CREDIT LESS LOAN APPROVAL PREDICTOR

Dr. Gaurav Singal
Department of Computer Science and Engineering
Netaji Subhas University of Technology
New Delhi, India

Chitra Agrawal
Department of Computer Science and Engineering (AI)
Netaji Subhas University of Technology
New Delhi, India

Kanika Tiwari
Department of Computer Science and Engineering (AI)
Netaji Subhas University of Technology
New Delhi, India

***Abstract*—There are alternative methods for loan approval that move beyond the conventional reliance on credit scores and job stability, which often exclude self-employed individuals and newcomers to the credit system. By using machine learning techniques, our study explores the integration of non-traditional financial data, such as transaction histories, spending patterns and income stability, to assess financial reliability. We aim to develop a robust and inclusive model for predicting loan approval eligibility, addressing gaps in traditional methods that fail to recognize the creditworthiness of individuals without established credit histories The study looked at how different factors, such as irregular income and financial habits, influence the model's predictions. We provide insight into the efficacy of alternative financial metrics in improving loan accessibility through data analysis and model validation. The research is expected to include a data-driven approach for loan approval that can be implemented to serve diverse financial profiles, contributing to financial inclusion and economic well-being. Our findings could have significant implications for the credit industry.**

***Index Terms*—Machine learning, Logistic Regression, Random forest, Naive bayes, Loan prediction**

I. Introduction

Access to credit is critical to financial security and growth, but many people struggle to get credit because they don't have a traditional credit history or stable employment. Banks and financial institutions rely on credit scores and employment records to assess credit eligibility; this works for those who are employed, but not for freelancers, the self-employed, and new immigrants. Non-traditional earners are often rejected not because they can't repay the loan, but because their finances don't meet the old standards. This highlights the need for a more integrated and flexible way to measure financial credibility.

This work is motivated by solving these problems using other data such as fixed income, spending patterns, and utility payment history to evaluate money more honestly and effectively. Using machine learning, our goal is to create a predictive tool for getting loans for people who are excluded from the traditional lending process. Following these are the contribution for the loan approval predictor-

- Inclusive credit assessment model by machine learning.
- It will integrate non-traditional data points such as spending behavior and payment history.
- Algorithm evaluation-aiming at providing the best predictive models.
- A system transparent and comprehensible to populations financially exploited.

This paper begins with the domain, problem, motivation, and contributions. Then follow methodology that described the data and models used. Finally, results and discussions offer insight into the model's performance, and the paper concludes with future directions.

II. Literature/Related work

A. Background Study

Since the global financial crisis, risk management in banks Bank has become increasingly important with guidance To decide. Sanction the loans to potential The individuals play a leading role in risk management. Yet, As machine learning algorithms are essentially black boxes, Lots of lenders vary in their actions regarding some factors. general and systematic literature reviews focused on the application of powerful machine learning in banking Risk management.

The authors put forward a machine learning-based loan prediction model. the Decision Tree and Random Forest algorithms. The The main aim is to determine the landscape, validity, and Background of the client applying for multiple loans. The technique investigative data analysis is used to transaction In addressing the problem of either approving or rejecting the loan application. -or the loan prediction. From this, it centers its focus mostly on In deciding, regardless of the loan extended to an excellent individual. Whether permitted or not, the person or organization concerned shall be notified.[1]

The authors apply logistic regression to the loan. loan approval prediction with a set of samples applications. [2]. The authors ran a machine Learning approach to predict credit recovery. Credit recovery The next important risk for the banking industry in question is predicting. is rather challenging. Many machine learning techniques were It was used to predict credit retrieval, and the gradient expansion algorithms (GBM) outperformed the other machine Learning techniques. [3]. The authors proposed an approach that automatically collects data for a candidate and calculates His credit score. To collect the information of a user, this model uses the social media. [4].

