Loan Approval Prediction using Credit Card Score Analysis

J. Sri Sai Samhitha[1], K. Adarsh Sagar[2], Kundula Haritha[3], K . Dinesh Kumar[4]

[1 2 3 4]Dept. of Computer Science and Engineering, Amrita School of Computing, Bengaluru, Amrita Vishwa Vidyapeetham, India.

Contributing authors: bl.en.u4cse21072@bl.students.amrita.edu; bl.en.u4cse21075@bl.students.amrita.edu; bl.en.u4cse21107@bl.students.amrita.edu; kk_dinesh@blr.amrita.edu;

Abstract

Banking sector offers a large variety of financial services out of which the significant amenities include payment services, custodial and safety support, loan application assistance and credit card options. Machine Learning and Data Science has a vast application in the field of Banking sector. These domains play significant role in the field of loan prediction which helps in fraud detection, risk management system, maintaining efficiency, transparency and helping in continuous improvement. Due to the increasing rate of loan defaults, it's a difficult task for the banking authorities to assess the loan requests and deal with the risks of customers defaulting the loan. An automatic loan prediction system is developed using machine learning models where the machine learns and predicts whether to approve the loan or not based on the eligibility criteria. Six prediction models including Logistic Regression Model, Decision tree, Random Forest, Support Vector Machine, K-Nearest Neighbor Classifier and XG-Boost algorithm have been performed. A comparative analysis is carried out between the models where the performance metrices for each model is calculated and compared based on which the best suitable model for predicting the loan approval status is concluded.

Keywords: Machine Learning, Banking Sector, Loan Approval Prediction

1 Introduction

The banking sector is a keystone of the financial system across the globe for promoting economic growth and manage the financial matters. Banking sector aids in providing various financial services. The domains of Machine Learning, Data Science and Data Analytics all play an important role in the banking sector for credit card score analysis and in the process of predicting the loan approval status. Many predicting techniques aid in assessing creditworthiness, which help in evaluating the risk associated with lending loans to a particular individual. Using Data Analytical tools, banks are able to find patterns and derive insights from the customer data, which help banks to decide on whom to lend loans and make their customers satisfied. By understanding the customers credit score details, history and the risk associated in lending loans to them, the banking sector will be able to make more optimized decisions regarding lending loans.

Considering the current situation of banking sector, there is a tremendous increase in the trend of taking loans, by which the banking sectors are facing new challenges. One of the main challenges faced by the banks is determining which customer should be granted loan. Assessing the loan requests, evaluation of customer details such as credit card score, loan history, annual income, loan amount and other crucial features has become a complex and challenging task for all the banks to decide on which customer to lend the loan. These procedures are lengthy and time consuming which makes the customers dissatisfied and cause delays in the process of decision-making.

This study aims to analyse the customers risk level and perform a relative assessment between the prediction models to analyse and opt the best suitable model for approving loans to the customers. Performance evaluation metrices of six machine learning models like Logistic Regression Model, Decision tree, Random Forest, Support Vector Machine, K-Nearest Neighbour Classifier, XG-Boost algorithm have been assessed to figure out the most reliable and accurate model for loan approval prediction.

2 Literature Survey

U. E. Orji et al. [1] The authors in this work have discussed the significance of advanced Machine Learning models in the loan approval prediction in the banking sector. A parallel assessment of six machine learning models which includes Logistic Regression Algorithm, K-Nearest Neighbor, Support Vector Machine (SVM), Decision Tree, Bagging and Boosting algorithms was performed by comparing the evaluation indicators calculated for each model. The techniques used for pre-processing the data are SMOTE, one -hot encoding and Normalization. The comparative study showed that Random Forest model is the most reliable model to predict the loan approval. All the six models have outperformed in comparison with the three models highlighted in the paper's literature.

J. Sinha et al. [2] The authors studied the importance of machine learning models in implementing an accurate model for predicting the approval of loan to a customer by analyzing the customers credit risks. The Machine Learning techniques used by the authors are Logistic Regression, K nearest neighbour and Random Forest Classifier.

