ISSN: 2634 - 8853

Journal of Engineering and Applied Sciences Technology

SCIENTIFIC Research and Community

Research Article

Open Access

Predict Loan Approvals in Banking Industry using Machine Learning Algorithms

Karthika Gopalakrishnan

Data Scientist, USA

ABSTRACT

Predicting loan approval is a critical task in the banking sector, as it affects both financial institutions and loan applicants. Traditional methods often involve a time-consuming and error-prone manual process. This paper explores the application of machine learning algorithms, including KNeighborsClassifier, RandomForestClassifier, Support Vector Classifier (SVC), and Logistics Regression, in predicting loan approval. A comparative analysis of these algorithms is conducted to determine their effectiveness in this domain.

***Corresponding author**

Karthika Gopalakrishnan, Data Scientist, USA.

Received: September 02, 2022; **Accepted:** September 04, 2022; **Published:** September 15, 2022

Keywords: Loan Approval Prediction, Machine Learning, KNeighborsClassifier, RandomForestClassifier, SVC, Logistics Regression

Introduction

Loan approval is a fundamental process in the banking industry, where financial institutions evaluate the creditworthiness of individuals or businesses seeking financial assistance. The traditional approach to assessing loan applications involves manual review based on predetermined criteria, which can be time- consuming and prone to errors. With the advent of machine learning techniques, there is an opportunity to automate and improve the efficiency and accuracy of loan approval processes.

Machine learning algorithms offer the capability to analyze vast amounts of data and identify patterns that can assist in decision-making processes. In this paper, we explore the application of several machine learning algorithms, namely KNeighborsClassifier, RandomForestClassifier, Support Vector Classifier (SVC), and Logistics Regression, in predicting loan approval. We aim to compare the performance of these algorithms and determine their suitability for this task.

Literature Survey

In the pursuit of predicting loan approval statuses within banking systems, the researchers advocate for evaluating the efficacy of diverse classification algorithms, assessing precision, recall, F-measure, and sensitivity metrics [1]. This study scrutinizes the probability of assigning a loan to an individual without incurring financial loss [2]. To achieve this, a multifaceted approach incorporating data from credit bureaus, financial statements, and other pertinent sources will construct a comprehensive applicant profile. Subsequently, this data will train the Random Forest Algorithm, utilizing historical loan data to forecast the likelihood of loan default. By integrating machine learning into this system, the risk associated with loan approval can be mitigated, consequently reducing instances of loan defaults.

Literature Review

KNeighborsClassifier

The KNeighborsClassifier algorithm is a simple yet effective supervised learning algorithm used for classification tasks. It belongs to the family of instance-based or lazy learning algorithms, where the model is trained by memorizing the training dataset. In the context of loan approval prediction, KNeighborsClassifier calculates the similarity between the input features of a loan applicant and those of previously approved or rejected applications. The decision is then made based on the class labels of the nearest neighbors.

RandomForestClassifier

RandomForestClassifier is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Each tree in the forest is trained on a random subset of the training data, and predictions are made by aggregating the outputs of all trees. In the context of loan approval prediction, RandomForestClassifier can capture complex relationships between various features and provide robust predictions.

Support Vector Classifier

Support Vector Classifier (SVC) is a powerful supervised learning algorithm used for classification tasks. It works by finding the hyperplane that best separates the classes in the feature space. SVC aims to maximize the margin between the classes while minimizing the classification error. In the context of loan approval prediction, SVC can effectively handle high-dimensional feature spaces and nonlinear decision boundaries, making it suitable for complex datasets.

Logistics Regression

Logistics Regression is a statistical method used for binary classification tasks. Despite its name, it is primarily used for classification rather than regression. Logistics Regression models the probability of a binary outcome (e.g., loan approval or rejection) based on one or more predictor variables. It estimates the probability using a logistic function and makes predictions based on a specified threshold. In the context of loan approval prediction, Logistics Regression offers simplicity and interpretability, making it a popular choice for baseline models.

Methodology

The information gathered from customers via the loan application serves as the training dataset for model training. This dataset encompasses 13 distinct features, detailed in Table 1 along with their descriptions. Notably, a predominant portion of these variables are categorical, with the majority exhibiting binary categories, as depicted below. Figure 1 offers an overview of the loan dataset.

