extending credit to individuals is essential for markets and society to act efficiently. Estimating the probability that an individual would default on their loan, is useful for banks to make a decision whether to approve a loan to the individual or not. In this paper, we find the accuracy of several models in R language and evaluate it to establish the finest model to forecast the finance status for an organization. We did the experiment five times on the same data set and find the experimental results that show the Tree Model for Genetic Algorithm is the best model for forecasting the finance for costumers.
Keywords: - Accuracy, Prediction, Genetic algorithm, Finance.

I. INTRODUCTION

A. Introduction to Machine Learning

Machine learning is a arena of computer science that involves the learning of pattern identification and computational learning theory in AI. Machine learning generally refers to the changes in systems that carry out tasks linked with artificial intelligence (AI). Such tasks include recognition, analysis, planning, robot control, forecasting, etc. It explores the study and construction of algorithm that can make prediction on data. Machine Learning is used to build programs with its tuning parameters that are adapted consequentially so as to increase their functioning by adapting to earlier data.

Machine learning can be broken into two categories:

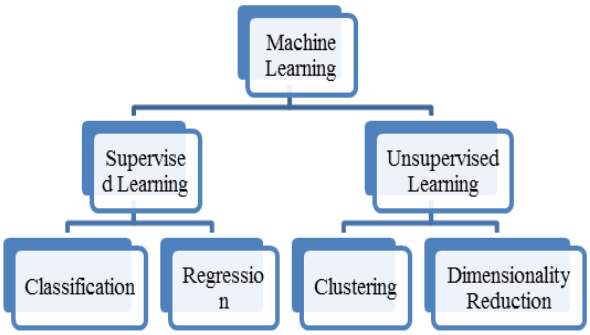


Fig.1 Categories of Machine Learning

1) **In Supervised Learning**, a data set includes both *features* and *labels*. The task is to build an estimator which is able to forecast the label of an object with the set of features. Supervised learning is further broken down into two parts: classification and regression.

Classification is the task of forecasting the value of a categorical variable given some input variables.

Regression is the task of forecasting the value of a continuously changeable variable (e.g. a price, a temperature) given some input variables.

2) **In Unsupervised Learning,** a data set has no label and we find similarities among the objects. We can use this technique to display the best arrangement of data. It includes tasks such as dimensionality reduction, clustering.

Dimensionality reduction is the task of derive a set of new features that is smaller than the original feature set while hold large of the changing of the original data.

Clustering is the method of gathering samples into groups of analogous samples according to some predefined similar or dissimilar measure.

B. Introduction to R Language

R is a programming language produced by Ross Ihaka and Robert Gentleman at the University of Auckland. R is a GNU project and the source code of R is written in

C, FORTRAN. R is basically a data analysis software environment for statistical computing i.e. interface between the statistics and computer science. R can be extended easily by packages that are available in CRAN package repository. R and its libraries implement a large variety of statistical and graphical techniques including time series analysis, clustering, classification, linear and non linear modeling. Basically R is freely available under GNU General Public License and precompiled binary versions are provided for various operating system. This language is mostly used by data miners and statisticians for developing the software.

1) The R environment

R is integrated software for manipulation of data, doing calculation and for graphical display. It includes

- an effective storage resource
- an effectual data handling resource
- a well developed programming language that include loops, functions and other input output facilities.
- Simple and effective language for users
- a huge and integrated collections of tools for analysis of data

C. Accuracy

Accuracy depends on how data is collected, and judged on basis of comparison of several parameters. True positive (TP) depicts amount of predictions which are positive, the actual value being positive. Similar in the case of true negative (TN). The accuracy is computed as:

$$Accuracy = \frac{TP + TN}{Toatal\ Data} * 100$$

When we build a predictive model, we need a way to evaluate the capability of the model on the data. This is done by estimating the accuracy of the model. The Caret Package in R provides a large no. of methods for estimating the accuracy of machine learning algorithms.

II. DATA SET AND FEATURES

The data set include 13 attributes such as Gender, Marital Status, Education, Income, Loan Amount, Credit History and others which are shown in table1 & the sample date set is shown in table2:

Table1: Feature Description

Feature ID	Features	Information
F1	Loan ID	Unique Loan ID
F2	Gender	Male/ Female
F3	Married	Applicant married (Y/N)
F4	Dependents	Number of dependents
F5	Education	Applicant Education (Graduate/ Under Graduate)
F6	Self-employed	Self employed (Y/N)
F7	Applicant Income	Applicant income
F8	Co applicant Income	Co applicant income
F9	Loan Amount	Loan amount in thousands
F10	Loan Amount Term	Term of loan in months
F11	Credit History	credit history meets guidelines
F12	Property Area	Urban/ Semi Urban/ Rural
F13	Loan Status	Loan approved (Y/N)

III. MACHINE LEARNING MODELS

Various machine learning models that have been applied for the prediction of accuracy as explained below:

1. Decision Tree Model

A decision tree model is one of the most frequent data mining models. It is popular because it is easy to understand. Decision trees are one of the useful algorithms that are used for regression and classification. They are also known as glass-box model. When the model once found the template in the data then we can see what the decision will be made for that data which we want to predict.

2. Linear model

A linear model is the one of the method for fitting a statistical model to data. It is appropriate when the target variable is numeric and persistant. This model helps to analyze the data and also helps to recognize and predict the performance of the complicated system.

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	0
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	1
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	1
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	1
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	1

Table 2: Sample dataset

Table 3: Machine learning models used.

Models	Packages	Tuning Parameter(s)
Bagged CART	Iperd	None
Random Forest	Randomforest	Mtry
Tree Model For Genetic Algorithm	evtree	Alpha
Decision Trees	rpart	Min split, max depth, min bucket
Linear Model	Car, nnet	Size
Neural Network	nnet	Size, decay
SVM	Kernlab	C
Extreme Learning Machine	elmNN	Nhid,actfun
Multivariate Adaptive Regression spline	earth	Degree
Bayesian Generalized Linear Model	arm	None
Model Tree	tree	None

3. Random Forest Model

A random forest model is basically a collection(i.e. ensemble) of tens or hundrerds of decision trees. These models are mainly used if we have large no. of input variables i.e. in hundreds and thousands and if we have very vast dataset. This model is very efficient if we have large no. of variables and it distributes the variable into different subsets. Ensemble models are robust to variance and bias.

4. Neural Network Model

This model is basically based on various layers that are connected to each other like neurons. This model combines the numbers and provides the numeric data to produce the final results throughout the network. These models are identical to biological neural network in order to perform functions parallel and collectively rather than individually.

5. Support vector machine

SVM is supervised machine learning model with learning algorithms which examine the data and uses that data for regression and classification. This model uses a technique namely a kernel trick to transform the data and based on these transforms of data, it finds the best optimum results. It is not considered as better as than the other machine learning models because it works on less data set.

6. Extreme learning machines

ELM is a modification is a feed forward network with single layer which have a hidden nodes for single layer. The Weights are randomly given to hidden nodes and it never be updated. The name to this model was given by Guang-Bin Huang. Different from other traditional models, the extreme learning model not only provide the smaller training error but also better performance.

7. Multivariate Adaptive Regression Splines

This model is established by Jerome H. Friedman in 1991. This model is used for both regression and classification type problem with the purpose to predict the values. The 'earth' package is used in implementation of this model. The earth source code is licensed under the GPL. This technique has popular in data mining because it is used to find the difficult data mining problems.

8. Model tree

Model tree is a classification model that is combination of decision tree learning and logistic regression model. The package named 'tree' is used in implementation of this model. This model tree works on when have to predict the numeric quantities. It is a tree that include linear regression function at their leaves.

9. Bayesian Generalized Linear Model

BGLM is most generally used technique for creating the relationship. This model is used when have huge dataset and BGLM is used to fit the dataset into

pragmatic size and remove the problem of over fitting. This model is included in package "arm" in r language.

10. Bagged Cart Model

This model is used for classification and regreession problems. This model build under the package 'iperd' and 'plyr'. Bagging for classification and regression trees were suggested by Breiman in 1996.

11. Tree model form Genetic Algorithm

Genetic algorithm is a search heuristic i.e. it is an algorithm for finding and solving a problem more quickly and produces the result in reasonable time. This model is very efficient, flexible and finds optimal solutions for given problem. This model builds under the package 'evtree'. This algorithm is usually based on theory of natural selection and survival of fittest. The larger the value of fitness is the most the optimal result will be.

IV. RESULTS

In it, we analyze the results of various different machine learning models which are implemented in R to find the accuracy of each model and find the best model for the bank that provides loan to the costumers. Accuracy is the most important aspect for any organization. The experimental results show that the Tree Model from Genetic Algorithm is the best model among the entire model for the given dataset for predicts the loan .Table 4 shows the accuracy of all the models.

Table 4: Accuracy Results

Models / No. of runs	**RUN1**	**RUN2**	**RUN3**	**RUN4**	**RUN5**
Decision Tree	75	80.56	76.39	81	79.86
Linear Model	73.61	75	70.14	79.17	78.47
Neural Network	75.69	81.25	77.02	83	79.86
Random Forest	76.39	79.86	77.08	82.64	80.56
SVM	77	79.56	76.39	81.94	80.56
Bagged Cart	77.08	74.47	76.39	79.17	78.47
Tree model for genetic algo	79.5	81.75	78.5	83.33	81.25
model tree	73.61	74.31	69.44	70.83	79.86
Extreme learning machine	57.64	69.44	65.97	73.61	68.75
Multivariate Adaptive Regression Spline	75.69	79.17	77.5	81.23	79.86
BGLM	76.39	81.25	76.39	83.33	79.86