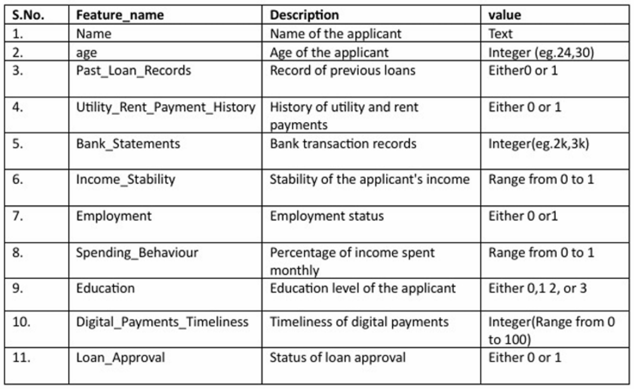
The primary goal is to determine whether the advance of the loan to a particular person is safe or not. The perpetrators offer To reduce the risk factor, various machine learning models are used. is allied with identifying a trusted person and saving much banking time and effort[5]. The An author explored an ML algorithm for loan sanctioning. Process prediction. The objective is to minimize the possibility of discovery A good character to repay the loan on time, so that the Bank can keep its nonperforming assets on hold. This can be Achieved by providing the bank's previous accounts of Customers who have bought loans are also very trained. The exact result can be produced by making machine learning models. [6]

III. Methodology

A. *Approach/Technique/Model/Improvement*

Dataset Description

This data was synthetically generated using the Faker within this project.Library to simulate a set of records for people applying for loans. The dataset Comprising ten features or attributes for each applicant, it encompasses different facets of their personal and financial backgrounds. To achieve the objective of forecasting, a dataset is compiled. The chances of getting an approval for a loan depend on these characteristics. The target variable loan approval depends on a set of financial indicators. This database allows applying learning machine algorithms. This will help predict loan approval based on these attributes to simulate a simplified Credit scoring model.



S.No.	Feature_name	Description	value
1.	Name	Name of the applicant	Text
2.	age	Age of the applicant	Integer (eg.24,30)
3.	Past_Loan_Records	Record of previous loans	Either0 or 1
4.	Utility_Rent_Payment_History	History of utility and rent payments	Either 0 or 1
5.	Bank_Statements	Bank transaction records	Integer(eg.2k,3k)
6.	Income_Stability	Stability of the applicant's income	Range from 0 to 1
7.	Employment	Employment status	Either 0 or1
8.	Spending_Behaviour	Percentage of income spent monthly	Range from 0 to 1
9.	Education	Education level of the applicant	Either 0,1 2, or 3
10.	Digital_Payments_Timeliness	Timeliness of digital payments	Integer(Range from 0 to 100)
11.	Loan_Approval	Status of loan approval	Either 0 or 1

Fig. 1. Dataset Description.

Data Pre-processing

In data preprocessing, first missing values with 30Standard methodology is utilized to remove and impute records. missing methods. Fig. 2 represents the missing fraction. number of data. All the variables undergo exploratory data analysis. end will be carried out. The EDA helps to unearth the latent where the Sub-Grade plot shows that with increasing grade A growing loan defaulters ratio further adds to distress. The Home-Owner ship plot identifies that more home owners Rental tenants are usually prone to loan defaulters. Consequently, one may identify The categorical data has been transformed into a format that can be processed utilizing machine learning techniques.

Feature Selection

Feature Selection focuses on pinpointing the most significant features that enhance the model's predictive ability, while eliminating irrelevant or redundant features to boost performance and minimize overfitting.

Recursive Feature Elimination (RFE): This method systematically removes less important features, ranking them according to their influence on the model's predictions. It is particularly beneficial for datasets with numerous features, aiding in the identification of the main predictors.

Principal Component Analysis (PCA): PCA works by reducing the data's dimensionality, converting correlated features into uncorrelated principal components. This process helps to eliminate multicollinearity and enhances model performance.

Correlation Analysis: This technique identifies and discards highly correlated features that offer redundant information, ensuring the model does not overly focus on any single feature, thereby improving its generalization.

Model selection

Choosing the right machine learning models is essential for achieving the best performance. This research focuses on few algorithms: Random forest, logistic regression, kNN and Naive Bayes. Each of these algorithms has unique benefits for predicting loan approval without relying on traditional credit scores.

Logistic Regression: This method is favored for its straightforwardness and effectiveness in binary classification tasks. It establishes a connection between features and the likelihood of an outcome, making it particularly useful for predicting loan approvals based on various criteria.

