Loan Analysis and Prediction

Literature Survey

By:

Muhammad Nihaal Ur Rahmaan	1NT21CS109
Lovia E B	1NT21CS095
Mukesh Raj Sah	1NT21CS110
Karthik	1NT21CS084

1

Sameerunnisa SK, Sai Rama Harsha M.N.V, Tarun Teja K, Swathi K.V, Manikanta K.T.V (March 2024) Machine Learning Based Classification Model for Prediction of Bank Loan Approval IJRASET59231, Volume 12 Issue III Mar 2024, ISSN: 2321-9653 available: https://doi.org/10.22214/ijraset.2024.59231

The Gradient Boosting Classifier, an ensemble learning method, is a technique that brings into play various prediction models to build a robust statistical estimator sequentially joining forecasts of numerous decision trees referred to as weak learners. This structure was custom-fitted to our loan approval prediction task, wherein they employed specific numerical terms and computations.

Base Learners (Decision Trees): Decision trees were the most common building blocks utilized within the Gradient Boosting framework.

In the end, hyperparameter tuning is done and then we take our model's performance for a test spin across a several evaluation metrics including but not limited to the ability of the model to distinguish between defaulted and non-defaulted loans

They got a formidable precision of 98.03% applying the Gradient Boosting Classifier with 25 trees for a decision. This excellent precision was realized through model training implementation of diverse learning algorithms, accurate hyperparameter tuning which is a procedure of optimizing machine learning algorithms

Classification Report:

	precision	recall	f1-score	support
Approved	0.98	0.99	0.98	678
Rejected	0.98	0.97	0.97	390
accuracy			0.98	1068
macro avg	0.98	0.98	0.98	1068
weighted avg	0.98	0.98	0.98	1068

Confusion Matrix:

```
[[670   8]
 [ 13 377]]
```

Fig.1. Accuracy for our project

2

v, Viswanatha & Ac, Ramachandra. (2023).
Prediction of Loan Approval in Banks using Machine Learning Approach.
International Journal of Engineering and Management Research.
13. 7-19. 10.31033/ijemr.13.4.2.

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 KNearest Neighbors Classifier, the Decision Tree Classifier, the Random Forest Classifier, and the Gaussian Naive Bayes Classifier.

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%.

3

Mr. V. Sravan Kiran, B. Teja Reddy, D. Uday Kumar, K. Sai Avinash Varma, T. Sheshi Kiran (Aug 2023)
Loan Eligibility Prediction Using Machine Learning
IJRASET55132 | Volume 11 Issue VIII | ISSN: 2321-9653
https://doi.org/10.22214/ijraset.2023.55132

Model	Accuracy
Logistic Regression	classification report of training data precision recall f1-score support 0 0.95 0.45 0.61 154 1 0.80 0.99 0.88 337 accuracy 0.82 491 macro avg 0.87 0.72 0.75 491 weighted avg 0.84 0.82 0.80 491 classification report of testing data precision recall f1-score support 0 0.83 0.39 0.54 38 1 0.78 0.96 0.86 85 accuracy 0.79 123 macro avg 0.81 0.68 0.70 123 weighted avg 0.80 0.79 0.76 123
Support Vector Machine	cost of training model precision recall f1-score support 0 0.96 0.47 0.63 154 1 0.80 0.99 0.89 337 accuracy 0.83 491 macro avg 0.88 0.73 0.76 491 weighted avg 0.85 0.83 0.81 491 cost of testing model precision recall f1-score support 0 0.79 0.39 0.53 38 1 0.78 0.95 0.86 85 accuracy 0.78 123 macro avg 0.78 0.67 0.69 123 weighted avg 0.78 0.78 0.75 123
Random Forest Classifier	cost of training model precision recall f1-score support 0 0.99 0.90 0.94 154 1 0.95 1.00 0.98 337 accuracy 0.97 491 macro avg 0.97 0.95 0.96 491 weighted avg 0.97 0.97 0.96 491 cost of testing model precision recall f1-score support 0 0.68 0.39 0.50 38 1 0.77 0.92 0.84 85 accuracy 0.76 123 macro avg 0.73 0.66 0.67 123 weighted avg 0.74 0.76 0.73 123

As a part of the project, the models under our consideration are:

1) Support Vector Machine
2) Logistic Regression
3) Random Forest Classifier

From the above designed models, we have considered Random Forest Classifier as a model with 97% train accuracy to predict the eligibility criteria of loan applicants.