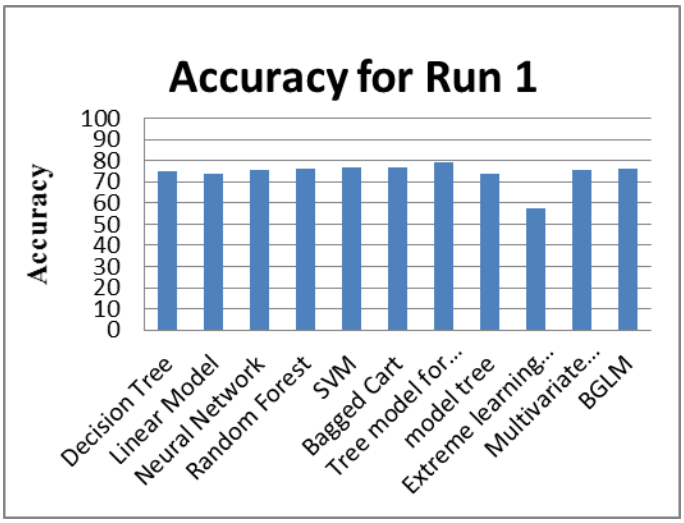


Fig. 2 Accuracy for Run 1

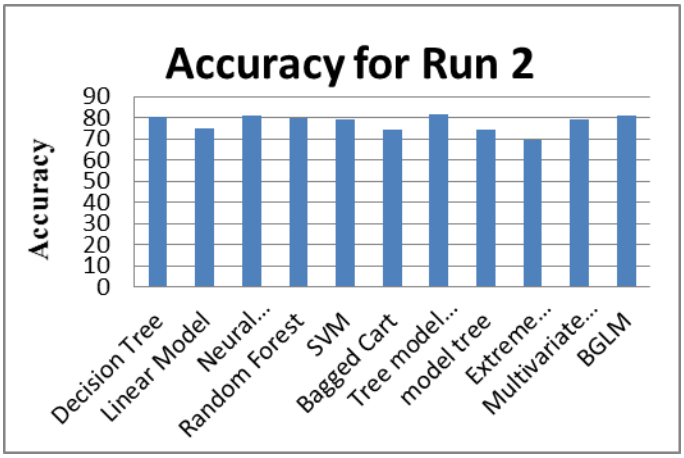


Fig. 3 Accuracy for Run 2

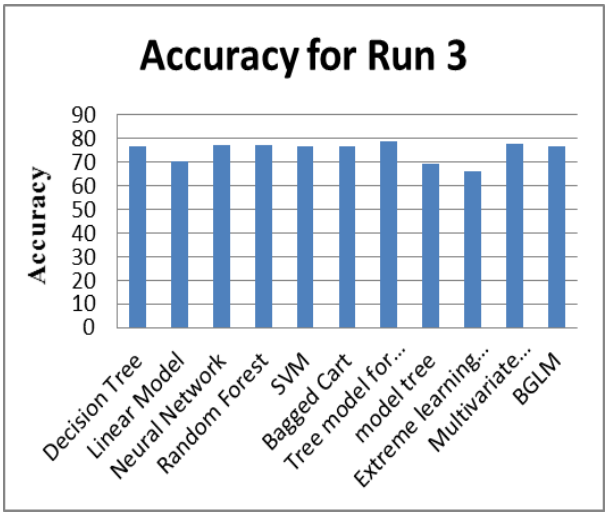


Fig. 4 Accuracy for Run 3

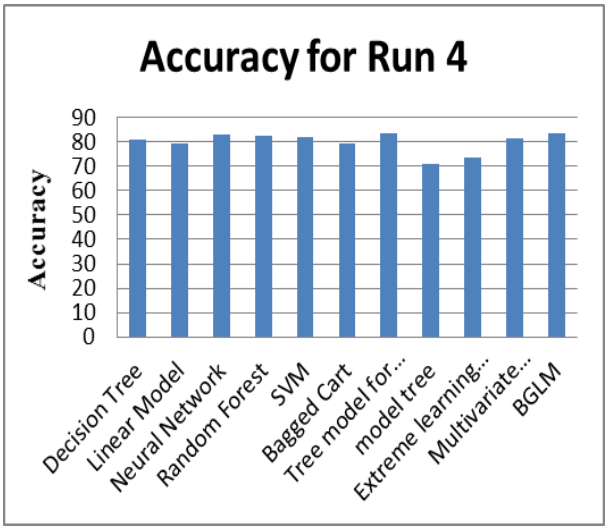


Fig. 5 Accuracy for Run 4

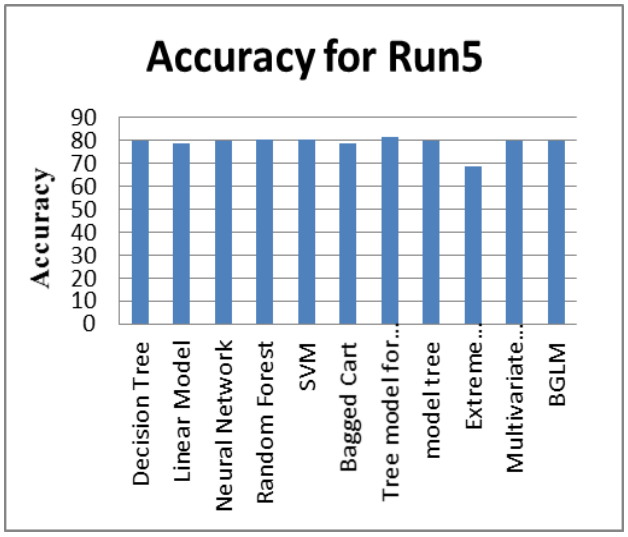


Fig. 6 Accuracy for Run 5

V. CONCLUSION

In this paper, we find the accuracy of several models in R language and evaluate it to establish the best model to predict the finance status for an organization. We did this experiment five times on the same data set having different seed values and the accuracy varies according to its seed value that is shown in figures 2 to 6. The experimental results that show the Tree Model for Genetic Algorithm is the best model for forecasting the loan for costumer.

REFRENCES

[1] Wo- Chiang Lee, " Genetic Programming Decision Tree for Bankruptcy Prediction", JCIS-Oct 2006, 1951-6851.

[2] Breiman, L, "Bagging predictors". Machine Learning. .

[3] Breiman, L, "Random forests". Machine Learning, 2001, 5–32.

[4] Lim, T.-S., Loh, W.-Y., & Shih, Y.-S. ." A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms,0ct 2000, 203–228.

[5] Niculescu-Mizil, A., & Caruana, R." Predicting good probabilities with supervised learning". Aug2005.

[6] Jason Brownleek. "Compare Machine Learning Models", Sep24, 2014.

[7] K. Bache and M. Lichman, UCI machine learning repository.

[8] H. Faris, B. Al-Shboul and N. Ghatasheh, "A Genetic Programming Based Framework for Churn Prediction in Telecommunication Industry",Sep 2014, 253-362.

[9] D. Fantazzini and S. Figini, "Random Survival Forests Models For SME Credit Risk Measurement," Methodology and Computing in Applied Probability, Jan 2009, 29-45.

Indonesian Journal of Electrical Engineering and Computer Science
Vol. 28, No. 2, November 2022, pp. 1002~1011
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v28.i2.pp1002-1011

An improved light gradient boosting machine algorithm based on swarm algorithms for predicting loan default of peer-to-peer lending

Much Aziz Muslim[1,2], Yosza Dasril[1], Muhammad Sam'an[1], Yahya Nur Ifriza[3]
[1]Department of Technology Management, Faculty of Technology Management and Business, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia
[2]Department of Computer Science, Universitas Negeri Semarang, Semarang, Indonesia

Article Info

Article history:

Received Apr 18, 2022
Revised Aug 26, 2022
Accepted Sep 5, 2022

Keywords:

Ant colony optimization
Bee colony optimization
Features selection
LightGBM
Loan default
P2P lending

ABSTRACT

Internet finance and big data technology are booming in the world. The launch of peer to peer (P2P) lending platforms is a sign and a great opportunity for entrepreneurs to easily increase their capital injection. However, this great opportunity has a high risk of impacting the sustainability and security development of the platform. One way to minimize loan risk is to predict the possibility of loan default. Hence, this study aims to find the best predictive model for predicting loan default of P2P Lending Club dataset. An improved light gradient boosting machine (LightGBM) via features selection by using swarm algorithms i.e. Ant colony optimization (ACO) and bee colony optimization (BCO) to the prediction analysis process. The best feature selection process is selected 6 out of 18 features. The synthetic minority oversampling technique (SMOTE) method is also provided to solve the unbalance class problem in the dataset, then a series of operations such as data cleaning and dimension reduction are performed. The experimental results prove that the LightGBM algorithm has been successfully improved. This success is shown by the prediction accuracy of LightGBM+ACO is 95.64%, LighGBM+BCO is 94.70% and LightGBM is 94.38%. This success also demonstrates outstanding performance in predicting loan default and strong generalizations.

Corresponding Author:

Much Aziz Muslim
Postgraduate Student, Faculty of Technology Management and Business
Universiti Tun Hussein Onn Malaysia (UTHM)
Batu Pahat, Johor, 86400, Malaysia
Email: a212muslim@mail.unnes.ac.id

1. INTRODUCTION

Big data and Internet finance (Fintech) are currently trending and are often discussed in the world as internet lending industry is known as peer to peer (P2P) Lending. P2P lending as a Fintech platform has a unique characteristic in transactions, namely connecting individual loan borrowers to individual lenders or investors to make credit agreements and complete transaction procedures directly through the online platform, without commercial bank intermediaries. Gradually, the existence of P2P lending has become a solution for small and medium businesses to get loan capital so easily that every year the loan amount is very large. As reported by LendingClub Corporation [1] about "Fourth Quarter and Full Year 2019 Results" show that the loan amount had achieved US $ 12,290.1 billion at the end of 2019. While Stern *et al*. [2]'s data showed that China's government noted that China became the most P2P loan platforms predicate of the investment market with quantitating to around 2.300 as of March 2017 and CNY 9.208 loan volume.

P2P lending presents an opportunity as well as a challenge, including China as a developed economy. In order to a large extent, P2P lending meets China's current economic needs as well as its risks. Financial risk can be seen from liquidity risk caused by insufficient liquidity funds, unbalanced information as a cause of credit risk and legal risk caused by unclear laws governing Fintech. In brief, the risk characteristics of Fintech are more complex than conventional finance. In addition, technical and virtual-based Fintech also triggers special risks that arise such as conventional financial risks as financial risks are sudden and spreading, besides that the increase in destructive risks is very serious and uncontrollable Challa *et al*. [3] said risk aversion is one of the hot topics, interesting and very important to be discussed among investors, policymakers, financial practitioners and made studies by researchers.

Generally, research and application of loan evaluation in P2P lending platforms are given two main directions. First, the use of credit scores to evaluate the credit risk of loans and second, transforming loan evaluations into a binary classification. A credit scorecard is a conventional loan evaluation method. Usually, Chen and Han [4] explained these scorecards are self-launched by P2P lending platforms for business needs, for example, Fair Isaac Corporation (FICO) score and LendingClub score. However, according to Malekipirbazari and Aksakalli [5], credit scorecards cannot distinguish between defaulter and non-defaulter. As big data technology matures many researchers use machine learning techniques to predict whether a loan can be returned or a loan repayment is due in P2P lending platform. Light gradient boosting machine (LightGBM) is a machine learning algorithm that is used as a classification. LightGBM is an improved version of the gradient learning framework based on decision trees and "weak" learner ideas. Since being developed by Microsoft in 2017 [6]. Since LightGBM was introduced in 2016, several researchers have applied the big data machine learning Algorithm in various fields and produce predictions with very high accuracy, fast-computationally and well-performance in minimizing relative over-fitting. Such as, web search, Breast cancer to identify miRNAs [7], the default accuracy prediction of P2P lending platform [8]-[11], music recommendation [12], the classification of acoustic scene [13], smart grid load forecasting [14], estimation of reference evapotranspiration of agricultural or hydrological [15], construction cost prediction [16], predict customer loyalty Fintech [17] and stream processing prediction [18].

LightGBM is known as an algorithm that is fast data learning, faster when handling big data, high accuracy, good model precision, low data memory consumption so that this algorithm is considered more effective and efficient than other machine learning techniques [8], [19]. According to Rao *et al*. [20], feature selection in a big data set as a significant phase performs several tasks such as image classification, cluster analysis, data mining, pattern recognition, and image capture [21], [22]. Many methods have been proposed, improved and discussed for feature selection. Alickovic and Subasi [23]-[24] improved whale optimization algorithm (WOA) to optimize features in the dataset. Zhu *et al.* [25] presented a method of uncontrolled spectral feature selection to maintain local and global features of the feature during the redundant feature removal process. Wan and Freitas [26] evaluated the hierarchy method in optimizing the feature selection of aging related gene data sets. Rao *et al*. [20] used artificial bee colony and gradient boosting decision tree to select features of eight UCI data sets and produced and the experimental results proven that Rao's method is able to reduce the dimensions of the data set and achieve superior classification accuracy. Ghosh *et al*. [27] improved the wrapper-filter feature selection method based on ant colony optimization to reduce computational complexity.

Based on the previous research described above, increasing prediction accuracy via feature selection techniques is focus of this study. Therefore, we use two swarm algorithms, i.e. ant colony optimization (ACO) algorithm and Bee Colony Optimization (BCO) algorithm as a feature selection and LightGBM as a tool to evaluate P2P lending data sets. This study aims to determine the two swarm algorithms performance in the feature selection process, then the prediction performance of the LightGBM algorithm. In addition, we also use the synthetic minority oversampling technique (SMOTE) to address data class imbalances. This technique is believed to also be able to improve the accuracy of predictions as has been proven by Faris *et al.* [28] to predict the bankruptcy of companies with highly imbalanced data classes.

2. RELATED KNOWLEDGE AND THEORY

2.1. Lending club

The lending club has lanced an impact on risk management. Loan applications can be approved are very small, around 10% of all applications. In addition, there are lending club levels i.e. A to G to classify loans based on risk. The main role of the Lending club is to make it easier for borrowers and lenders or investors to transact and provide information related to However. In fact, there are many problems in this transaction model, such as loan money is not returned by the borrower according to the agreement so that investors experience losses. Determining loan interest rates according to loan credit and loan term. The Lending club business pattern is shown in Figure 1.

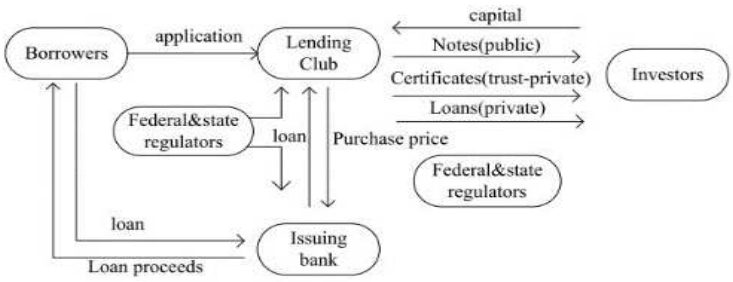


Figure 1. Lending club business pattern [29]

2.2. Bee colony optimization

Bee colony optimization (BCO) is one of bio-inspired methods by bees gathering nectar behavior [30]. Akay and Karaboga [31] said that the value of global optimum is determined by neighborhood search optimization of each bee. Wen *et al*. [32] said that artificial bee colony method able to locate the global optimum solution for the global optimization problems. When compared to other bio-inspired heuristic algorithms, BCO has many strengths i.e. a simple structure, requires few control parameters and is easy for implementing [33]. Because of this strength, BCO has attracted the attention of researchers to study and apply it in various fields.

2.3. Ant colony optimization

Ant colony optimization (ACO) is one of bio-inspired algorithms by ant colony behavior [34]. Ants cannot see. However, through indirect communication, the ants can find the shortest route from nest to the food source [35]. Ants modify their environment (by disguising pheromone) to influence another ant behavior is named Stigmergy. The concept of ACO algorithms for foraging ant behavior. Algorithms often discussed and applied are ant system (AS), ant colony system (ACS), max-min ant system (MMAS). In solving the optimization problem using the ACO algorithm, several artificial ants are used to model the solution iteratively. For each iteration, the ants will store a certain amount of pheromone which is proportional to solution quality. In each rarity, Tabakhi and Moradi [36] explained that the ant calculates a series of feasible solutions to the current partial solution and one of the choices depends on two factors i.e. local heuristics and prior knowledge, three phases need to be addressed i.e. Graph representation, Heuristic desirability and Pheromone update rule.