K-Nearest Neighbors (KNN): This is a non-parametric approach that classifies data by looking at the majority class among the nearest neighbors. It is particularly effective for complex decision boundaries and performs well when the decision regions are not linearly separable.

Random Forest: This ensemble technique merges multiple decision trees to enhance prediction accuracy. It mitigates overfitting by averaging the predictions and is resilient to noisy data, making it well-suited for managing large and intricate datasets, such as those used for predicting loan approvals.

Naive Bayes is a probabilistic model that is effective with concepts that are continuously changing, as well as ones that are already classified and are irreversible.

B. *Algorithm and Flow Chart*

Logistic Regression

Logistic Regression is one of the most commonly used algorithms for binary classification.Works by mimicking the probability of a binary outcome, that is, loan approval or rejection as the linear combination of the features at input. The logistic function This linear output translates to a value between 0 and 1, so it can be interpreted as a probability.It

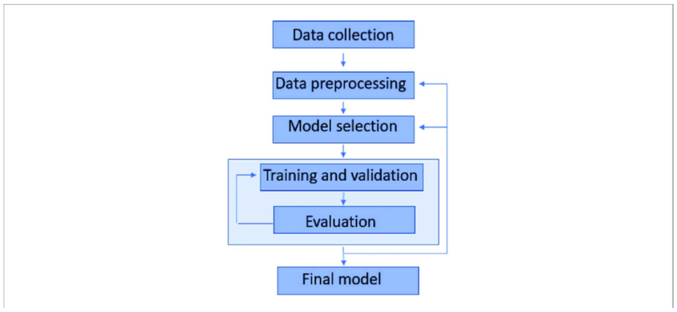


Fig. 2. Flowchart for Building ML model

provides insight into the impact of individual predictors, among which is income. Stability and utility payment history, on loan approval probability. This inter Predictability is critical to credit less loan situations where financial institutions Approval decisions also require to be justified. Its coefficients were decomposed Be familiar with which factors are most associated with getting a loan approved.

K-Nearest Neighbors (KNN)

In credit less loan approval, KNN is a predictor that provides a f.This approach captures the complexities and non-linear relationships in the data flexibly by classifying applicants according to similar previous ones. By setting the k parameter and standardization of features such as Bank Statements and Income Stability, KNN Well used local patterns, gathering similar financial behaviors (such as income Stability and spending patterns that can influence loan approval predictions. Assessed on accuracy, precision, recall, and F1-score, KNN performed very strongly. But it requires a delicate balance between model complexity and computational feasibility. KNN is very resource-intensive with large datasets. In general, the adaptability of KNN This makes it a useful tool in understanding subtle applicant behavior; however, Practical considerations on the computational intensity involved in Proper use. The graph Figure Fig.2 represents the accuracy of a k-nearest neighbors model as a Function of the number of neighbors. From the plot, the highest accuracy It seems, the accuracy is around at k=2 and k=13 is 0.76.

