International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 03 | Mar 2023 www.irjet.net p-ISSN: 2395-0072

Loan Default Prediction Using Machine Learning Techniques

Dr. E. Praynlin[1], Madesh S[2], Mohammed Thafeez H[3], Venu K V[4], Vinodh Kumar K[5]

[1]Assoc. Professor, Computer Science and Engineering, T. John Institute of Technology, Bengaluru, Karnataka, India
[2345] UG Students, Computer Science and Engineering, T. John Institute of Technology, Bengaluru, Karnataka, India

---

Abstract - *Loan business is one of the major income sources for bank. Loan default problem is a major issue for loan business. Loans, specifically whether borrowers repay the loan or default on it, have a significant impact on a bank's profitability. By anticipating loan defaulters, the bank is able to reduce its non-performing assets. Three primary predictive analytics techniques—I Data Collection, II Data Cleaning, and III Performance Assessment—are used to research the prediction of loan defaulters. Experimental investigations reveal that when it comes to loan forecasting, the KNN model performs better than the Decision tree model.*

Key Words: Machine learning, Loan prediction, Banking, Decision tree, KNN.

1.INTRODUCTION

Both applicants and bank workers find Loan Prediction to be highly useful. The goal of the article is to present a rapid, easy and instantaneous way of choosing the competent people. A Finance Company handles all loans. They provide services to all urban, semi-urban, and rural areas. The client submits their loan application once the business or bank confirms their qualification for a loan. The business or bank wants to automatically determine if a customer is eligible for a loan based on the information, they provide on the application form (in real time). There contains information about the borrower's gender, marital status, educational background, number of dependents, income, loan amount, and credit history. Data from former clients of multiple banks, whose loans were approved in line with a set of norms, were used in this programme. The machine learning models is trained on the data to provide accurate results. Predicting loan safety is the primary goal of this study. With the SVM technique, loan safety may be predicted. The data is first cleaned to remove any missing values from the data collection. The goal of this project is to create a ML model that, given the loan and personal data provided, can predict if a borrower would fail on the loan. The technique of the model is intended to be used by the client and his financial institution as a reference tool to help with loan issuing choices in order to reduce risk and optimize profit.

1.1 Motivation

There are several reasons why predicting loan default is important for lenders, investors, and borrowers. Here are some of the key motivations: Risk Management, Cost Reduction, Investment Decisions, Regulatory Compliance, Social Impact.

For financial companies, the loan approval procedure is crucial. The loan applications were accepted or rejected by the system. Loan recovery is a major determinant of a bank's financial performance. Predicting whether a consumer will pay back a debt is exceedingly tough.

2. LITERATURE SURVEY

Loan default prediction is a crucial task in the banking and finance industry, and it has received significant attention from researchers and practitioners in recent years. Here are some of the relevant studies and literature on loan default prediction:

1."Predicting Credit Card Defaults Using Machine Learning Techniques" by Wei et al. (2009): This study compared different types of ML algorithms, like neural networks and decision tree, for predicting credit card defaults. The authors found that ensemble methods, such as random forests, were the most effective for this task.

2."A Comparative Study of Machine Learning Methods for Loan-Default Prediction" by Brown & Thomas (2011): This study compared different types of ML algorithms, including support vector machines, decision trees and neural networks, for predicting loan defaults. The authors found that gradient boosting and random forests performed the best.

3."Credit Scoring and Loan Default" by Thomas et al. (2016): This book provides an overview of credit scoring and loan default prediction. It covers the traditional statistical methods, such as logistic regression and discriminant analysis, as well as more recent machine learning techniques.

4."Loan Default Prediction Using Bayesian Networks: A Comparative Study" by Azevedo et al. (2019): This study compared Bayesian networks with other ML techniques, like SVM and decision-tree, for predicting loan defaults. The authors found that Bayesian networks outperformed the other methods.

"Predicting Loan Default: An Analysis of Variables and Techniques" by Adeyemo and Adeleke (2021): This study

investigated the factors that contribute to loan default and compared various machine learning algorithms for predicting defaults. The authors found that gradient boosting and random forests performed the best, and that variables such as income, loan amount, and loan term were significant predictors of default.