2.4. Light gradient boosting machine

Light gradient boosting machine (LightGBM) is a fast and efficient gradient boosted decision tree (GBDT) algorithm with an open-source promotion work objective that was created by Microsoft MSRA in 2016. This algorithm is used for sorting, classification, regression, and many other machine learning techniques assignments and supports efficient parallel training. In contrast to Xtream Gradient Boosted (XGBoost), LightGBM algorithm uses a histogram to speed up the training process, reduce memory space, and implement a wise growth strategy with depth constraints. The basic idea of LightGBM using a histogram is to discrete the continuity of floating-point eigenvalues to k bins and create a histogram with a width of k. LigthGBM does not require large storage of pre-sorted results, can store 8-bit integers and can also reduce memory consumption to 1/8 of the original. This rough partition does not reduce the mode of LigthGBM accuracy. The LightGBM is a boosting type that has three steps. For simplicity, X is given as a pre-processed streaming data set.

Step 1. Initialize the weak learner by (1).

$$f_0(x) = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, c) \quad (1)$$

where: $f_0(x)$ as the weak learner basis function, $L(y_i, c) = L(y, f(x)) = (y - f(x))^2$ as the function of loss, n as the amount of samples.

Step 2. Calculate weak learners M times, Iteratively.

a. For the sample $x_i \in X \forall i = 1,2, \dots, n$ calculate the negative gradient of loss function evaluated in the existing model in (2).

$$r_{mi} = -[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}]_{f(x)=f_{m-1}(x)} \quad (2)$$

where r_{mi} as negative gradient of the loss function.

b. The residual r_{mi} resulted is taken as sample new real value. Fit a regression tree for;

$\{(x_1, r_{m1}), \dots, (x_n, r_{mn})\}$ and make a new regression tree $f_m(x)$.

c. Calculate the best-fit value of the leaf area $j = 1,2, \dots, J$. By using c_{mj} in (3) as linear search to predict leaf node region value for minimizing the loss function.

$$c_{mj} = \text{argmin}_c \sum_{x_i=R_{mj}} L(y_i, f_{m-1}(x_i + c)), i = 1, \dots, M. \quad (3)$$

d. Update the robust learner by using (4).

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} c_{mj} I(x \in R_{mj}) \quad (4)$$

where $f_m(x)$ as the the existing weak leaner, $f_{m-1}(x)$ as pre-weak leaner, I as the indicator function.

Step 3. Determine the final regression tree by using (5).

$$F(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I(x \in R_{mj}) \quad (5)$$

The significance of a feature is calculated as the normalized total reduction of criterion brought by that feature. It is also known as the Gini significance Gini is denoted by Gini (p) in (10).

$$Gini(p) = \sum_{l=1}^{L} P_l(1 - p_l) = 1 - \sum_{k=1}^{L} p_l^2 \quad (6)$$

where: L as the number of labels p_k as the weight of l-label.

3. RESEARCH METHOD

The research method of loan default of P2P lending prediction analysis uses several phases, i.e. Dataset pre-processing, data oversampling, ensemble classification and performance evaluation. Generally, the research framework can be shown in Figure 2.

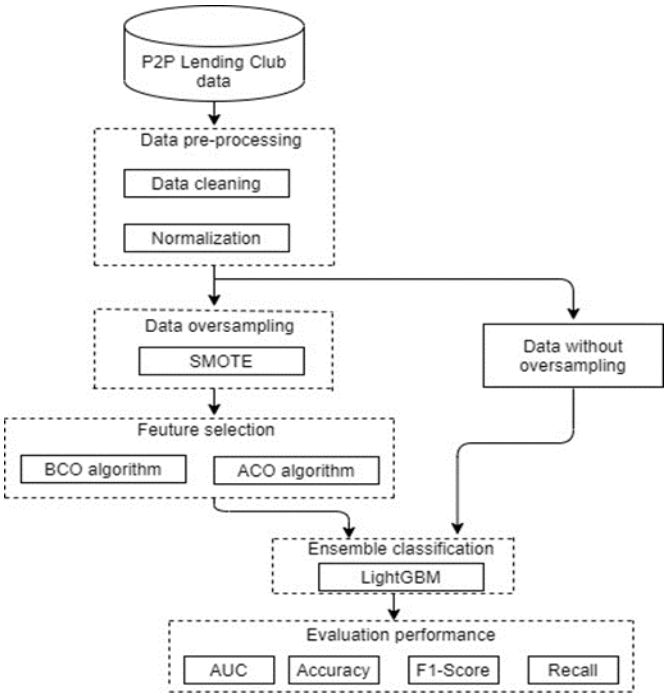


Figure 2. Framework of study

3.1. Dataset pre-processing

Data pre-processing is a sequence of process parts that are practiced to prepare the dataset for analysis and modeling. Therefore, this phase is believed to be an important step in the data mining process [37]. In this study, data preprocessing includes data cleaning, data normalization, and data retrieval. In data cleaning, missing values, inconsistencies and noise (e.g., incorrect data input) are eliminated [38]-[41]. We use the Lending club data set for the 2019 quarter downloaded from Kaggle.com containing 20.875.146 original user loans with 18 attributes. Furthermore, after data pre-processing, the missing value is filled via interpolation mode and multiple or not effect attributes are removed so that we get six attributes and Table1 shows the attributes used in the experiment.

Table 1. The selected attributes and pre-processing

Feature name	Description and pre-processing	Type	Algorithm
amount_borrowed	the principal amount of the loan upon which interest will accrue	numeric	A,B,C
borrower_rate	the interest rate at which money may be borrowed	numeric	A,B,C
installment	the monthly payment owed by the borrower if the loan originates	numeric	A,B,C
principal_paid	a payment toward the original amount of a loan that is owed	numeric	A, B, C
interest_paid	a payment of interest on a loan or mortgage	numeric	A, B, C
grade	lending club assigned loan grade	nominal	B
term	the loan repayment amount the Value represented 36 months from binary number to discretization	numeric	A
loan_status	the source of our answer to the core question if people are paying the loans they take out	nominal	C

Note: A = LighGBM without swarm algorithms, B = LighGBM with BCO algorithm, C = LighGBM with ACO algorithm

3.2. Synthetic minority oversampling technique

Based on data pre-processing 3.1, significant differences in the number categories of normal and default on target variable 'loan_status' can complicate learning modeling. SMOTE is an oversampling method to overcome imbalanced data sets, the SMOTE rationale as follows [29],

a) To calculate the K-nearest neighbor of each minority sample with the Euclidean Distance as the standard, the neighbor algorithm is used.
b) Adjusting a sampling proportion with the unbalance sample proportion and each sample x minority class, a few samples are randomly selected from its K-neighbors.
c) Suppose x_n is the selected neighbor. For each randomly selected neighbor x_n, a new sample can be generated using (11) with the respective original samples.

$$x_{new} = x_i + rand(0,1) * |x - x_n| \quad (7)$$

By iteratively, for each sample x_i, the original sample size of minority class can be widened to an ideal ratio.

3.3. Feature selection

First, we define the "installment" feature to represent the user's monthly fee payment as a percentage of their monthly revenue. The greater the "installment" value, the more loans provided by investors will be more burdened and tend to default. Second, feature abstraction. We encoded the loan status 'Current',' Completed' as usual=0, encoding 'Default',' Charge off' and' Canceled' as default=1. Next, we plot loan_status. That 89% of loan_status is "Default" and the rest is only 11% for "Normal". Based on these results, it indicates a serious imbalance of datasets. After scaling the features, third is feature selection. The selected feature attributes have high relevance or correlation value and remove irrelevant features or low correlation. This elimination can reduce difficulties in the training process. We use swarm algorithms i.e. ACO and BCO to select 6 features with the strongest correlation with the target variable and remove features step by step to achieve the reduction of the first dimension with variables 18 to 6. We illustrate a Pearson correlation graph of 18 features, as shown in Figure 3.

Meanwhile, the results of the reduction of the first dimension, the redundant features are selected and removed using the Pearson correlation graph based on the swarm algorithm used. The feature dimensions reduced from 18 to 6 are shown in Figure 4, Figure 4(a) shown that features selection of BCO algorithm is amount_borrowed, borrower_rate, installment, principal_paid, grade and on Figure 4(b) shown that features selection of ACO algorithm is amount_borrowed, borrower_rate, installment, principal_paid and loan_status.

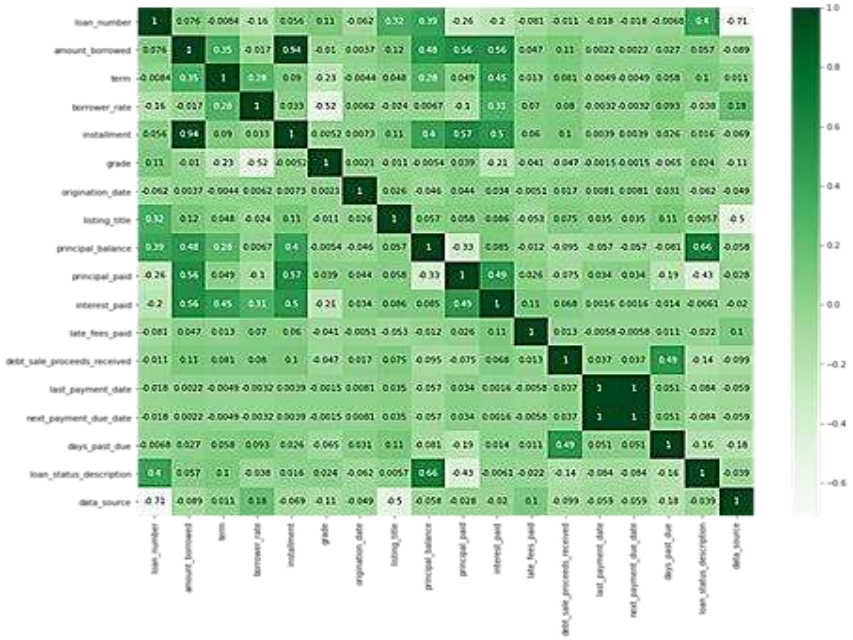
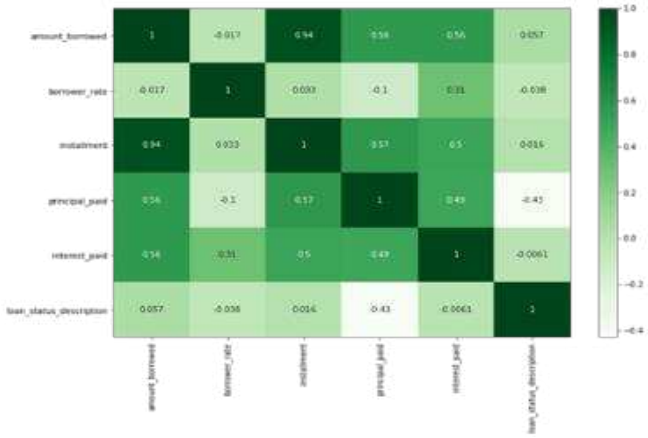


Figure 3. Person correlation of 18 features



(a)

(b)

Figure 4. Person correlation of 6 features, (a) The features selection of BCO algorithm and (b) The features selection of ACO algorithm

Population correlation coefficient is formulated as the covariance and standard deviation between two variables. Predict covariance and standard deviation of sample to determine the Pearson correlation coefficient of sample. Finally, we use the swarm algorithm i.e. ACO and BCO to select the importance of the feature and reduce the learning difficulty to optimize the model calculation.

3.4. The evaluation performance model

In this study, we use three parameters i.e. accuracy, AUC and ROC to evaluate and assess the performance of our proposed model. Accuracy is the ratio of the number of correct sample classifications to the total number of samples for a particular test data set as shown in (8).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

where: TP=True Positives, TN=True Negatives, FP=False Positives and FB=False Negatives.

Recall is called the fraction of all positive instances (default) where the classifier categorizes true as positive or known as the TP ratio. A balanced F score or F1-score is called the balanced average of Precision and Recall.

3.4.1. Receiver operating characteristic curve

In statistics, receiver operating characteristics or ROC known as a two-dimensional graphical plot illustrates the performance of a binary classifier. The curve of ROC is made in various threshold settings by plotting true positive ratio (TPR) to the false positive ratio (FPR) by using (9). Intuitively, this curve represents the performance of the classifier.

$$TPR = \frac{TP}{TP+TN} FPR = \frac{FP}{FP+TN} \tag{9}$$

3.4.2. AUC value

The AUC represents the area under the curve of ROC in the test data-set. Suppose that the curve of ROC is formed by a sequential relationship of points with coordinates of $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m\}$. Thus, the value of AUC can be formulated by using (10).

$$AUC = \frac{1}{2}\sum_{i=1}^{m-1} (x_{i+1} - x_i) \bullet (y_i + y_{i+1}) \tag{10}$$

where the AUC value range is [0.5,1.0] and if the AUC value is almost 1.0 then the classifier has a good performance.