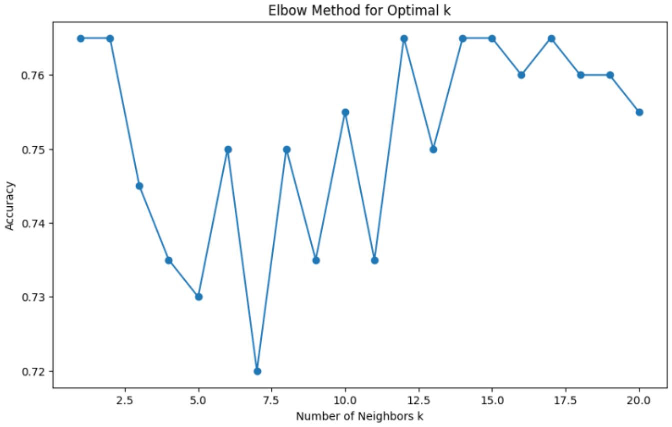


Fig. 3. Variation in accuracy with K neighbours

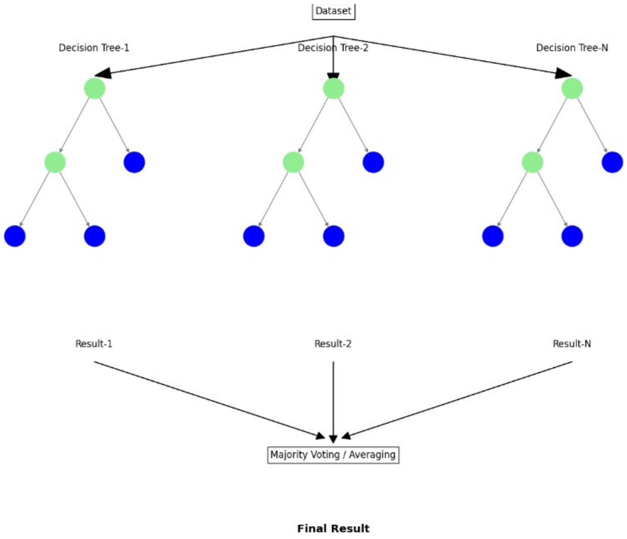


Fig. 4. Random forest flowchart

Naive Bayes

Naive Bayes is indeed a very simple probabilistic classifier assuming features are independent - and by 'independent, I mean there that each feature contributes independently to the target variable's probability. Very much in contradiction with this really very strong assumption, very effective it remains for many practical aims mainly for amounts of large data, even for high required computing efficiency. Naive Bayes provides a fast and more interpretable solution for loan approval prediction when credit history is not available. Being based on conditional probabilities of some features such as age, employment status, and how timely digital payments are settled, it calculates the conditional probability of loan approval for one class and another, presenting a simple method for classification: easy to interpret. Though not as successful in the capturing of tightly complex relationships among features as with SVC, Naive Bayes is pretty successful where features are only somewhat independent or speed and simplicity are important.

Random forest

Random Forest can treat the non-linear relationships, does a good job with biased data, and provides for easy interpretability using feature importance. Accuracy and loss testing were performed using the loan prediction datasets. The RF method achieved 100precision and 21Loss,and 90precision.

C. Mathematical Formulation

Random Forest is one of the strong algorithms in machine learning with ensemble techniques for classification and regression. It is known to handle issues of high dimensionality and overfitting and thus achieves accurate prediction based on ensemble techniques. Below follows a brief presentation of its theoretical background along with necessary formulas:.

- $E(S)$ – represents the entropy function
- $IG(S,A)$ – represents the Information Gain
- $Gini(S)$ – represents the Gini index
- $p(\text{Yes})$ – represents the proportion of positive terms
- $p(\text{No})$ – represents the proportion of negative terms

$S, A, S_v, |S|, |S_v|$ represent the current set of data, candidate feature to split on, subset of S, sizes of S and S_v respectively.

$$E(S) = -p(\text{Yes}) \log_2 p(\text{Yes}) - p(\text{No}) \log_2 p(\text{No}) \tag{1}$$

$$IG(S,A) = E(S) - \sum_{v \in \{0,1\}} \left(\frac{|S_v|}{|S|} E(S_v) \right) \tag{2}$$

$$Gini(S) = 1 - \sum_{i=1}^{k} p(i)^2 \tag{3}$$

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{4}$$

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \tag{5}$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \tag{6}$$

$$F1\text{-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Naive Bayes is based on Bayes' Theorem, which gives the posterior probability of a class given a set of features:

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

Where:

- $P(Y|X)$: Posterior probability of class Y given features X
- $P(X|Y)$: Likelihood of the features given the class
- $P(Y)$: Prior probability of class Y
- $P(X)$: Evidence, the total probability of the features

Euclidean Distance In k-NN, the distance between two points $x = (x_1, x_2, \ldots, x_n)$ and $x' = (x'_1, x'_2, \ldots, x'_n)$ in an n-dimensional space is commonly measured using the Euclidean distance:

$$d(x, x') = \sqrt{\sum_{i=1}^{n} (x_i - x'_i)^2}$$

Where: - x_i and x'_i are the i-th features of points x and x', respectively.