4

Dharavath Sai Kiran, Avula Dheeraj Reddy, Suneetha Vazarla, Dileep P.
Loan Approval Prediction using Adversarial Training and Data Science.
TURCOMAT [Internet]. 2023Jul.11 [cited 2024Jun.20];14(03):356-60. Available from:
https://turcomat.org/index.php/turkbilmat/article/view/13978

```
print('Accuracy:', accuracy)
print('Precision:', precision)
print('Recall:', recall)
print('F1-score:', f1)
```

```
Accuracy: 0.6521739130434783
Precision: 0.6593406593406593
Recall: 0.9836065573770492
F1-score: 0.7894736842105262
```

Fig 6: Evaluation Metrics

This paper proposes an approach to predict loan approval using Adversarial Training and Data Science techniques. We described the methodology, which collects and preprocesses the data, creates the model and uses adversarial training to improve the model's performance. We implemented the model in Python and evaluated its performance. The algorithm reaches 65% accuracy. The accuracy may increase when the model is trained with vast amounts of data as input to it. We also drew a confusion matrix to understand the model's performance better.

5

Nazim Uddin, Md. Khabir Uddin Ahamed, Md Ashraf Uddin, Md. Manwarul Islam, Md. Alamin Talukder, Sunil Aryal, An ensemble machine learning based bank loan approval predictions system with a smart application, International Journal of Cognitive Computing in Engineering, https://doi.org/10.1016/j.ijcce.2023.09.001. (https://www.sciencedirect.com/science/article/pii/S2666307423000293)

The goal of this system is to improve the accuracy and efficiency of loan approval processes in banks by leveraging advanced machine learning models. The study employs a variety of machine learning models, including, **Logistic Regression, Decision Tree**, **Random Forest**, **Extra Trees**, **SVM, KNN**, **Gaussian Naive Bayes, AdaBoost**, **Gradient Boosting.**

The performance of each model was evaluated based on metrics such as accuracy, recall, and F1-score. Among the individual models, the Extra Trees classifier demonstrated superior performance. The ensemble voting classifier, which included the best-performing models, achieved an accuracy of 87.26%. This represented a 0.62% increase in accuracy compared to the Extra Trees classifier alone.

Table 10
Model wise highest accuracy comparison with existing work.

	Model	Accuracy (%)	Paper
Existing Work	Extra Tree	86.2	(Anand et al., 2022)
	Decision Tree with AdaBoost	84	(Kumar et al., 2022)
	Logistic Regression	78.5	(Dosalwar et al., 2021)
	Multilayer Perceptron	80	(Alsaleem & Hasoon, 2020)
	Naïve Bayesian	80.4	(Blessie & Rekha, 2019)
Proposed work	Model Wise(ExTree)	86.64	Table 9
	Ensemble with All	84.07	
	Ensemble with best Three	87.26	

6

Suliman Mohamed Fati
Machine Learning-Based Prediction Model for Loan Status Approval
Journal of Hunan University（Natural Sciences） Vol. 48 No. 10 October 2021
http://jonuns.com/index.php/journal/article/view/783

Table 4 Summary results of the three machine learning algorithms

Algorithms	Precision	Recall	F1	Accuracy	AUC
Logistic Regression	0.79	0.98	0.88	0.91	0.80
Decision tree	0.77	0.83	0.80	0.82	0.75
Random forest	0.78	0.93	0.84	0.86	0.79

Loan application processing is one of the main tasks for banks. Many approaches are proposed in the literature for loan prediction. Among those approaches, machine-learning algorithms are proposed to predict the loan status based on different factors and parameters. Therefore, in this paper, we have trained and tested a dataset to predict loan approval. There are 13 features in the dataset, and we found that Credit_History is the most important feature for the prediction of loans

Three machine-learning algorithms were trained and tested in the proposed prediction model on the data: linear regression, decision tree, and Random Forest. Logistic regression showed better performance than the others with 81% accuracy, while Decision tree and Random Forest got 72% and 76% accuracy, respectively. The recorded results have been validated using the ROC curve.

7

Nitesh Pandey, Ramanand Gupta, Sagar Uniyal, Vishal Kumar (June 2021).
Loan Approval Prediction using Machine Learning Algorithms Approach
IJIRT | Volume 8 Issue 1 | ISSN: 2349-6002 available:
https://ijirt.org/Article?manuscript=151769

In this paper, machine learning was used to predict loan acceptance. The prediction method begins with data pre-processing, filling the missing values, experimental data analysis. After evaluating model on test dataset, each of these algorithms obtained a precision rate between **70% and 80%.** Although here it can be concluded with certainty that the **Support Vector Machine model** is very efficient and produces superior results than other models.