4. RESULTS AND DISCUSSION

In research, the improved LightGBM algorithm as a classifier via features engineering or feature selection using an swarm algorithm i.e. ACO and BCO are evaluated and assessed their performance using several parameters i.e. accuracy, AUC, F1-Score, recall and ROC curves. The results obtained are shown in Table 2.

Table 2. The evaluation metrics comparison of the proposed model

Classifier model	Accuracy	AUC	F1-score		Recall		Rank
			0	1	0	1	
LighGBM+ACO	95.64 %	0.956	0.97	0.97	0.96	0.97	1
LighGBM+BCO	94.70 %	0.947	0.93	0.93	0.94	0.93	2
LighGBM	94.38 %	0.943	0.90	0.90	0.90	0.92	3

Table 2 shows that the performance of the LightGBM algorithm increases after the application of feature selection using the swarm algorithm. The performance of LightGBM+ACO algorithm is superior to LightGBM+BCO algorithm and LightGBM without swarm algorithm. Precision and Recall prediction models based on LightGBM using either the evaluation algorithm or not, all above 0.90. This value indicates that the model has strong generalizability. Meanwhile, the ROC curve graph is illustrated in Figure 5. This table shows that the closer the ROC curve is to upper left corner, the higher the prediction rate of model. The point of the ROC curve closest to upper left corner is best classification with lowest error based on the maximum threshold and the least total number of FPR and TPR. So from the curve, we can conclude that the LightGBM+swarm is superior to LightGBM.

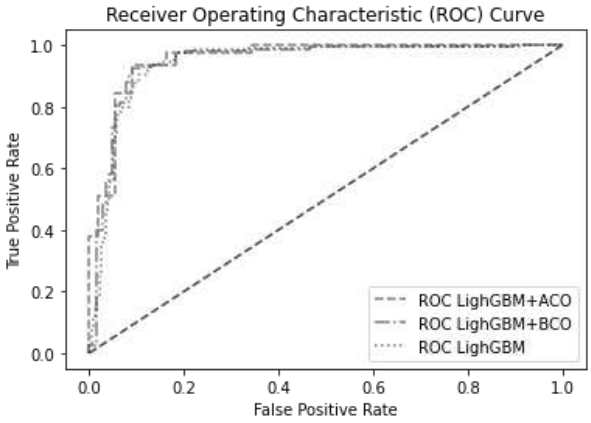


Figure 5. The ROC curve performance comparison LightGBM+ACO, LightGBM+BCO and LightGBM

5. CONCLUSION

In this study, the LightGBM algorithm is improved through feature engineering or feature selection using the BCO algorithm and the ACO algorithm to create a P2P loan evaluation model, especially the prediction of credit defaults. The experiment uses data sets from kaggle.com to show that improved LightGBM is successful. The best feature selection process is selected 6 out of 18 features. The SMOTE method is also provided to solve the unbalance class problem in the dataset, then a series of operations such as data cleaning and dimension reduction are performed. The experimental results prove that the LightGBM Algorithm has been successfully improved. This success is shown by the prediction accuracy of LightGBM + ACO is 95.64%, LighGBM + BCO is 94.70% and LightGBM is 94.38%. This success also demonstrates outstanding performance in predicting loan default and strong generalizations.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude and appreciation to the Universiti Tun Hussein Onn Malaysia (UTHM) through the research grant TIER 1 (H777).

REFERENCES

[1] G. Attigeri, M. M. Manohara Pai, and R. M. Pai, "Framework to predict NPA/Willful defaults in corporate loans: A big data approach," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 5, pp. 3786–3797, Oct. 2019, doi: 10.11591/ijece.v9i5.pp3786-3797.

[2] C. Stern, M. Makinen, and Z. Qian, "FinTechs in China – with a special focus on peer to peer lending," *Journal of Chinese Economic and Foreign Trade Studies*, vol. 10, no. 3, pp. 215–228, Oct. 2017, doi: 10.1108/JCEFTS-06-2017-0015.

[3] M. L. Challa, V. Malepati, and S. N. R. Kolusu, "Forecasting risk using auto regressive integrated moving average approach: an evidence from S&P BSE Sensex," *Financial Innovation*, vol. 4, no. 1, p. 24, Dec. 2018, doi: 10.1186/s40854-018-0107-z.

[4] D. Chen and C. Han, "Comparative Study of online P2P Lending in the USA and China," *Journal of Internet Banking and Commerce*, 2012.

[5] M. Malekipirbazari and V. Aksakalli, "Risk assessment in social lending via random forests," *Expert Systems with Applications*, vol. 42, no. 10, pp. 4621–4631, Jun. 2015, doi: 10.1016/j.eswa.2015.02.001.

[6] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 3147–3155, 2017.

[7] D. Wang, Y. Zhang, and Y. Zhao, "LightGBM: An effective miRNA classification method in breast cancer patients," in *ACM International Conference Proceeding Series*, 2017, pp. 7–11, doi: 10.1145/3155077.3155079.

[8] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu, "Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning," *Electronic Commerce Research and Applications*, vol. 31, pp. 24–39, Sep. 2018, doi: 10.1016/j.elerap.2018.08.002.

[9] J. Zhou, W. Li, J. Wang, S. Ding, and C. Xia, "Default prediction in P2P lending from high-dimensional data based on machine learning," *Physica A: Statistical Mechanics and its Applications*, vol. 534, p. 122370, Nov. 2019, doi: 10.1016/j.physa.2019.122370.

[10] Y. Wang and X. S. Ni, "Improving investment suggestions for peer-to-peer lending via integrating credit scoring into profit scoring," in *ACMSE 2020 - Proceedings of the 2020 ACM Southeast Conference*, Apr. 2020, pp. 141–148, doi: 10.1145/3374135.3385272.

[11] Y. Song, Y. Wang, X. Ye, D. Wang, Y. Yin, and Y. Wang, "Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending," *Information Sciences*, vol. 525, pp. 182–204, Jul. 2020, doi: 10.1016/j.ins.2020.03.027.

[12] W. Zhang, H. Quan, and D. Srinivasan, "Parallel and reliable probabilistic load forecasting via quantile regression forest and quantile determination," *Energy*, vol. 160, pp. 810–819, Oct. 2018, doi: 10.1016/j.energy.2018.07.019.

[13] E. Fonseca, R. Gong, D. Bogdanov, O. Slizovskaia, E. Gomez, and X. Serra, "Acoustic scene classification by ensembling gradient boosting machine and convolutional neural networks," *Detection and Classification of Acoustic Scenes and Events (DCASE)*, no. November, pp. 1–5, 2017.
[14] Q. Zhang, N. Cui, Y. Feng, D. Gong, and X. Hu, "Improvement of Makkink model for reference evapotranspiration estimation using temperature data in Northwest China," *Journal of Hydrology*, vol. 566, pp. 264–273, Nov. 2018, doi: 10.1016/j.jhydrol.2018.09.021.
[15] J. Fan, X. Ma, L. Wu, F. Zhang, X. Yu, and W. Zeng, "Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data," *Agricultural Water Management*, vol. 225, p. 105758, Nov. 2019, doi: 10.1016/j.agwat.2019.105758.
[16] D. Chakraborty, H. Elhegazy, H. Elzarka, and L. Gutierrez, "A novel construction cost prediction model using hybrid natural and light gradient boosting," *Advanced Engineering Informatics*, vol. 46, p. 101201, Oct. 2020, doi: 10.1016/j.aei.2020.101201.
[17] M. R. Machado, S. Karray, and I. T. De Sousa, "LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry," in *14th International Conference on Computer Science and Education, ICCSE 2019*, Aug. 2019, pp. 1111–1116, doi: 10.1109/ICCSE.2019.8845529.
[18] Z. Chu, J. Yu, and A. Hamdulla, "LPG-model: A novel model for throughput prediction in stream processing, using a light gradient boosting machine, incremental principal component analysis, and deep gated recurrent unit network," *Information Sciences*, vol. 535, pp. 107–129, Oct. 2020, doi: 10.1016/j.ins.2020.05.042.
[19] M. A. Muslim, A. Nurzahputra, and B. Prasetiyo, "Improving accuracy of C4.5 algorithm using split feature reduction model and bagging ensemble for credit card risk prediction," in *2018 International Conference on Information and Communications Technology, ICOIACT 2018*, Mar. 2018, vol. 2018-January, pp. 141–145, doi: 10.1109/ICOIACT.2018.8350753.
[20] H. Rao *et al.*, "Feature selection based on artificial bee colony and gradient boosting decision tree," *Applied Soft Computing Journal*, vol. 74, pp. 634–642, Jan. 2019, doi: 10.1016/j.asoc.2018.10.036.
[21] B. Prasetiyo, Alamsyah, and M. A. Muslim, "Analysis of building energy efficiency dataset using naive bayes classification classifier," *Journal of Physics: Conference Series*, vol. 1321, no. 3, p. 032016, Oct. 2019, doi: 10.1088/1742-6596/1321/3/032016.
[22] A. Nurzahputra, M. A. Muslim, and B. Prasetiyo, "Optimization of C4.5 algorithm using meta learning in diagnosing of chronic kidney diseases," *Journal of Physics: Conference Series*, vol. 1321, no. 3, p. 032022, Oct. 2019, doi: 10.1088/1742-6596/1321/3/032022.
[23] E. Aličković and A. Subasi, "Breast cancer diagnosis using GA feature selection and rotation forest," *Neural Computing and Applications*, vol. 28, no. 4, pp. 753–763, Apr. 2017, doi: 10.1007/s00521-015-2103-9.
[24] M. Mafarja and S. Mirjalili, "Whale optimization approaches for wrapper feature selection," *Applied Soft Computing*, vol. 62, pp. 441–453, Jan. 2018, doi: 10.1016/j.asoc.2017.11.006.
[25] X. Zhu, S. Zhang, R. Hu, Y. Zhu, and J. Song, "Local and global structure preservation for robust unsupervised spectral feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 3, pp. 517–529, Mar. 2018, doi: 10.1109/TKDE.2017.2763618.
[26] C. Wan and A. A. Freitas, "An empirical evaluation of hierarchical feature selection methods for classification in bioinformatics datasets with gene ontology-based features," *Artificial Intelligence Review*, vol. 50, no. 2, pp. 201–240, Aug. 2018, doi: 10.1007/s10462-017-9541-y.
[27] M. Ghosh, R. Guha, R. Sarkar, and A. Abraham, "A wrapper-filter feature selection technique based on ant colony optimization," *Neural Computing and Applications*, vol. 32, no. 12, pp. 7839–7857, Jun. 2020, doi: 10.1007/s00521-019-04171-3.
[28] H. Faris *et al.*, "Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: a case from the Spanish market," *Progress in Artificial Intelligence*, vol. 9, no. 1, pp. 31–53, Mar. 2020, doi: 10.1007/s13748-019-00197-9.
[29] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Procedia Computer Science*, vol. 162, pp. 503–513, 2019, doi: 10.1016/j.procs.2019.12.017.
[30] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," *Applied Soft Computing Journal*, vol. 8, no. 1, pp. 687–697, Jan. 2008, doi: 10.1016/j.asoc.2007.05.007.
[31] B. Akay and D. Karaboga, "A modified Artificial Bee Colony algorithm for real-parameter optimization," *Information Sciences*, vol. 192, pp. 120–142, Jun. 2012, doi: 10.1016/j.ins.2010.07.015.
[32] G. K. Wen, Y. Bin Dasril, N. Bujang, M. Mohamad, M. D. H. Gamal, and L. C. Soon, "Hybridization gradient descent search with artificial bees colony algorithm in general global optimization problems," *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 4, pp. 999–1008, 2021.
[33] M. Schiezaro and H. Pedrini, "Data feature selection based on Artificial Bee Colony algorithm," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 47, Dec. 2013, doi: 10.1186/1687-5281-2013-47.
[34] M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri, "Text feature selection using ant colony optimization," *Expert Systems with Applications*, vol. 36, no. 3 PART 2, pp. 6843–6853, Apr. 2009, doi: 10.1016/j.eswa.2008.08.022.
[35] Y. N. Ifriza and M. Sam'an, "Performance comparison of support vector machine and gaussian naive bayes classifier for youtube spam comment detection," *Journal of Soft Computing Exploration*, 2021, [Online]. Available: https://shmpublisher.com/index.php/joscex/article/view/42.
[36] S. Tabakhi and P. Moradi, "Relevance-redundancy feature selection based on ant colony optimization," *Pattern Recognition*, vol. 48, no. 9, pp. 2798–2811, Sep. 2015, doi: 10.1016/j.patcog.2015.03.020.
[37] C. F. Tsai and K. C. Cheng, "Simple instance selection for bankruptcy prediction," *Knowledge-Based Systems*, vol. 27, pp. 333–342, Mar. 2012, doi: 10.1016/j.knosys.2011.09.017.
[38] G. Mogos and N. S. Mohd Jamail, "Study on security risks of e-banking system," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 2, pp. 1065–1072, Feb. 2020, doi: 10.11591/ijeecs.v21.i2.pp1065-1072.
[39] S. Kim and K. You, "Data analysis of financial burden index through KBO league FA pitcher's performance and contract amount size," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 3, pp. 2525–2562, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2555-2562.
[40] M. A. Muslim and Y. Dasril, "Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 6, pp. 5549–5557, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5549-5557.
[41] K. Budiman and Y. N. Ifriza, "Analysis of earthquake forecasting using random forest," *Journal of Soft Computing Exploration*, 2021, [Online]. Available: https://www.shmpublisher.com/index.php/joscex/article/view/51.