The evaluation metrices are assessed for each of the predicting approaches and the predictions made by the model are depicted using confusion matrix.

R. Priscilla et al. [3] The authors performed a comparative analysis between machine learning algorithms for selecting a credible loan approval prediction model to ease the task of banks to decide on the trustworthy customers for sanctioning loans and reduce the risk of financial loss for banks. They studied various Machine Learning algorithms to build a loan approval prediction model which are Logistic Regression (LR), Support Vector Classification (SVM), Random Forest Classifier, K-Neighbor Classifier, Gradient Boosting, Perceptron, Decision Tree Classifier. Then they have compared the evaluation metrices of each ML algorithm which shows that Random Forest Classifier has the highest accuracy of 96.82 percentage. The Perceptron has the lowest accuracy of 90.52 percentage. Through this study they concluded Random Forest Classifier as the best algorithm to implement the model.

V. Singh et al. [4] The authors have made use of machine learning algorithms like XG- Boost, Random Forest and Decision Tree to predict loan approval for the new applicants in the bank based on the hitorical data of the customers. They have performed feature selection on the data and figured out that ”zipcode” and ”credit-history” are the most contributing features to predict the loan status of a customer.

P. S. Saini, et al. [5] The authors conducted a comparative analysis between 4 machine learning(i.e, random forest classifier ,k-nearest neighbors, logistic regression, support vector classifier) algorithms on the loan eligibility. The dataset has been pre-processed by handling missing values, encoding categorical variables and performing feature scaling. Out of these algorithms random forest model achieved greatest accuracy of 98.04 percentage whereas support vector classifier had lower accuracy of 68.71 percentage.

S. K. Hegde et al. [6] In this work, a machine learning model has been constructed by comparing the performance of decision tree, XG-Boost, Random Forest, and Logistic regression techniques for approval of loan based on different factors.Compared to other machine learning algorithms, the Logistic regression technique has provided superior prediction result. The data is initially separated using k-fold cross-validation and to evaluate the characteristics and target variables of data they employed various visualization approaches.

M. A. Sheikh et al. [7] The authors have built a Logistic regression model by training the dataset and the different measures of performances are computed. In data pre-processing, apart from the traditional data cleaning techniques, data reduction technique is used to deal with huge data, such has PCA (principal Component Analysis) and along with it data mining techniques are performed on the data.

A. Gupta et al. [8] The author addresses the increasing challenges in the banking sector by proposing a machine learning-based system for loan approval prediction, and also has discussed the importance of Machine Learning in the field of banking sector and explained the types of machine learning: supervised learning and unsupervised learning. The Algorithms utilized for prediction in this article include Logistic regression, Random Forest and the Correlation between parameters is obtained.

Yashna Sayjadah et al. in the paper [9] of Credit card default prediction using machine learning techniques made use of a data set which consisting of 30000 instances and 24 attributes generated by credit card users. The process of prediction is done using the steps data pre-processing, data partitioning, data visualization with machine learning algorithm. The algorithms they have tested includes decision tree, logistic regression model and random forest to predict the default credit card score and these algorithms have 82 percentage accuracy.

Ch. Naveen kumar et al. [10] in the paper titled Customer loan eligibility prediction using machine learning algorithms in banking sector performed data cleaning on the dataset followed by feature selection. They made use of the machine learning algorithms including decision tree, random forest, support vector machine, K-nearest neighbour and Ada-Boost.

3 Methodology

This section describes the dataset, proposed architecture and the data preprocessing steps used in developing the model. Python language is used to train and test the various machine learning models on the considered dataset. The experimental environment used to implement the python codes is Jupyter notebook on windows Operating System. Jupyter notebook is a very user-friendly and open -source platform where the python codes can be executed using large number of inbuilt libraries with ease.