Logistic Regression is essentially a probabilistic model with primarily usage in the realm of binary classification. Basically, it models the probability assigned by giving an input to a given class. Using logistic, or sigmoid function transforms a linear combination of features on inputs to take the values between 0 and 1 so as to interpret as a probability.

1. Sigmoid Function

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Where:

- $h_\theta(x)$: Predicted probability
- $\theta^T x$: Linear combination of input features and their weights

2. Decision Boundary

$$\text{Predict 1 if } h_\theta(x) \geq 0.5, \text{ otherwise 0}$$

3. Cost Function (Log Loss)

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right]$$

Where:

- m: Number of training examples
- $y^{(i)}$: True label for the i-th example
- $h_\theta(x^{(i)})$: Predicted probability for the i-th example

D. Hardware/Software Setup

State-of-the-art hardware is the necessary pre-requisite in the development mechanism of machine learning (ML) tasks and it's main function is to have fluid data processing, model training, and validation. Such major hardware devices as the Central Processing Unit (CPU), Graphics Processing Unit (GPU), Random Access Memory (RAM) and storage remain the most indispensable. Each of them together handles the high amount of data related to training the ML models and the algorithms' running with efficiency.

- For basic machine learning tasks such as preparing data and checking models, multi-core CPUs like the Intel Core i7 or AMD Ryzen 7 are sufficient. Faster processors shorten model training times.
- For deep learning and large scale neural networks, GPUs are ideal due to their parallel processing capabilities. They significantly accelerate model training and optimization.
- To work smoothly with big datasets, 16GB to 32GB of RAM is required. It prevents crashes during real-time activities.
- SSDs with a capacity of at least 1TB are preferable for storing complex datasets and ensuring rapid data loadings necessary for ML tasks.

The Creditless Loan Approval Predictor works on Windows 10/11, Ubuntu 20.04 LTS, or macOS Monterey. The software uses Python 3.8+ with NumPy, Pandas, Scikit-Learn, TensorFlow/Keras, and PyTorch. Jupyter Notebook, VSCode, or PyCharm is used for coding. Data processing is done using Apache Spark and Dask. Data is stored in MySQL, PostgreSQL, MongoDB, and AWS RDS. Flask/FastAPI, Docker, and Kubernetes handle deployment, while MLflow, TensorBoard, and Prometheus/Grafana perform monitoring. Git and GitHub/GitLab manage version control, ensuring effective machine learning model development.

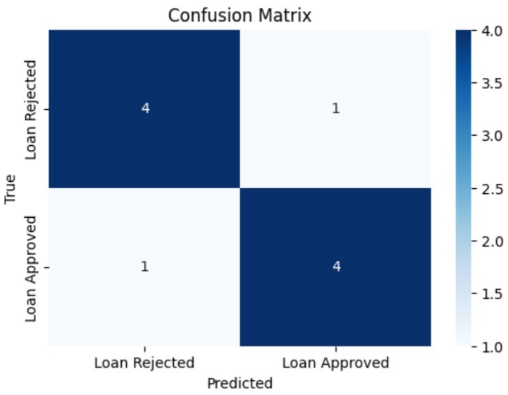


Fig. 5. Confusion metrics

IV. Result Analysis

A. Experiment setup

We are making a computer model to foretell loan approval based on person and financial facts. The facts we have for this are who applied, financial details, and whether the loan was permitted. In preparing the knowledge for the model, we need to fill in any empty spots, code names, make the numbers consistent, and select the very best features. After that, we teach Logistic Regression, Naive Bayes, Random Forest and KNN on 70% of our data and if they work well. The results we get will be checked through accuracy, precision, recall, F1-score. We will use the model with the best results to make it simpler for banks to say 'yes' or 'no' to your loan application.

B. Evaluation metrics

Evaluation of the model is a method used in the assessment of the model's performance by setting up some constraints among which we should note while evaluating the model that it can't underfoot or overfit the model. Several ways are prescribed to measure the performance of the model such as Confusion metrics, Accuracy, Precision, Recall, F1 score etc.