BIOGRAPHIES OF AUTHORS

Much Aziz Muslim PhD candidate in the faculty of management technology at Universiti Tun Hussein Onn Malaysia (UTHM). The scope of research he is currently working on is in the fields of Data Mining. besides that he is also a lecturer in the computer science department of the Universitas Negeri Semarang. He can be contacted at email: a212muslim@mail.unnes.ac.id.

Yosza Dasril received PhD and master's degree degree in Applied Mathematics from Univeriti Putra Malaysia and Bachelor Degree in Mathematics from Universitas Riau, Indonesia. He is a Lecturer at Faculty of Technology Management and Business, Universiti Tun Hussein Onn Malaysia. His research interests are in Optimization, Engineering Mathematics. He can be contacted at email: yosza@utem.edu.my.

Muhammad Sam'an graduated in Master of Mathematics from Universitas Diponegoro in 2018. Currently Lecturer at department of computer sciences, Universitas Muhammadiyah Semarang. He has interested in fuzzy optimization and operation research. He can be contacted at email: muhammad.92sam@gmail.com.

Yahya Nur Ifriza graduated in Master of Informatic System from Universitas Diponegoro in 2017. Currently, he is a Lecturer at department of computer sciences, Universitas Negeri Semarang. He has interested research in data mining and wireless sensor network. He can be contacted at email: yahyanurifriza@mail.unnes.ac.id.

Predicting Bank Loan Risks Using Machine Learning Algorithms

Maan Y. Alsaleem
maanyounis1983@gmail.com
Directorate of Education in Nineveh
Tenahî, Duhok, Ramy Land B9, Iraq

Safwan O. Hasoon
dr.safwan1971@uomosul.edu.iq
College of Computer Science and Mathematics
University of Mosul, Mosul, Iraq

Received on: 10/03/2020 **Accepted on: 25/03/2020**

ABSTRACT

Bank loans play a crucial role in the development of banks investment business. Nowadays, there are many risk-related issues associated with bank loans. With the advent of computerization systems, banks have become able to register borrowers' data according their criteria. In fact, there is a tremendous amount of borrowers' data, which makes the process of load management a challenging task. Many studies have utilized data mining algorithms for the purpose of loans classification in terms of repayment or when the loans are not based on customers' financial history. This kind of algorithms can help banks in making grant decisions for their customers. In this paper, the performance of machine learning algorithms has been compared for the purpose of classifying bank loan risks using the standard criteria and then choosing (*Multilayer Perceptron*) as it has given best accuracy compared to *RandomForest*, *BayesNet*, *NaiveBayes* and *DTJ48* algorithms.

Keywords: Bank loans, machine learning algorithms, Multilayer Perceptron, classification, accuracy, ROC.

التنبؤ بمخاطر القروض المصرفية باستخدام خوارزميات تعلم الالة

معن يونس عناد
مديرية تربية نينوى، الموصل، العراق

صفوان عمر حسون
كلية علوم الحاسوب والرياضيات
جامعة الموصل، الموصل، العراق

تاريخ استلام البحث: ٢٠٢٠/٠٣/١٠ **تاريخ قبول البحث: ٢٠٢٠/٠٣/٢٥**

الملخص

تلعب القروض المصرفية دوراً حاسماً في تطوير الأعمال الاستثمارية للبنوك. في الوقت الحاضر ، هناك العديد من القضايا المرتبطة بمخاطر القروض المصرفية. مع ظهور أنظمة الحوسبة ، أصبحت البنوك قادرة على تسجيل بيانات المقترضين وفقًا لمعاييرها. في الواقع ، هناك كمية هائلة من بيانات المقترضين ، مما يجعل عملية اتخاذ القرار مهمة صعبة. استخدمت العديد من الدراسات خوارزميات تنقيب البيانات لغرض تصنيف القروض من حيث الايفاء أو عدم الايفاء بسداد القرض بالاعتماد على بيانات المقترضين السابقة . يمكن لهذا النوع من الخوارزميات مساعدة البنوك في اتخاذ قرارات المنح لعملائها. في هذه الورقة ، تمت مقارنة أداء خوارزميات التعلم الآلي لغرض تصنيف مخاطر القروض المصرفية باستخدام المعايير القياسية ثم اختيار الشبكات العصبية متعددة الطبقات (Multilayer

Perceptron) حيث أنها أعطت أفضل دقة مقارنة بخوارزميات (NaiveBayes , BayesNet , RandomForest & DTJ48).

الكلمات المفتاحية: القروض المصرفية، خورزميات تعلم الالة، شبكات العصبية متعددة الطبقات، التصنيف، الدقة ,ROC.

I Introduction:

Granting loans is an essential part of the work of any bank. Most of the banks' profits come from the benefits that are taken on these loans and most of the capital is in them. These days many banks agree on the loan after verification and validation, but there is still no whether the applicant is the appropriate applicant [13, 14].

Machine learning algorithms have been used in many areas of business, business administration, human resource management and medical purposes and have shown good success in data mining and decision support systems [3].

In the proposed paper, the use of neuronal networks as one of the machine learning algorithms for the purpose of classifying bank loans in terms of risks depending on the data in banks for previous loans and training algorithms in classifying loans using the characteristics of the borrower and comparing the performance of the neuronal network algorithm with decision trees j48, random forests and statistical methods (NaiveBayes , BayesNet).

II Methods:

In this section, we summarize the machine learning algorithms that were used for the purpose of classifying and forecasting loan risks.

A. DT J48, is one of algorithms used for making a decision tree developed by Ross Quinlan [1, 8]. The tree is built in the same way as building Iterative Dichotomiser 3 (ID3), where the contract is chosen based on the concept of gain, where the attribute with the highest classification ability (highest gain) is considered as the root of the tree that is branched into leaves. These leaves also choose (in the same way) the attribute with the ability to rank higher than the remaining attribute at the next level. This separation continues until the entire tree is built [5, 11]. The attribute selection for each node is based on the following three measurements:

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i|t)\, log_2\, p(i|t) \quad (1)$$

$$Gini\ (t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2 \quad (2)$$

$$Classification\ error(t) = 1 - [p(i|t)] \quad (3)$$

Where c is the number of class and $p(i|t)$ indicates the probability of records belonging to that class.

B. Random Forest the RF approach, it is based on the creation of many taxonomy trees based on different subsets of data using random subsections of available variables. The overall result of this approach creating and refining a set of correct theories and assumptions represented by trees, and combine trees in a "forest of classifiers" that its final decision depends on the results of the different decision trees. An additional powerful advantage of this approach, it is based on decentralized group behavior without any central

or hierarchical learning structure [11]. Each tree is built similar to J48 and the final result depends on the average output of those trees, as in the following equation:

$$RF_i = \frac{\sum T_j}{N} \quad (4)$$

RF sub (*i*): the end result of feature *i* from all trees in the RF model
T sub (*j*): the output of tree for *i* in tree *j*

C. Bayes's theorem

It is one of statistical probability theories used to predict the occurrence of a particular event based on the attribute of that event. It can be performed by calculating the probability of each attribute and its impact on the occurrence of that event [10]. Bayes's theorem is mathematically represented by the following formula:

$$p(x) = \frac{p(y)p(y)}{p(x)} \quad (5)$$

Where is:

$p(x)$: The probability that *y* will occur if event *x* occurs.
$p(y)$: The overall probability of a result of that class.
$p(x)$: The probability that event will occur for all events within a particular attribute**.**

In this paper, two types of Bayes' theory have been used

1. *BayesNet*, it is a probabilistic graphical model that uses Bayesian reasoning to calculate probabilities. The Bayesian network relies on conditional dependence, causation, and inferring from random variables by calculating these probabilities and according to the influence of each factor.
2. *NaiveBayes*, is one of probabilistic classifiers family that based on Bayesian theorem. This model is based on the principle of a maximum a posteriori decision rule and takes. The probability of each attribute independently without considering the relationships between those attributes.

D. Multilayer Perceptron, it is a mathematical model that derives its principle from the way neurons work in the human brain. The network consists of a group of artificial cells linked by connections. The work of Neuronal networks is based on the principle of parallelism that enables the network to analyze many problems with multiple variables [8, 11]. The multi-layered neural network is composed of an input layer that in turn receives the input values for the network and a number of hidden layers depending on the network structure. In this research we use a network with one hidden layer and this layer is called hidden because it is considered as a black box for the user as its inputs are the outputs of the input layer and its results are the inputs of the last layer in the sense that its inputs and outputs are not visible to the user and finally the output layer which consists of one [11]. The mathematical formula of a neuron is the following equation:

$$Y_k = f\left(\sum W_{jk}X_j + BK\right) \quad (6)$$

Where is:

$\{X_1, X_2, \dots \dots X_j\}$: Input signal.
$\{W_1, W_2, \dots \dots W_j\}$: Weights for the neuron k.
BK : It represents bias that can be counted as one of the weights.
f : Activation function.

III Dataset

The dataset used for classification purpose entitled "German Credit data" collected from UCI repository that contains 1000 Instances, 11 attributes as shown in Table (1).

Table 1: German Credit dataset

No	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
1	67	male	2	own	NA	little	1169	6	radio/TV	good
2	22	female	2	own	little	moderate	5951	48	radio/TV	bad
3	49	male	1	own	little	NA	2096	12	Education	good
4	45	male	2	free	little	little	7882	42	furniture/equipment	good
5	53	male	2	free	little	little	4870	24	Car	bad
6	35	male	1	free	NA	NA	9055	36	Education	good
7	53	male	2	own	quite rich	NA	2835	24	furniture/equipment	good
8	35	male	3	rent	little	moderate	6948	36	Car	good
9	61	male	1	own	rich	NA	3059	12	radio/TV	good
10	28	male	3	own	little	moderate	5234	30	Car	bad

After the preprocessing step, dataset become 24 numerical attributes and 1 binary classifier because when converting columns with categorical data to numeric, they will become more than one column [4]. The dataset is divided into two subsets, 80% of the data for training, and then 20% of these data was used for testing. The chosen dataset contains two formats of data (original data, numerical data). The numerical dataset was used to compare it with various machine learning algorithms.

VI Implementation

The performance comparisons were applied to 1,000 cases, including 700 loan repayments and 300 payment default loans. The *weka* Version 3.8.4 environment was used for this research for the purpose of model building and testing [6]. The proposed algorithms were trained using a dataset consisting of 800 instances of loans through supervised learning and targeted data (YES, NO) YES in the case of loan repayment and NO in case of payment defaulted after training the algorithms, a trainer model for each algorithm was obtained. The algorithms were tested using a test set of 200 Instances then we obtained the results of each algorithm and analyzed those results as shown in the figure 2.

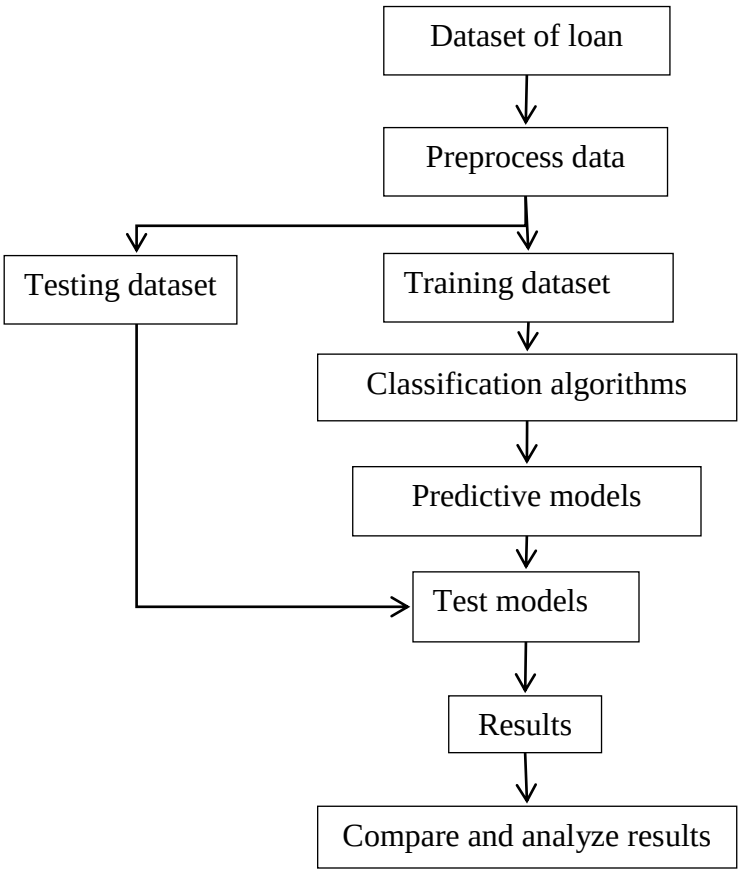


Figure (2): Proposed frame work

The results were analyzed by comparing the performance of each algorithm using several measurements as shown in Table 2 where it appeared that neuronal networks possess the highest accuracy compared to the other of the algorithms.