3.1 Dataset

The dataset utilised for this investigation is acquired from Kaggle. It has 4219 customer details and 12 significant attributes that play a crucial role in accurately forecasting the loan approval status. The dataset includes the following features: loan id, number of dependents, education status, self-employment status, annual income, loan amount, loan term, credit score, residential assets value, commercial assets value, luxurious assets value, bank asset value, and the target variable loan status. The loan approval status of all 4219 consumers is predicted by the machine learning models based on these variables.

3.2 Proposed Architecture

This section gives a pictorial representation of the flow of the architecture proposed to develop a loan approval prediction model. Intially the dataset is collected followed by performing pre-processing steps on the dataset. The dataset is then split in the ratio of 80:20 for training and testing the machine learning models. The performance evaluation metrices are calculated and compared based on which the best suitable model for predicting the loan approval prediction is adopted. Fig. 1 gives a clear view about the flow of the proposed system.

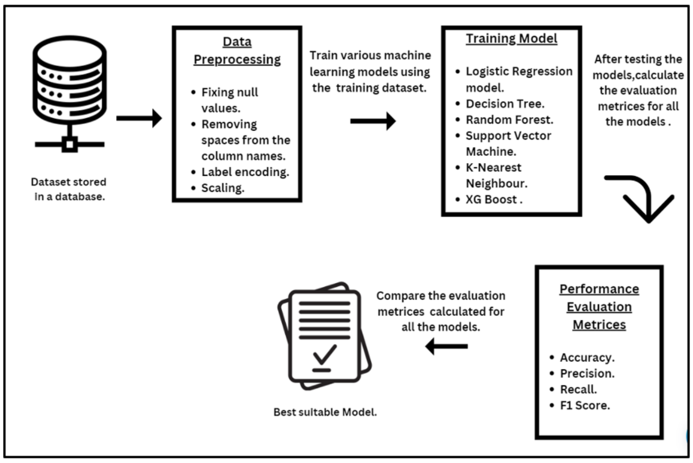
Dataset stored in a database.
Data Preprocessing
• Fixing null values.
• Removing spaces from the column names.
• Label encoding.
• Scaling.
Train various machine learning models using the training dataset.
Training Model
• Logistic Regression model.
• Decision Tree.
• Random Forest.
• Support Vector Machine.
• K-Nearest Neighbour.
• XG Boost .
After testing the models,calculate the evaluation metrices for all the models .
Performance Evaluation Metrices
• Accuracy.
• Precision.
• Recall.
• F1 Score.
Compare the evaluation metrices calculated for all the models.
Best suitable Model.

Fig. 1 Architecture of the model

3.3 Data Preprocessing

Several data pre-processing steps have been performed on the dataset which includes checking for null values in dataset, removing the spaces in the column names, performing label encoding of the data and scaling the data.

1. Using the isnull().sum() function it is found that there are no null values in any of the columns of the dataset and any spaces present in the names of the columns have been removed.
2. Label encoding has been performed on the columns of education, self-employed and the target variable loan-status where all the categorical values in the columns have been mapped to numerical values. Label encoding converts the data into numerical format, which the machine learning algorithms can understand and work easily with. Label encoding also reduces the memory space used.
3. Scaling is performed on all the numerical input features where all the independent variables are standardized in a fixed range of values, so that all the input features contribute equally for the models to learn and determine the loan status (target variable).

4 Model Development

This section deals with the various machine learning models used in the project along with the performance evaluation metrices. The dataset is spilt in the ratio of 80:20, where 80 percentage of the data is used for training the models and 20 percentage of the data is used to test the models. Several Machine Learning models have been trained and tested on the dataset to find the best suitable model for loan price prediction.