1) Confusion Metrics

2) Accuracy: Predefined metrics were used to measure the accuracy of the model with accuracy. A model that is balanced shows very high accuracy, whereas in an unbalanced class, the accuracy is very low.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{4}$$

3) Precision: Positive instance positivity ratio is the division of the number of cases when the predicted value is true positive by the discreetly planned estimated healthy individuals TH. In the equation provided, the denominator denotes the positive predictions of the model relating to the entire input dataset. Precision describes the rigor of our model. Our data set obtained precision value of 0.72, which is the good precision value.

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \tag{5}$$

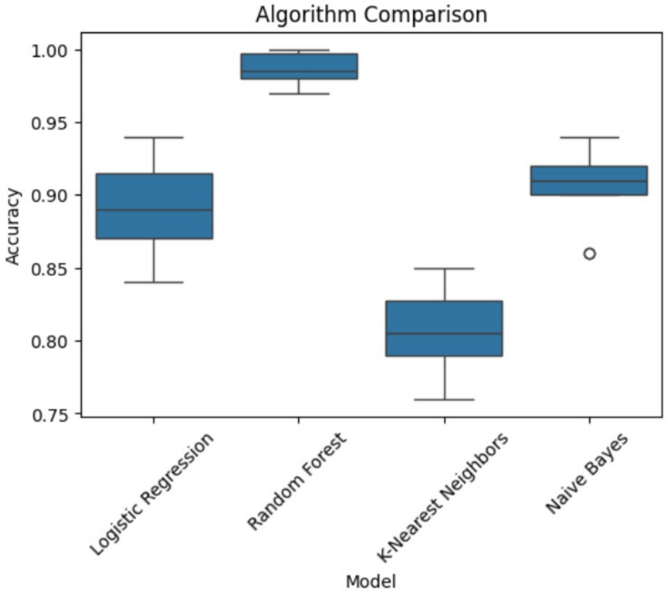


Fig. 6. Algorithm Comparison

4) Recall: Recall value is the proportion of genuine positive instances in the whole dataset that are actually positive. This value is represented as the numerator of the recall value. The denominator represents the total number of instances of which there are positive relevance in the whole data. Hence it is said, 'how much extra right ones, the model will fail if it shows maximum right ones'.

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \tag{6}$$

5) F1 Score: The F1 Score is defined as the harmonic mean of precision and recall. The best model is the one with the highest F1 score. So the values of the numerator are the one product of precision and recall and such condition does not meet, i.e. for example the final F1 score goes down significantly. So a model does well in F1 score if the positive predicted (precision) having positive value and doesn't miss out on positives and predicts them negative (recall).

$$F1\text{-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

C. Comparison/ Result graphs

Random forest achieves high accuracy (0.895) with a good balance between precision (0.911765) and recall (0.632653), but its auc-roc (0.300810) suggests room for improvement in class separation. Random forest performs slightly better in accuracy (0.910) and perfect precision (1.0), though its recall (0.632653) and auc-roc (0.205826) indicate struggles with class differentiation. Naive bayes, despite perfect recall (1.0), shows poor overall performance with low accuracy (0.250), precision (0.246231), and f1-score (0.395161), due to many false positives. K-nearest neighbors (best k) has moderate accuracy (0.765) but low precision (0.520833) and recall (0.510204), leading to imbalanced performance, and its