Table (2): Compare the performance of the algorithms

Classification algorithm	Root relative squared error	Relative absolute error	Root mean squared error	Mean absolute error	Kappa statistic	Incorrectly Classified Instances	Correctly Classified Instances
DT J48	105.3969%	76.8996%	0.4762	0.3198	0.3599	26.5 %	73.5 %
BayesNet	89.0378%	78.5019%	0.4023	0.3264	0.33	25 %	75 %
NaiveBayes	87.3059%	65.147 %	0.3945	0.2709	0.4268	22.5 %	77.5 %
RandomForest	85.048%	74.587 %	0.3843	0.3101	0.4271	21.5 %	78.5 %
Multilayer Perceptron	85.3731%	58.26%	0.3823	0.2411	0.4631	20 %	80 %

In machine learning algorithms, there are standard measurements used to explain the performance of each algorithm with respect to the targeted data. In this paper, there are two targeted values YES in the case of payment and NO in the case of default. The performance of each algorithm was as shown in Table 3.

Table (3): Performance measures of machine learning standard

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
DT J48	0.804	0.439	0.821	0.804	0.813	0.360	0.699	0.820	yes
	0.561	0.196	0.533	0.561	0.547	0.360	0.699	0.467	no
Weighted Avg	0.735	0.369	0.739	0.735	0.737	0.360	0.699	0.719	
BayesNet	0.881	0.579	0.792	0.881	0.834	0.338	0.788	0.899	yes
	0.421	0.119	0.585	0.421	0.490	0.338	0.788	0.550	no
Weighted Avg	0.750	0.448	0.733	0.750	0.736	0.338	0.788	0.799	
NaiveBayes	0.867	0.456	0.827	0.867	0.846	0.428	0.824	0.922	yes
	0.544	0.133	0.620	0.544	0.579	0.428	0.824	0.629	no
Weighted Avg	0.775	0.364	0.768	0.775	0.770	0.428	0.824	0.839	
RandomForest	0.902	0.509	0.816	0.902	0.857	0.436	0.819	0.915	Yes
	0.491	0.098	0.667	0.491	0.566	0.436	0.819	0.678	No
Weighted Avg	0.785	0.392	0.774	0.785	0.774	0.436	0.819	0.848	
Multilayer Perceptron	0.902	0.470	0.835	0.902	0.867	0.469	0.831	0.924	Yes
	0.530	0.098	0.673	0.530	0.593	0.469	0.831	0.659	No
Weighted Avg	0.800	0.367	0.791	0.800	0.792	0.469	0.831	0.851	

One important metric for measuring the performance of binary classification algorithms is Receiver Operating Characteristic (ROC), which showed the separatability of each algorithm depending on (true positive rate) and (false positive rate). Multi-layer neuronal networks have the potential for higher separation compared to the rest of the algorithms on this type of dataset as shown in Figure (3-7).