5."Credit Risk Assessment Using Machine Learning Techniques: A Review" by Sathyadevan et al. (2021): This review article covers various machine learning techniques for credit risk assessment, including loan default prediction. The authors discuss the strengths and weaknesses of different methods and provide recommendations for future research.

Overall, this research indicates that ensemble approaches like gradient boosting technique and random forest approach, as well as machine learning techniques, are useful for predicting loan failure. Moreover, key indicators of default include things like income, loan size, and loan length.

3. IMPLEMENTATION

- Data Collection and Preparation: Collect relevant data from various sources such as credit bureaus, bank statements, loan applications, etc. The data should include both demographic and financial information about the borrower, such as age, income, credit history, employment status, loan amount, interest rate, loan term, etc. Clean and preprocess the data to remove missing values, outliers, and other errors that may affect the accuracy of the model.

- Feature engineering: The practice of adding new features to the data that currently exist in order to improve the model's accuracy is known as feature engineering. For instance, given the borrower's financial data, one may calculate metrics like the loan-to-value, debt-to-income, payment history, and credit utilization ratio. With feature selection, the most important features for the prediction model may also be discovered.

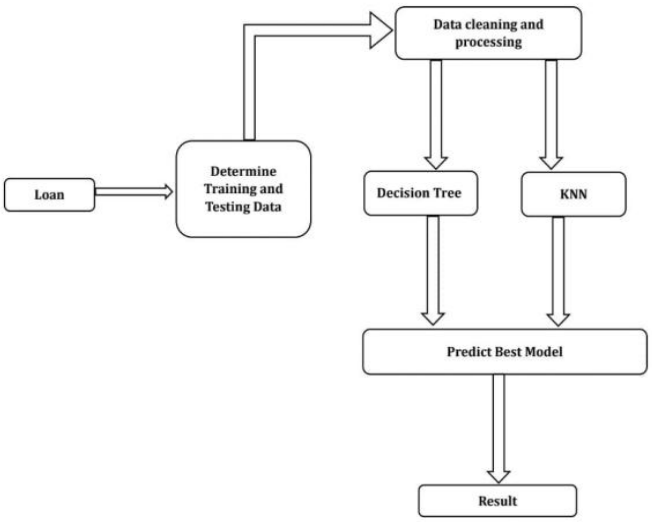
- Model Selection: Based on the available data and the issue at hand, select a suitable machine learning algorithm. Many techniques, including as decision-tree, gradient boosting technique, random - forest, neural networks & bayesian networks, are frequently used to predict loan failure. Both categorical and continuous features should be able to be handled by the algorithm.

- Model Validation: To assess the effectiveness of the ML-model, the datasets will be divided into training datasets and testing datasets. When the model has been created using the training set, its performance is assessed using the testing set. To evaluate the model's effectiveness, utilize appropriate methods such as accuracy, precision, F1 score, ROC-AUC and recall.

- Model Deployment: The model may be put into use in a production setting after it has been trained and verified. This entails making sure the model is scalable and stable as well as incorporating it into the bank's loan processing system.

- Model Monitoring and Maintenance: It's crucial to keep track of the model's performance over time and update it as required. Updating can entail updating the model's data or changing its features or methods.