4.1 Performance Evaluation Metrices

Evaluation metrices or the performance metrices are used the analyse the performance and efficiency of the machine learning models on the given dataset. The research takes accuracy, precision, recall, and F1-score into account as performance metrics. The confusion matrix offers us a picture about the amount of accurate and wrong predictions generated by the model by dividing the entire number of test cases into several groups of True Positive(TP), True Negative(TN), False Positive (FP) and False Negative (FN) based on the predictions made by the model out of all data instances classified by the model. The metrices used to compare the performance of the models includes Accuracy, Precision, Recall and F1-Score.

4.2 Machine Learning Models

The various machine learning models trained and tested are explained in detail.

4.2.1 Logistic Regression Algorithm

A supervised machine learning model that is frequently used to solve classification issues is the logistic regression algorithm. This method works out the link between the independent (input) variables and the binary dependent (output) variable and finds out the likelihood of a new data instance of belonging to a certain class. For the considered dataset, this model has been tested on the testing set and has predicted the output of loan-approval class of each of the testing dataset as yes (1) or no (0).

4.2.2 Decision Tree

Using decision trees, a supervised machine learning algorithm is used to solve regression and classification problems. It represents a hierarchical structure similar to a tree and consists of the root node representing the entire dataset, internal node which represents the choice made with respect to a input attribute, leaf nodes representing final class labels and branches which represent the outcome based on the conditions on the internal nodes. This algorithm works by splitting the entire dataset into smaller sets recursively by selecting the suitable feature by calculating the entropy and information gain at each step. The information gain is calculated using the below formula:

$$\mathrm{IG}(T, a) = H(T) - H(T \mid a) \tag{1}$$

In Equation (5),

$$\begin{aligned} \mathrm{IG}(T, a) &: \text{Information gain of dataset } T, \text{ with respect to attribute } a. \\ a &: \text{Value of attribute.} \\ H(T \mid a) &: \text{Conditional entropy of dataset } T \text{ given attribute } a. \\ H(T) &: \text{Entropy on the dataset } T. \end{aligned}$$

4.2.3 Random Forest

Random Forest is a supervised machine learning technique which is based on ensemble learning. This technique works by mixing varying number of decision trees which act on subsets of the provided dataset. The average of the findings of each decision tree is taken as the final result. This method contributes to both the model's accuracy and performance improvements. As the number of decision trees incorporated rises, the total accuracy of random forest likewise increases. The accuracy, precision, recall, and F1-score of this method are computed after it has been trained and tested on the dataset in order to assess the model's performance.

4.2.4 Support Vector Machine

Support Vector Machine (SVM) is a type of supervised machine learning method that may be applied to solve classification and regression issues, but works best on binary classification problems. This algorithm aims at finding out most optimal hyperplane which will be classifying the points into their corresponding classes. Support vectors are the points which lie closest to the hyperplane, and this algorithm figures out the hyperplane such that the distance between the support vectors and the hyperplane will be maximum. This algorithm works for both linearly separable data and on non - linearly separable data by using kernel functions. This model has been tested on the dataset and the performance metrices of this model are evaluated. The decision function of SVM algorithm is given as :

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \tag{2}$$

In Equation (6),

$$\begin{aligned} f(x) &: \text{Decision function.} \\ N &: \text{Number of support vectors.} \\ \alpha_i &: \text{Lagrange multipliers.} \\ y_i &: \text{Class label of the } i\text{-th support vector.} \\ K(x_i, x) &: \text{Kernel function.} \\ b &: \text{Bias term.} \end{aligned}$$

4.2.5 K-Nearest Neighbor

The supervised machine learning technique, K-Nearest Neighbour (KNN) classifier is mostly used to solve classification issues. This approach takes into account the classes of k points that are closest to the test instance. The k nearest points to the test instance are found out by calculating the Euclidean distance or Manhattan distance. The majority or the mean class of all the classes of k nearby points is the predicted class of new instance. The performance of KNN classifier on the dataset considered is assesed by calculating the performance evaluation metrices.