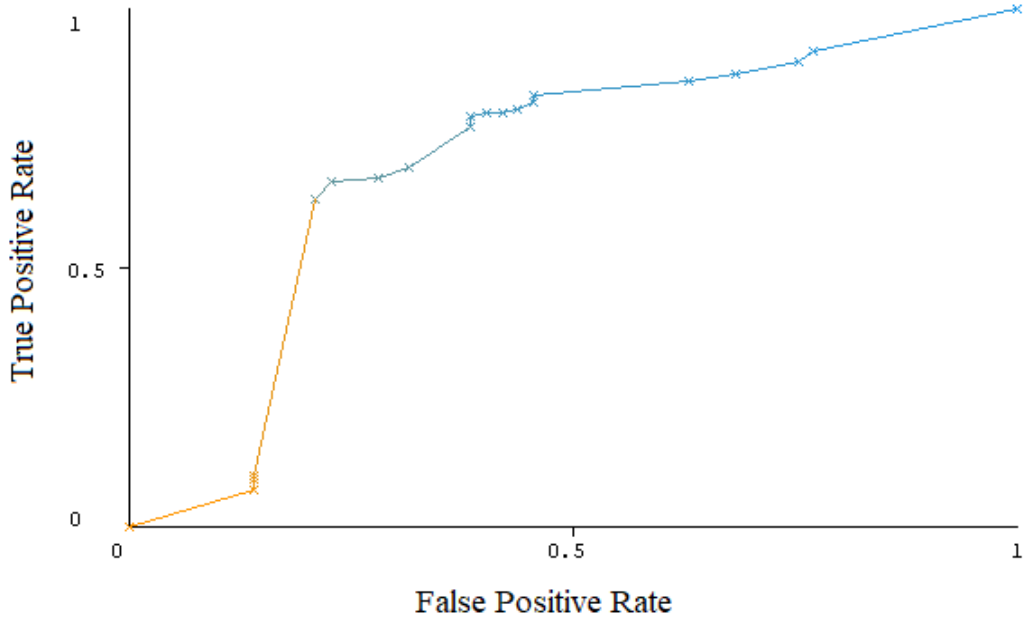


Figure (3): ROC curve for the model DT j48

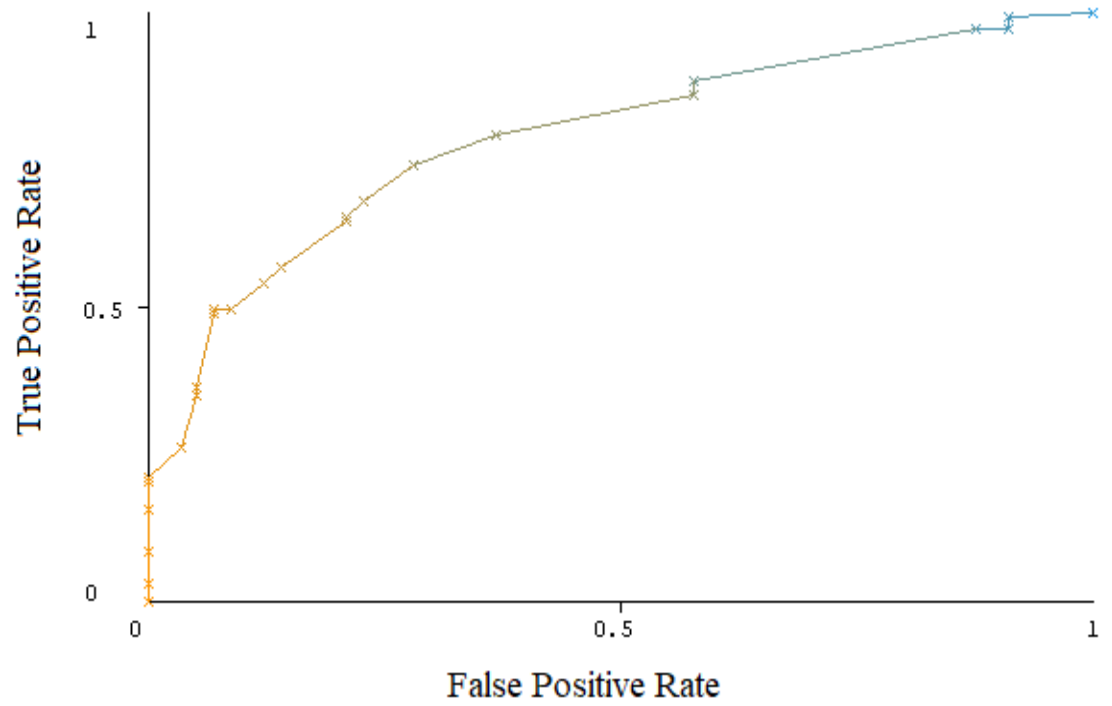


Figure (4): ROC curve for the model BayesNet

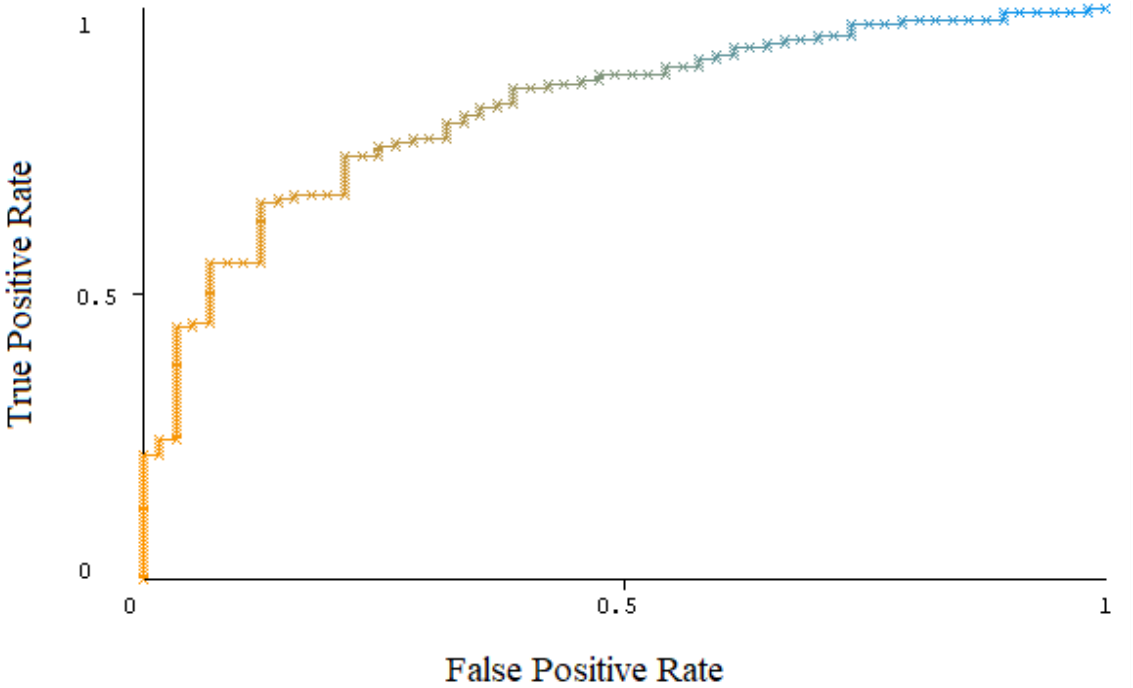


Figure (5): ROC curve for the model NaiveBayes

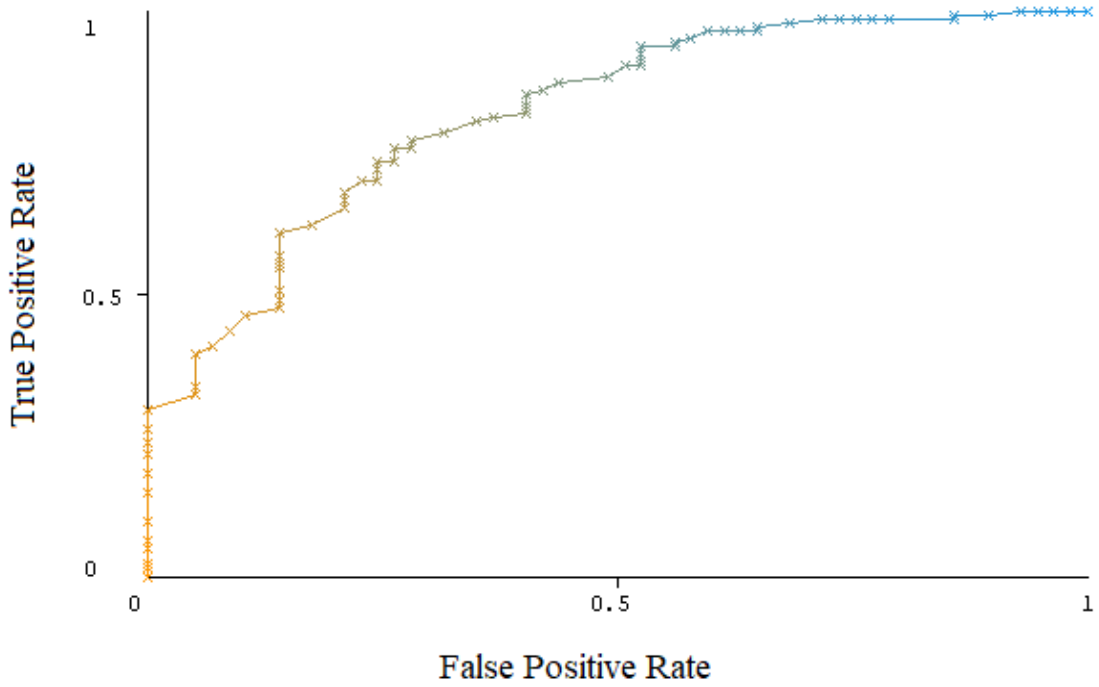


Figure (6): ROC curve for the model RandomForest

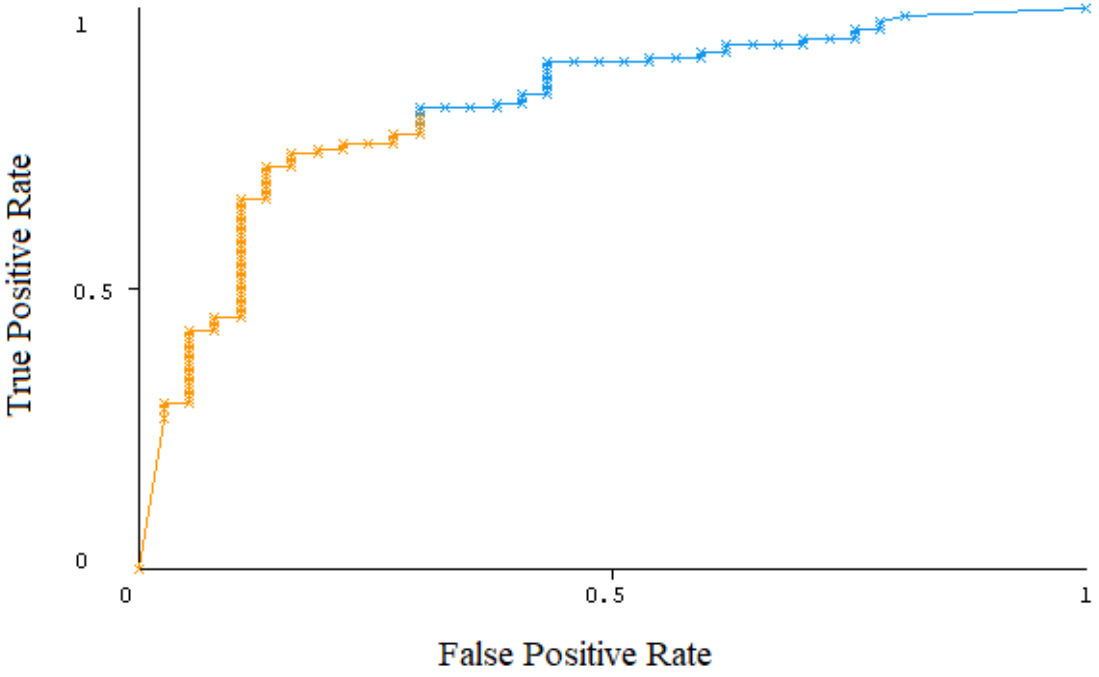


Figure (7): ROC curve for the model Multilayer Perceptron

Conclusion:

Machine learning algorithms play a significant role in predicting the risks of bank loans and decision support systems. The choice of the algorithm used to make the decision (whether the borrower will default), which is the key to addressing decision management when issuing a loan. In this paper, the performance of machine learning algorithms has been tested and their performance compared to standard measurements used on a dataset that includes 1000 loans and their repayment status. Finally, the results showed the possibility of using the proposed algorithms for this purpose with acceptable accuracy rates and superiority of the neural networks for this purpose.

REFERENCE

[1] Alsaleem, Maan & Hasoon, Safwan. (2020). COMPARISON OF DT& GBDT ALGORITHMS FOR PREDICTIVE MODELING OF CURRENCY EXCHANGE RATES. EUREKA: Physics and Engineering. 1. 56-61. 10.21303/2461-4262.2020.001132.

[2] Belcastro, L. & Marozzo, Fabrizio & Talia, Domenico & Trunfio, Paolo. (2016). Using Scalable Data Mining for Predicting Flight Delays. ACM Transactions on Intelligent Systems and Technology. 8. 10.1145/2888402.

[3] Chen, M. C., & Huang, S. H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. Expert Systems with Applications, 24, 433–441.

[4] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[5] Eletter, Shorouq & Yaseen, Saad. (2010). Applying Neural Networks for Loan Decisions in the Jordanian Commercial Banking System. 10.

[6] Hall et al.. The WEKA data mining software: an update. SIGKDD Explorations, 11(1). 2009.

[7] Huang Wei , Lai K. K. , Nakamori Y. and Wang Shouyang, (2004). Forecasting Foreign Exchange Rates with Artificial Neural Networks: A Review, International Journal of Information Technology & Decision Making Vol. 3, No.1.

[8] J. R. Quinlan. Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research, 4:77-90, 1996.

[9] Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Computational Statistics and Data Analysis, 50, 111

[10] Mattei, Pierre-Alexandre. (2019). A Parsimonious Tour of Bayesian Model Uncertainty.

[11] Salzberg, Steven. (1996). Book Review: "C4.5: Programs for Machine Learning" by J. Ross Quinlan.

[12] Stahl, F., & Jordanov, I. (2012). An overview of the use of neural networks for data mining tasks. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(3), 193–208. doi: 10.1002/widm.1052

[13] Supriya, P., Pavani, M., Saisushma, N., Vimala, N. and Vikas, K. (2019). Loan Prediction by using Machine Learning Models. International Journal of Engineering and Techniques, 5(22), pp.144-148.

[14] Wang, Di & Wu, Qi & Zhang, Wen. (2019). Neural Learning of Online Consumer Credit Risk.

Fraud prediction in bank loan administration using decision tree

IOP Conf. Series: Journal of Physics: Conf. Series **1299** (2019) 012037 doi:10.1088/1742-6596/1299/1/012037

Fraud prediction in bank loan administration using decision tree

I O Eweoya[1], A A Adebiyi[1,2], A A Azeta[1] and Angela E Azeta[3]

[1]Department of Computer and Information Sciences, Covenant University, Nigeria
[2]Department of Computer Science, Landmark University, Nigeria
[3]Department of PTTIM, FIIRO, Nigeria

ibukun.eweoya, ayo.adebiyi, ambrose.azeta{@covenantuniversity.edu.ng}, azetaangela@gmail.com

Abstract. The rate at which banks looses funds to loan beneficiaries due to loan default is alarming. This trend has led to the closure of many banks, potential beneficiaries deprived of access to loan; and many workers losing their jobs in the banks and other sectors. This work uses past loan records based on the employment of machine learning to predict fraud in bank loan administration and subsequently avoid loan default that manual scrutiny by a credit officer would not have discovered. However, such hidden patterns are revealed by machine learning. Statistical and conventional approaches in this direction are restricted in their accuracy capabilities. With a large volume and variety of data, credit history judgement by man is inefficient; case-based, analogy-based reasoning and statistical approaches have been employed but the 21st century fraudulent attempts cannot be discovered by these approaches, hence; the machine learning approach using the decision tree method to predict fraud and it delivers an accuracy of 75.9 percent.

Keywords: Confusion matrix, decision tree, fraud, machine learning, prediction.

1. Introduction

There are unsolved fraudulent practices in financial operations in the society, including bank credit administration, calling for a remedy through intelligent technology [1-4]. Existing fraud detection techniques in bank credit administration have not sufficiently met the desired accuracy, and avoidance of false alarm, and none focused on fraud in bank credit default. Also, fraudulent duplicates, missing data, and undefined fraud scenarios affect prediction accuracy [1-11].

Any unlawful act by human beings or invoked by machines that leads to personal gain at the expense of institutions or the legal human beneficiaries is a financial fraud, but an error must not be taken for a fraud [1],[12-14]. Considering the overall effect of financial frauds, it is referred to as an economic sabotage. The examples of financial fraud are money laundering, bank credit fraud, pension fraud, co-operative society fraud, tax evasion, telecommunications fraud, credit card fraud, inflated

contracts, financial reports fraud, health insurance fraud [15], automobile insurance fraud, and mortgage insurance fraud.

According to [16], there are many types of fraud including, credit card fraud, telecommunication fraud, computer intrusion, bankruptcy fraud, theft fraud or counterfeit fraud, and application fraud. The economy of nations do feel the impact of fraud and many approaches have been employed but with shortcomings. However, machine learning has proved to be more reliable. Machine learning uses data mining techniques to reveal hidden patterns in a large, volatile, and variety of data and make intelligent decisions through the revealed insights.

It is worthy of note that according to [17-19]; a high rate of default has been reported in different nations and this can be reduced using information technology. The rest of this paper is organized as follows: Section 2 is the materials and methods, followed by the results and discussion in Section 3, and section 4 is the conclusion of the paper.

The decision tree classifies data into discrete ones using tree structure algorithms [20-21]. It highlights the structural information contained in the data and classifies from root to the leaf node [22].The advantages of using decision trees include the fact that simplicity and speed of decision trees are second to none; there is no requirement for a domain knowledge or parameter setting; also, it comfortably handles high dimensional data where there are many attributes involved; the way it is represented allows for enhanced comprehensibility; it has a fantastic accuracy though this is dependent on the data in use; it supports incremental learning; they are unvaried, since they are used based on a single feature at each interval node. They work fine on both classification and regression problems; they can handle missing values; trees are plotted graphically, and can be easily interpreted; most interestingly, trees can be easily explained to people [23-25].

Credit default refers to the failure of a client to meet the legal obligations or conditions of a loan according to the promissory note. In other words, loan or credit default is the failure to repay a loan according to the terms agreed to initially before the approval of that loan. Non-performing loan refers to a specific amount of credit taken by a borrower but the debtor has declined in making agreed installment paybacks in 90 days for commercial banking loans and 180 days for consumer loans. Non-payment indicates neither the interest nor the principal gets paid with respect to that credit in 90 to 180 days depending on the type of loan, purpose or industry. Any definition of a non-performing loan is a function of the terms of that loan and the subsisting agreement as definition is not cast in stone but conditional based on promissory notes and agreements.

2. Materials and Methods

A credit dataset of 5000 instances and 9 attributes were employed for this research based on features extraction with the target attribute being the default status, and an effective data pre-processing before subsequent operations. The attributes include age, sex, income, employment status, the track of the last three payments (if any), and balance of loan taken. Python programming language was used for fraud prediction in credit or loan default using spyder 9.0. The classification was executed in Matlab 2017b [26] where cross validation and features extraction were employed. The training and testing was done in Matlab which gave a result of 75.9% accuracy. The scatter plot of the data is in Figure 1.Through the confusion matrix, the true positive rate, and false positive rate are as shown in Figure 2. Decision Tree Positive predictive values and false discovery rates are presented in Figure 3. Also, the ROC, is in Figure 4. Furthermore, 80.04 % was classified rightly, and 19.96% wrongly classified based on weka 3.8 [27] using stratified cross-validation