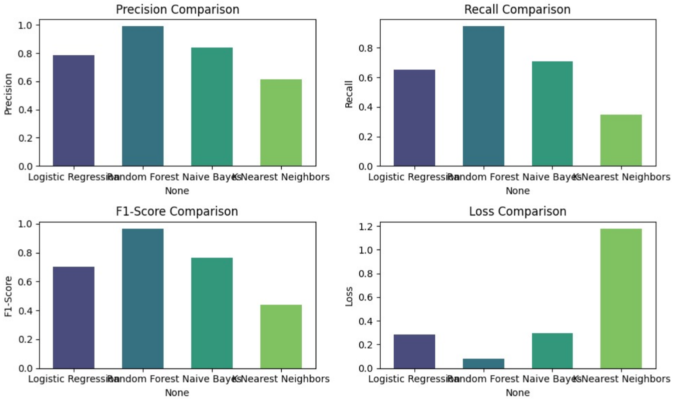
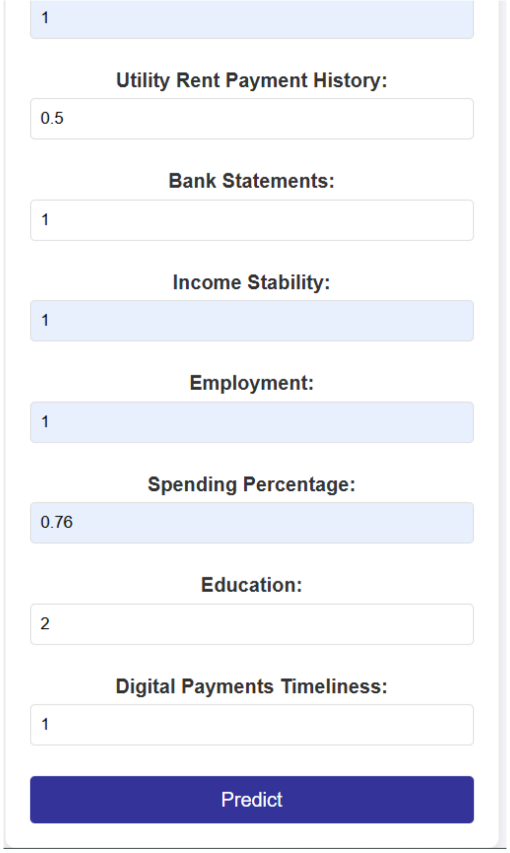


Fig. 7. Confusion metrics comparison



1

Utility Rent Payment History:

0.5

Bank Statements:

1

Income Stability:

1

Employment:

1

Spending Percentage:

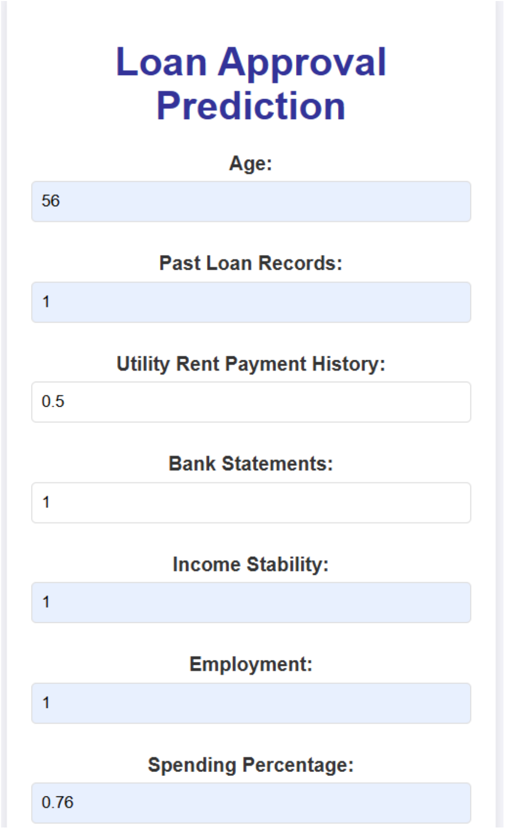
0.76

Education:

2

Digital Payments Timeliness:

1



Loan Approval Prediction

Age:

56

Past Loan Records:

1

Utility Rent Payment History:

0.5

Bank Statements:

1

Income Stability:

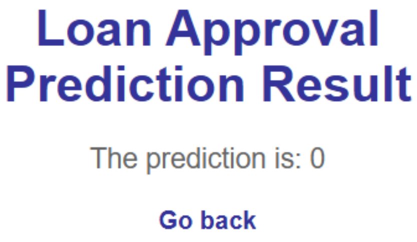
1

Employment:

1

Spending Percentage:

0.76



Loan Approval Prediction Result

The prediction is: 0

auc-roc (8.470259) suggests difficulty in distinguishing classes effectively.