Algorithm	Accuracy (in %)	Precision	Recall	F1 Score
Random Forest	76.42	0.46	0.80	0.58
Support Vector Machine	79.67	0.46	0.95	0.61
Decision Tree	70	0.51	0.60	0.55
Logistic Regression	75.60	0.6	0.69	0.64

8

Mridul Bhandari
Medium, “How to predict Loan Eligibility using Machine Learning Models" (2020, Sept 15) available: https://towardsdatascience.com/predict-loan-eligibility-using-machine-learning-models-7a14ef904057

Model	Accuracy (%)
Logistic Regression (without Feature Engg.)	78
LR, stratified k-folds cross-validation (w/o FE)	80
Logistic Regression	72
Decision Tree	71
Random Forest	79
XGBoost	77.5

9

Mridul Bhandari
Medium, “ML basics: Loan prediction" (2019, June 6) available:
https://towardsdatascience.com/ml-basics-loan-prediction-d695ba7f31f6

The article discusses using various machine learning models to predict loan eligibility, employing K-fold cross-validation to evaluate performance. Three models are tested: Logistic Regression, Decision Tree, and Random Forest. Logistic Regression, using predictors like Credit History and Education, achieves an accuracy of 80.945% and a cross-validation score of 80.946%. The Decision Tree, with a slightly different set of predictors, also achieves an accuracy of 80.945% but a lower cross-validation score of 78.179%, indicating overfitting. The Random Forest model, using a more extensive set of predictors, shows perfect accuracy at 100% but a lower cross-validation score of 78.015%, demonstrating significant overfitting. The article suggests reducing predictors or tuning model parameters to address overfitting issues.

10

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.

The process of prediction starts from cleaning and processing of data, imputation of missing values, experimental analysis of data set and then model building to evaluation of model and testing on test data. On Data set, the **best case accuracy obtained on the original data set is 0.811 (Logistic Regression)**. The following conclusions are reached after analysis that those applicants whose credit score was worst will fail to get loan approval, due to a higher probability of not paying back the loan amount. Most of the time, those applicants who have high income and demands for lower amount of loan are more likely to get approved which makes sense, more likely to pay back their loans. Some other characteristic like gender and marital status seems not to be taken into consideration by the company.

11

G. Arutjothi and C. Senthamarai,
"Prediction of loan status in commercial bank using machine learning classifier,"
2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 2017,
pp. 416-419, doi: 10.1109/ISS1.2017.8389442.

In this paper, a loan status model is proposed to predict whether a loan applicant is a valid customer or a default customer. The proposed model demonstrates an accuracy of 75.08% in classifying credit applicants using the R package. Credit lenders can utilize this model to make informed decisions on loan proposals. Furthermore, a comparison study has been conducted with different levels of iterations. The model with 30 iterations based on the k-NN algorithm provides significant accuracy compared to other levels. This model can help commercial banks avoid substantial losses.

TABLE 2. CLASSIFICATION RESULTS

S. No	K-Nearest Neighbor model (k=30)				
	Percentage Split	*Accuracy*	*RMSE*	*Correlation*	*Error Rate*
1.	80% :20%	74.5	2.04	0.99	0.256
2.	70% :30%	73.4	1.91	0.98	0.259
3.	60% :40%	74.9	2.29	0.92	0.25
4.	50% :50%	75.08	2.41	-0.26	0.24

12

Vishnu Vardhan
Medium, “Case Study – Loan Prediction" (2014) available:
https://medium.com/@vishnumbaprof/case-study-loan-prediction-ac035f3ec9e4

I have tried various techniques like Random Forest, Decision Tree, Decision Tree etc. and came to conclusion that the above code gave maximum accuracy. However there is still a lot of room to enhance accuracy which I have to figure it out still.

As we can see **average accuracy is 81.11%** . I have used the same thing for predicting test data variable. However, how much ever I try I ended up with maximum accuracy of **79.166%**

Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002).
An Introduction to Logistic Regression Analysis and Reporting.
Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.

In this paper, we demonstrate that logistic regression can be a powerful analytical technique for use when the outcome variable is dichotomous. The effectiveness of the logistic model was shown to be supported by
(a) significance tests of the model against the null model,
(b) the significance test of each predictor,
(c) descriptive and inferential goodness-of-fit indices,
(d) and predicted probabilities

During the last decade, logistic regression has been gaining popularity. The trend is evident in the JER and higher education journals. Such popularity can be attributed to researchers' easy access to sophisticated statistical software that performs comprehensive analyses of this technique. It is anticipated that the application of the logistic regression technique is likely to increase

14

Breiman, L. Statistics Department, University of California, Berkeley, CA, 94720
Random Forests. Machine Learning 45, 5–32 (2001).
https://doi.org/10.1023/A:1010933404324

The article "Random Forests" by Leo Breiman, published in Machine Learning, introduces the Random Forest algorithm, an ensemble learning method for classification and regression. The technique involves constructing multiple decision trees during training and outputting the mode of the classes for classification or mean prediction for regression from individual trees. This method enhances the model's accuracy and robustness, addressing issues like overfitting and improving generalization. The paper demonstrates the effectiveness of Random Forests through various empirical results, highlighting their superiority over single decision trees and other ensemble methods.

Our Models

- Decision Tree
- Random Forest
- Logistic Regression
- Support Vector Classifier
- KNN


Literature Survey

Thank You