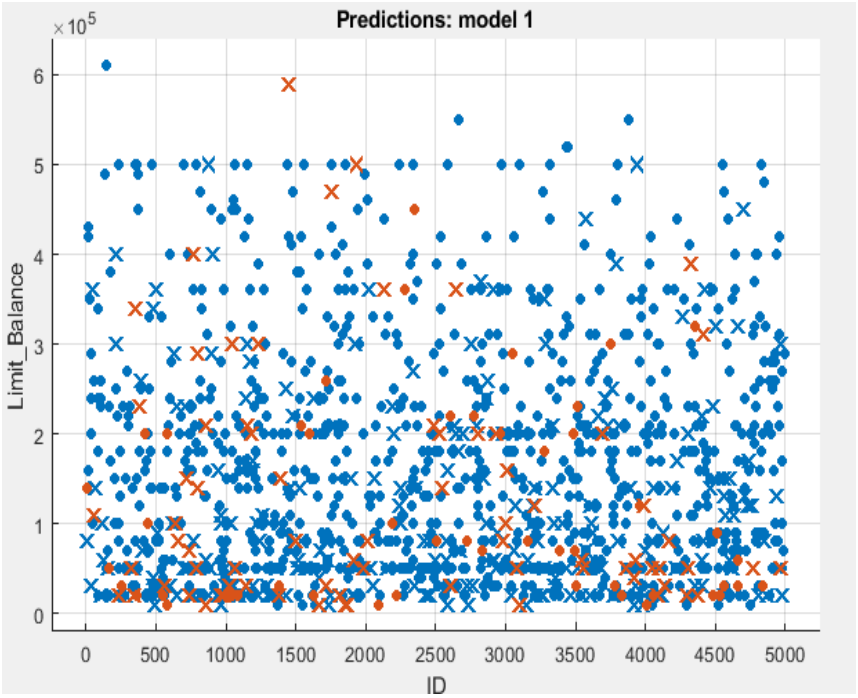


Figure 1: A scatter plot of the ID versus Limit balance

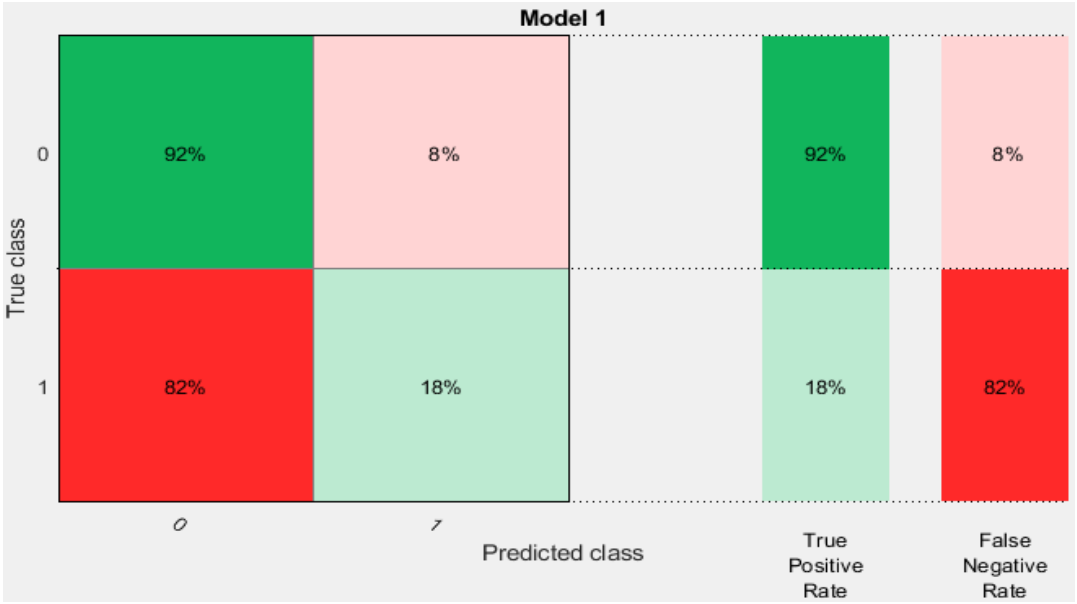


Figure 2: Decision Tree True positive rate and false negative rate

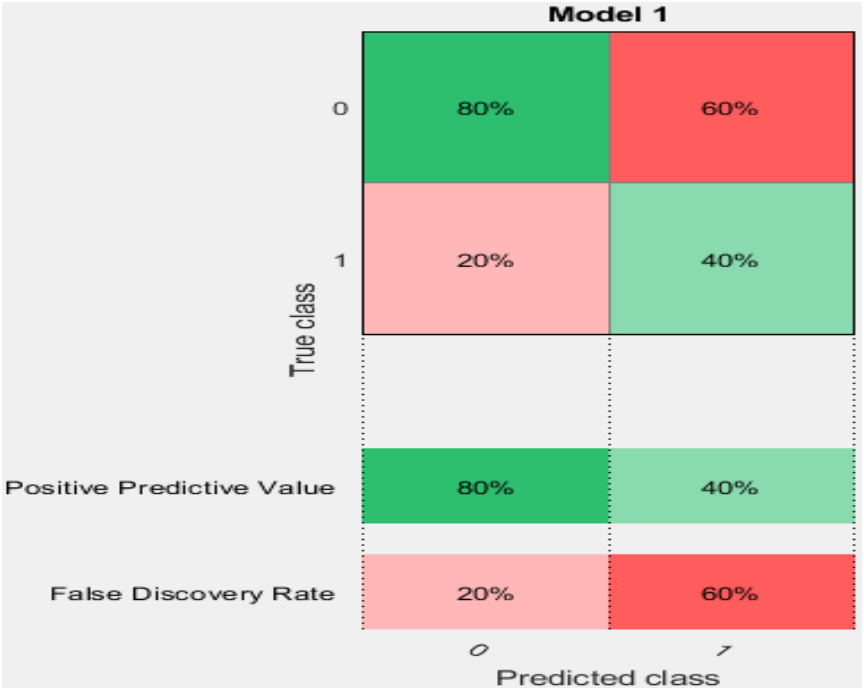


Figure 3: Decision Tree Positive predictive values and false discovery rates

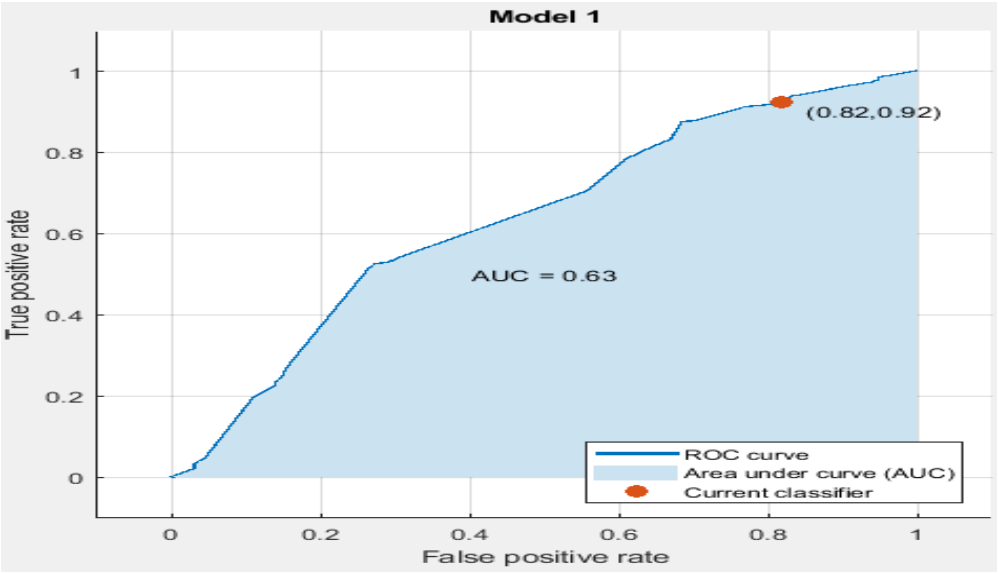


Figure 4: Decision Tree ROC curve

3. Results and Discussion

Credit or loan defaults have led to bank insolvency and nations entering recession, this has an untoward effect on people. For the purpose of extracting relevant features (features engineering), Principal Component Analysis was employed, cross validation was used to avoid overfitting in the model built; data splitting was done to separate testing data from training data. This allows the model to work on a fraction of the data not known before for the testing of the model. The training and testing yielded 75.9 % accuracy with a high true positive ratio. Also, 80.04 % of the instances were

correctly classified and 129 of the testing data identified to be fraudulent based on a python [28] program written for this work. As was presented in [29,30], this current study did not include hypothesis in its model formulation and testing, rather machine learning technique such as decision tree was engaged in the model formulation and prediction.

4. Conclusion

The anomaly of taking credit and ending up in a default to the detriment of the lender has been confirmed to have a remedy in machine learning. Using a real life dataset it has been revealed that false positives can be reduced with an employment of decision tree, thereby getting a highly reliable accuracy that financial institutions can depend on while scrutinizing loan applications.

5. Acknowledgement

In this study, we would like to express our deep appreciation for the support and sponsorship provided by Covenant University Centre for Research, Innovation and Discovery (CUCRID).

6. References

[1] Bagul P D Bojewar S and Sanghavi A 2016 Survey on hybrid approach for fraud detection in health insurance. *International Journal of Innovative Research in Computer and Communication Engineering,* **4**(4): 6918-*6922.*

[2] Hameed A A Karlik B and Salman M S 2016 Backpropagation algorithm with variable adaptive momentum *Knowledge-based Systems*,**114**: 79-87. DOI: https://doi.org/10.1016/j.knosys.2016.10.001

[3] Demla N and Aggarwal A 2016 Credit card fraud detection using svm and reduction of false alarms. *International Journal of Innovations in Engineering and Technology*, **7(**2): 176-182.

[4] Fahmi M Hamdy A and Nagati K 2016 Data mining techniques for credit card fraud detection: empirical study. *Sustainable Vital Technologies in Engineering and Informatics*, pp.1-9, Elsevier.

[5] Vaishali V 2014 Fraud detection in credit card by clustering approach. *International Journal of Computer Applications,* **98(**3)*: 29-32.*

[6] Abid L Masmoudi A and Zouari-Ghorbel S 2016 The consumer loan's payment default predictive model: an application in a Tunisian commercial bank. *Asian Economic and Financial Review,* **6**(1): 27-42.

[7] Sharma S and Choudhury A R 2016 Fraud analytics: A survey on bank fraud and fraud prediction using unsupervised learning based approach. *International Journal of Innovations in Engineering Research and Technology*, **3(**3): 1-9.

[8] Agaskar V Babariya M Chandran S and Giri N 2017 Unsupervised learning for credit card fraud detection. *International Research Journal of Engineering and Technology* (IRJET), **4**(3): 2343-2346.

[9] Rawate K R and Tijare P A 2017 Review on prediction system for bank loan credibility. *International Journal of Advance Engineering and Research Development,* **4(**12): 860-867.

[10] Rimiru R Wa S W and Otienoc C 2017 A hybrid machine learning approach for credit scoring using PCA and logistic regression. *International Journal of Computer,* **27(**1): 84-102.

[11] Boateng E Y and Oduro F T 2018 Predicting microfinance credit default: A study of Nsoatreman rural bank, Ghana. *Journal of Advances in Mathematics and Computer Science (JAMCS),* **26(**1): 1-9.

[12] Rawte V and Anuradha G 2015 Fraud detection in health insurance using data mining techniques. *International Conference on Communication, Information & Computing Technology (ICCICT),* pp. 1-5, Mumbai

[13] Naik J and Laximinarayana J A 2017 Designing hybrid model for fraud detection in insurance. *IOSR Journal of Computer Engineering*, **1**: 24-30.

[14] Akomolafe J A Eluyela D F Ilogho S O Egharevba J W and Aina O 2017 Financial crime in Nigeria public sector: A study of Lagos state ministries. *International Journal of Innovative Research in Social Sciences & Strategic Management Techniques*, **4** (1):13-21.
[15] Kose I Gokturk M and Kilic K 2015 An interactive machine learning-based electronic fraud and abuse detection system in healthcare insurance, *Applied Soft Computing*, **36**:283–299.
[16] Tripathi K K and Pavaskar M A 2012 Survey on credit card fraud detection methods. *International Journal of Emerging Technology and Advanced Engineering,* **2(**11): 721-726.
[17] Central Bank of Nigeria CBN 2016 Financial stability report - December 2016. Available from: https://www.cbn.gov.ng/out/2017/fprd/fsr%20december%202016%20(2).pdf. Retrieved March, 2018.
[18] Central Bank of Nigeria CBN 2017 Financial stability report - June 2017. Available from: https://www.cbn.gov.ng/Out/2018/FPRD/FSR%20June%202017.pdf. Retrieved March, 2018.
[19] World Bank 2018 Economic indicators for over 200 countries, https://www.theglobaleconomy.com/Nigeria/Nonperforming_loans/ Retrieved March, 2018.
[20] Quinlen J R 1986 Introduction of decision trees. *Machine Learning* **1**: pp. 81-106.
[21] Han J and Kamber M 2011 Data mining concepts and techniques. *Elsevier*, p. 744, Morgan Kaufmann.
[22] Kotsiantis S B 2007 Supervised machine learning: A review of classification techniques. *Informatica,* **31**: 249-268.
[23] Williams G J and Huang Z 1997 Mining the knowledge mine: The hot spots methodology for mining large real world databases. *Australian Joint Conference on Artificial Intelligence*, pp. 340-348.
[24] Liou F M Tang Y C and Chen J Y 2008 Detecting hospital fraud and claim abuse through diabetic outpatient services. *Health Care Management Science, **11***(4): 353-358.Available from: http://dx.doi.org/10.1007/s10729-008-9054-y
[25] Shin H Park H Lee J and Jhee W C 2012 A scoring model to detect abusive billing patterns in health insurance claims. *Expert Systems with Applications,* **39(**8), 7441-7450 Available from: http://dx.doi.org/10.1016/j.eswa.2012.01.105
[26] Matlab and Statistics Toolbox Release 2017b The MathWorks Inc Natick Massachusetts United States.
[27] Hall M Frank E Holmes G Pfahringer B Reutemann P and Witten H I 2009 The WEKA data mining software: An update. SIGKDD explorations, Volume 11, Issue 1.
[28]Python Software Foundation. Python language reference, version 2.7. Available at http://www.python.org
[29] Nicholas-Omoregbe O S Azeta A A Chiazor I A and Omoregbe N 2017 Predicting the adoption of e-learning management system: A case of selected private universities in Nigeria. *Turkish Online Journal of Distance Education-TOJDE **18(2)*** 106-121.
[30] Azeta A A Misra S Azeta V I Osamor V C 2019 Determining suitability of speech-enabled examination result management system. *Wireless Networks* 1-8.