4.2.6 XG-Boost Algorithm

Extreme Gradient Boost algorithm is a machine learning algorithm which uses an ensemble technique of decision trees and gradient boosting. This model is generally applied on structured data. It combines the predictions made by multiple decision trees, where one decision tree corrects the errors made by the previous decision tree. By doing this the overall accuracy of the predictions made by the model is increased. In this paper, the model is trained and tested using the dataset and the performance of this model is evaluated using accuracy, precision, recall and F1-score values.

5 Results and Discussions

Multiple machine learning models were trained and evaluated on the dataset by splitting it into an 70:30 ratio. Performance metrics such as accuracy, precision, recall and F1 Score were generated for each model.The XG Boost method had the greatest values for accuracy (0.9813), precision (0.9813), recall (0.9813), and F1 score (0.9812) compared to the Logistic Regression Model, Decision Tree, Random Forest, and Support Vector Machine. The Support Vector Machine model exhibits the lowest performance, with an accuracy of 0.6276, precision of 0.3939, recall of 0.6276, and F1 score of 0.4840. Therefore, it is evident that the XG-Boost Algorithm is the optimal and effective algorithm for accurately forecasting the loan acceptance status with superior performance. Comparison of performance evaluation metrices for various machine learning models is shown in Table 1.

Table 1 Comparison of performance evaluation metrices of machine learning models

Algorithm	Accuracy	Precision 3	Recall	F1-Score
Logistic Regression	0.908	0.908	0.908	0.908
Decision Tree	0.974	0.974	0.974	0.974
Random Forest	0.976	0.976	0.976	0.976
Support Vector Machine	0.627	0.393	0.627	0.484
K-Nearest Neighbor	0.594	0.574	0.594	0.579
XG-Boost	0.981	0.981	0.981	0.981

6 Conclusion

This work discusses the design of a robust and dependable loan approval prediction system employing a range of machine learning algorithms, such as the Logistic Regression Model, Decision Trees, Random Forest, Support Vector Machine, K Nearest Neighbour, and XG-Boost algorithm. The dataset considered has initially undergone several data preprocessing techniques, followed by training and testing various machine learning models and evaluating the performance evaluation metrices for each model. The result of performance evaluation metrices calculated for all the machine learning algorithms are compared, and it shows that XG-Boost Algorithm is the best and efficient

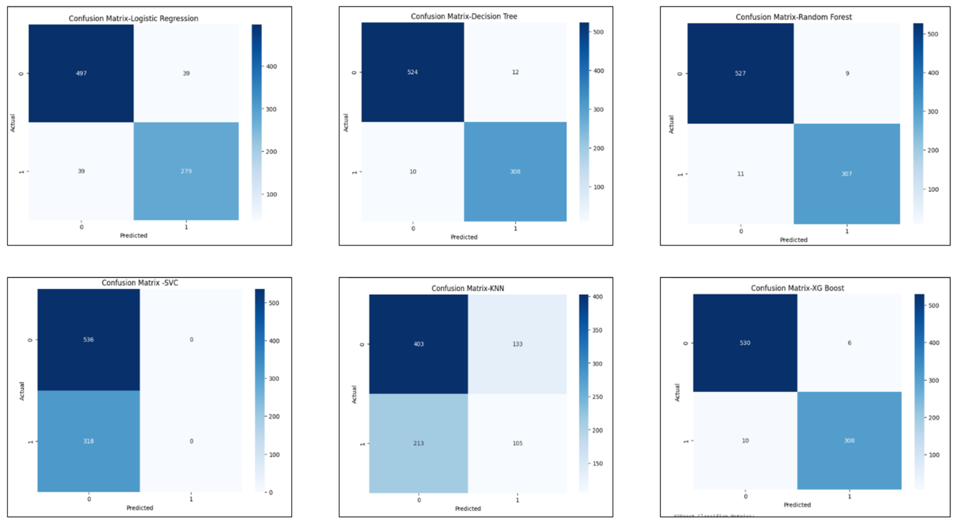

Fig. 2 Confusion Matrices of all the models

algorithm out of all models for predicting the approval of loan most accurately. Confusion matrix for the testing dataset of about 844 instances is visualized for which gives the idea regarding the number of true positives, true negatives, false positives and false negatives predicted by the model aiding in analysing the performance of the models. Fig. 2 represents the confusion matrices of the machine learning algorithms.