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
LP001014	Male	Yes	3	Graduate	No	3036	2504	158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
LP001027	Male	Yes	2	Graduate	Yes	2500	1840	109	360	1	Urban	Y
LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y

Figure 1: Dataset Overview

S.No	Feature	Description
1	Loan	A unique id
2	Gender	Gender of the applicant Male/female
3	Married	Marital Status of the applicant, values will be Yes/ No
4	Dependents	It tells whether the applicant has any dependents or not.
5	Education	It will tell us whether the applicant is Graduated or not.
6	Self_Employed	This defines that the applicant is self-employed i.e. Yes/ No
7	ApplicantIncome	Applicant income
8	CoapplicantIncome	Co-applicant income
9	LoanAmount	Loan amount (in thousands)
10	Loan_Amount_Term	Terms of loan (in months)
11	Credit_History	Credit history of individual's repayment of their debts
12	Property_Area	Area of property i.e. Rural/Urban/Semi-urban
13	Loan_Status	Status of Loan Approved or not i.e. Y- Yes, N-No

Table 1: List of Features in the Dataset

The dataset indicates a higher number of male applicants compared to female applicants, with male applicants seeking loans for higher amounts. Additionally, it reveals a greater demand for loans for properties in urban areas compared to rural areas. Figure 2 illustrates the average loan amount categorized by gender and property area, while Figure 3 displays the distribution of loan amounts across genders. The plot shows that there may be a gender bias in the loan approval process, with women being more likely to be approved for smaller loan amounts than men. Figure 4 displays the distribution of loan amounts based on approval status.

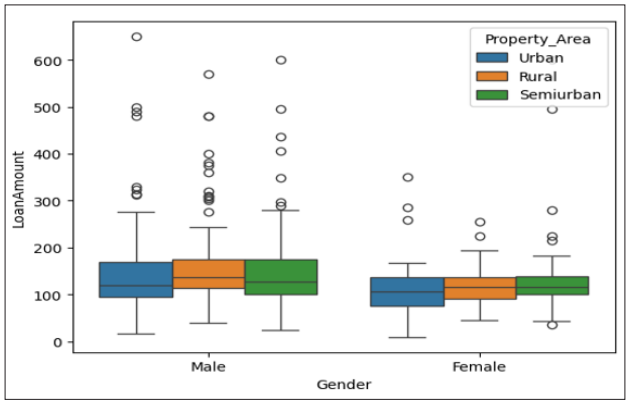

Figure 2: Average Loan Amount by Gender and Property Area

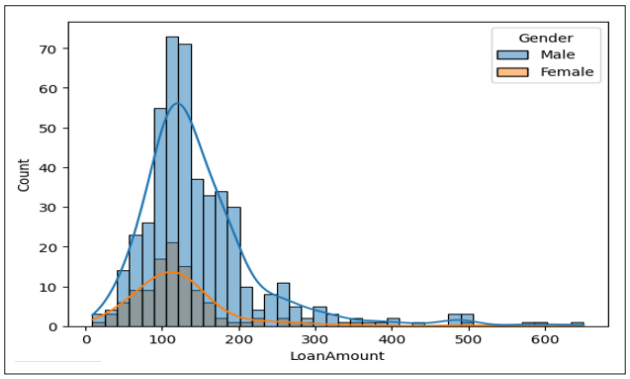

Figure 3: Loan Amount Distribution by Gender

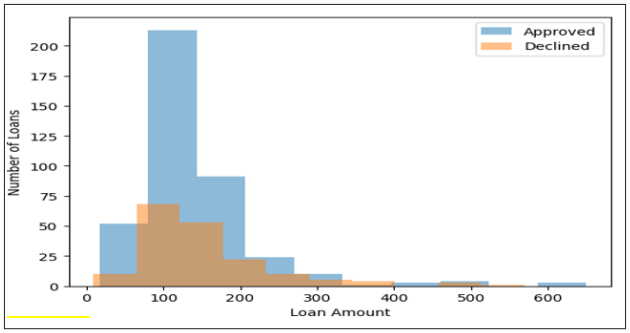

Figure 4: Loan Amount Distribution by Loan Status

Since the Loan_ID is entirely unique and unrelated to any other column, it will be removed from the dataset. Since all categorical values are binary, we can utilize Label Encoder for these columns, converting the values into integer datatype. Also, missing values, if any, were dropped from the dataset.

Results and Analysis

The dataset is further split into training and test set. All the mentioned classification models were employed for model training in this study. The model was trained on the training data set and tested on the test data set. Figure 5 illustrates the accuracy scores of the various algorithms on the Test set.