D. observation inferences

The observations reveal that Logistic Regression performs well in terms of accuracy and balance between precision and recall, but its low AUC-ROC score suggests it struggles with class separation, indicating a need for improvement, especially in handling imbalanced data. Random Forest demonstrates high accuracy and perfect precision, but its low recall and AUC-ROC imply overfitting and difficulty in distinguishing between classes, requiring further optimization. Naive Bayes achieves perfect recall but suffers from poor accuracy, precision, and F1-score due to a high number of false positives, making it unsuitable for this task. K-Nearest Neighbors shows moderate accuracy but low precision and recall, along with a low AUC-ROC, highlighting its challenges in distinguishing classes effectively. In conclusion, while Random Forest shows the most promise, it requires optimization to improve recall and AUC-ROC for better performance in predicting loan approval.

V. Conclusion and Future Work

From the model evaluations, Random Forest appears to be the most effective model for the credit loan approval prediction task, providing high accuracy and perfect precision. However, improvements can be made to the recall and AUC-ROC scores.To conclude the model has shown an maximum optimal accuracy of nearly 89% for the random forest algorithm which will work even better on larger datasets(Logistic Regression is also a strong contender with good balance between precision and recall. Naive Bayes and KNN showed limited effectiveness in predicting credit loan approvals, particularly due to low precision and recall values. Further model tuning and enhancements are necessary to improve performance, especially in handling class imbalances and optimizing recall and AUC scores.

VI. References

[1]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, 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.

[2]Yuelin Wang, Yihan Zhang, Yan Lu, Xinran Yu, A Comparative Assessment of Credit Risk Model Based on Machine Learning ——a case study of bank loan data, Procedia Computer Science, Volume 174, 2020, Pages 141-149, ISSN 1877-0509

[3]Sarkar T, Rakhra M, Sharma V, Singh A. An Empirical Comparison of Machine Learning Techniques for Bank Loan Approval Prediction. In2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE) 2024 May 9 (pp. 137-143). IEEE.

[4]Sarkar, Tiyas, Manik Rakhra, Vikrant Sharma, and Amanpreet Singh. ”An Empirical Comparison of Machine Learning Techniques for Bank Loan Approval Prediction.” In 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), pp. 137-143. IEEE, 2024.

[5]Tumuluru, P., Burra, L.R., Loukya, M., Bhavana, S., CSaiBaba, H.M.H. and Sunanda, N., 2022, February. Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms. In 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS) (pp. 349-353). IEEE.

[6]Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning——a case study of bank loan data. Procedia Computer Science, 174, 141-149.

[7]Singh, V., Yadav, A., Awasthi, R., & Partheeban, G. N. (2021, June). Prediction of modernized loan approval system based on machine learning approach. In 2021 international conference on intelligent technologies (CONIT) (pp. 1-4). IEEE.

[8] Singh, V., Yadav, A., Awasthi, R. and Partheeban, G.N., 2021, June. Prediction of modernized loan approval system based on machine learning approach. In 2021 international conference on intelligent technologies (CONIT) (pp. 1-4). IEEE.

[9]Uddin, N., Ahamed, M. K. U., Uddin, M. A., Islam, M. M., Talukder, M. A., & Aryal, S. (2023). An ensemble machine learning based bank loan approval predictions system with a smart application. International Journal of Cognitive Computing in Engineering, 4, 327-339.

[10]Uddin N, Ahamed MK, Uddin MA, Islam MM, Talukder MA, Aryal S. An ensemble machine learning based bank loan approval predictions system with a smart application. International Journal of Cognitive Computing in Engineering. 2023 Jun 1;4:327-39.

[11]Sheikh, Mohammad Ahmad, Amit Kumar Goel, and Tapas Kumar. ”An approach for prediction of loan approval using machine learning algorithm.” In 2020 international conference on electronics and sustainable communication systems (ICESC), pp. 490-494. IEEE, 2020.

[12]Sheikh, Mohammad Ahmad, Amit Kumar Goel, and Tapas Kumar. ”An approach for prediction of loan approval using machine learning algorithm.” 2020 international conference on electronics and sustainable communication systems (ICESC). IEEE, 2020.

[13]Wang Y, Zhang Y, Lu Y, Yu X. A Comparative Assessment of Credit Risk Model Based on Machine Learning——a case study of bank loan data. Procedia Computer Science. 2020 Jan 1;174:141-9.