References

1. Orji, U.E., Ugwuishiwu, C.H., Nguemaleu, J.C. and Ugwuanyi, P.N. (2022). *Machine Learning Models for Predicting Bank Loan Eligibility*. In *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*, pp. 1-5.
2. Sinha, J., Astya, R., Tripathi, K., Verma, A. and Verma, M. (2021). *Machine Learning based Loan Allocation Prediction System for Banking Sector*. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pp. 1614-1619.
3. Priscilla, R., Siva, T., Karthi, M., Vijayakumar, K. and Gangadharan, R. (2023). *Baseline Modeling for Early Prediction of Loan Approval System*. In *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, pp. 1-7.
4. 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.
5. Saini, P.S., Bhatnagar, A. and Rani, L., 2023, May. Loan Approval Prediction using Machine Learning: A Comparative Analysis of Classification Algorithms. In 2023

3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 1821-1826). IEEE.

6. Hegde, S.K., Hegde, R., Marthanda, A.V.G.A. and Logu, K., 2023, February. Performance Analysis of Machine Learning Algorithm for the Credit Risk Analysis in the Banking Sector. In 2023 7th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 57-63). IEEE.
7. Sheikh, M.A., Goel, A.K. and Kumar, T., 2020, July. 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.
8. Gupta, A., Pant, V., Kumar, S. and Bansal, P.K., 2020, December. Bank Loan Prediction System using Machine Learning. In 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 423-426). IEEE.
9. Sayjadah, Y., Hashem, I.A.T., Alotaibi, F. and Kasmiran, K.A., 2018, October. Credit card default prediction using machine learning techniques. In 2018 Fourth International Conference on Advances in Computing, Communication Automation (ICACCA) (pp. 1-4). IEEE.
10. Kumar, C.N., Keerthana, D., Kavitha, M. and Kalyani, M., 2022, June. Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector. In 2022 7th International Conference on Communication and Electronics Systems (ICCES) (pp. 1007-1012). IEEE.
11. Sathyan, Dhanya K B, Anand Jose, Chinnu Aravind, N. (2018). Modelling the minislump spread of superplasticized PPC paste using RLS with the application of Random Kitchen sink. IOP Conference Series: Materials Science and Engineering. 310. 012035. 10.1088/1757-899X/310/1/012035.
12. Gayathri R, P B Pati Tripty Singh, ”A Framework for the prediction of Diabetes Mellitus using Hyper-Parameter tuned XGBoost Classifier”, 13th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Virtual, Oct 2022.
13. S. SasankVarthakavi, Babu, D. Rohith Pra, Reddy, L. Kumar Redd, and Remya Ajai A. S., “Analysis of preprocessing algorithms for face detection using KNN and SVM classifiers”, in 10th International Conference on Advances in Computing, Control, and Telecommunication Technologies, ACT 2019, 2019.
14. Sathyan, Dhanya K B, Anand Jose, Chinnu Aravind, N. (2018). Modelling the minislump spread of superplasticized PPC paste using RLS with the application of Random Kitchen sink. IOP Conference Series: Materials Science and Engineering. 310. 012035. 10.1088/1757-899X/310/1/012035.
15. M.S.Sivagama Sundari D.Periyasamy, ”SVM based MPC of Maglev system using PSO” 2022, Second International Conference on Sustainable Infrastructure with Smart Technology for Energy and Environmental Management, Bannari Amman Institute of Technology, Sathiamangalam.
16. M. G. Deepika, Sarika P. (2021), A Comparative Analysis of MFIs in India Using ANOVA and Logistic Regression Model, Advances in Intelligent Systems and Computing, 1133, 503-515.