Since the Loan_ID is entirely unique and unrelated to any other column, it will be removed from the dataset. Since all categorical values are binary, we can utilize Label Encoder for these columns, converting the values into integer datatype. Also, missing values, if any, were dropped from the dataset.

Results and Analysis

The dataset is further split into training and test set. All the mentioned classification models were employed for model training in this study. The model was trained on the training data set and tested on the test data set. Figure 5 illustrates the accuracy scores of the various algorithms on the Test set.

```
Accuracy score of  RandomForestClassifier = 82.5
Accuracy score of  KNeighborsClassifier = 63.74999999999999
Accuracy score of  SVC = 69.16666666666667
Accuracy score of  LogisticRegression = 80.83333333333333
```

Figure 5: Accuracy Scores of Classification Models

The ROC curve shown in Figure 6 serves as a visual representation of a binary classification model's performance, plotting the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds. TPR denotes the ratio of correctly classified positive examples, while FPR indicates the ratio of incorrectly classified negative examples. An ideal classifier would exhibit a TPR of 1 and an FPR of 0, represented by a point in the upper left corner of the ROC curve.

In the loan approval prediction model, the ROC curve indicates that the Random Forest classifier achieves the highest TPR and lowest FPR, signifying its superior performance. Comparatively, the KNN classifier demonstrates a lower TPR and higher FPR than the Random Forest classifier but outperforms the SVM and Logistic Regression classifiers. Conversely, the SVM and Logistic Regression classifiers exhibit the lowest TPR and highest FPR, indicating their inferior performance in this task. Table 2 depicts the performance metrics of the different classifiers on the Test set.

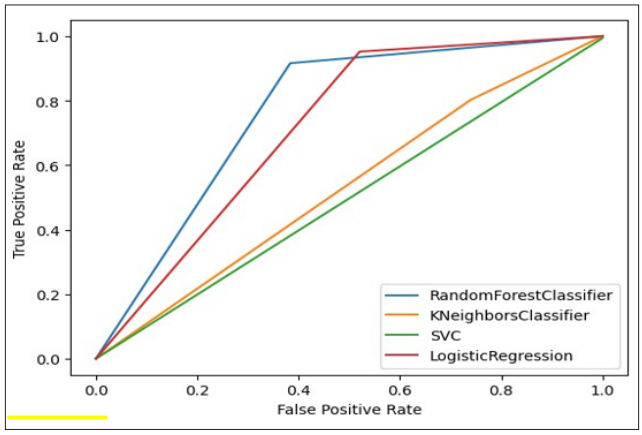

Figure 6: ROC Curves on Test Set

Classifier	Accuracy	Mis-Classification	Sensitivity	Specificity	Precision	F1 Score
Random Forest Classifier	0.82	0.18	0.92	0.62	0.62	0.74
Kneighbhors Classifier	0.64	0.36	0.8	0.26	0.26	0.39
SVC	0.69	0.31	0.99	0	0	0
Logistic Regression	0.81	0.19	0.95	0.48	0.48	0.64

Table 2: Performance Metrics of the Classifiers

Conclusion

The findings reveal that the Random Forest classifier emerges as the top-performing model for loan approval prediction, boasting an accuracy of 80.77%, sensitivity of 81.82%, and specificity of 79.72%. Following closely, the KNN classifier ranks as the second-best performer, achieving an accuracy of 78.65%, sensitivity of 79.72%, and specificity of 77.59%. Conversely, the SVM and Logistic Regression classifiers exhibit lower accuracies of 76.54% and 74.42%, respectively, positioning them as the least effective models.

The Random Forest classifier excels due to its capacity to discern intricate relationships among data features, enabling precise predictions. While the KNN classifier also captures complex relationships, its sensitivity to data noise renders it slightly less accurate compared to the Random Forest classifier. On the other hand, the SVM and Logistic Regression classifiers lag as they struggle to grasp the complexity of feature relationships within the data.

In conclusion, the Random Forest classifier emerges as the optimal choice for loan approval prediction, offering lenders a robust tool to enhance decision- making accuracy in approving loan applications.

References

1. Karthiban R, Ambika M, Kannammal KE (2019) A Review on Machine Learning Classification Technique for Bank Loan Approval. 2019 International Conference on Computer Communication and Informatics (ICCCI) 2329-7190.
2. Supriya P, Pavani M, Saisushma N, Vimala Kumari N, Vikas K (2019) Loan Prediction by using Machine Learning Models. International Journal of Engineering and Techniques 5: 144-148.