4. SYSTEM ARCHITECTURE



5. EXISTING MODEL

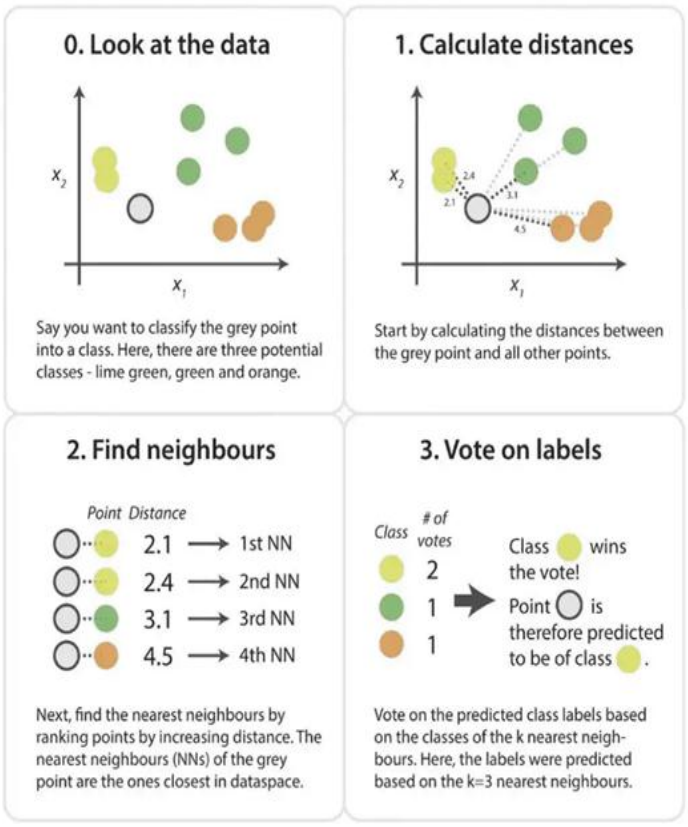
A popular machine learning approach for classification and also for regression applications is the decision-tree. The method divides the data recursively into subgroups depending on the most useful attributes until a halting requirement is satisfied. One benefit of decision trees is that they can handle both category and numerical data. Another benefit is that they are simple to understand and display. Nevertheless, if the tree is too complicated or the data is noisy, they may potentially experience overfitting and instability.

In decision-tree analysis, the values of competing choice are estimated like visual, analytical decision assistance tool using a decision-tree approach and the corresponding related impact diagram.

Operations research and management regularly use decision trees. when choices must be made online with little to no understanding beforehand, the optimal decision - tree model is to use a probability model. or online selection model. Moreover, conditional probabilities may be calculated using decision trees in a descriptive way.

6. PROPOSED MODEL

Both classification and regression are carried out using the supervised learning method known as K-nearest neighbours. KNN model predicts the class for the test data by calculating the separation of the test datasets & the training points. Then the 'K' points that most closely reflect the test datasets should be chosen. When the K-NN algorithm splits the datasets of test into one of the "K" training data class, the highest probability class will be chosen. The training points' average with the stated value of "K" is the value in a regression scenario.



Consider the case when we have a photo of a creature that resembles both cats and dogs but we are unsure of which it is. However, since the K nearest neighbour method is based on a similarity measure, we may use it for this identification. By comparing the new data set's characteristics with those shown in pictures of cats and dogs, our K nearest neighbour model will identify whether the new data set is a cat or dog.

The K-NN operates according to the following algorithm:

- Choose which of the K neighbours you want in step one.
- The second step is to calculate the Euclidean distance between K neighbours.
- In third step, on the basis of Euclidean distance determined, select the K-neighbours that are close.
- In fourth step, among the k-neighbours, count the no. of data points in each category.
- In fifth step, the category with the closest neighbours should receive the additional data points.
- At last step, the model completes the operation.

7. CONCLUSION

To enhance our loan application evaluation process, we plan to implement both decision tree and K-NN models to predict potential loan defaulters. Our objective is to compare and evaluate the performance of both models and determine which one yields more accurate results. By using multiple models, we aim to improve our prediction accuracy and minimize the risk of potential loan defaults. We will thoroughly analyze the results of both models to choose the most suitable one. It correctly predicts whether or not a loan application or customer will be approved. This study will claim that the dataset's prediction accuracy is excellent. When a customer experiences a calamity, for example, the algorithm may not be able to forecast the right outcome. This study can determine if a potential consumer would return a loan, and its accuracy is good. Age, income, loan length, and loan amount are the most crucial variables when determining (whether the client would have been). Zip code and credit history are the two most crucial variables in determining the loan applicant's category.

REFERENCES

[1] Shraddha R. Nikam and Ashwini S. Kadam, - "Prediction for Loan Approval using ML Algorithm," International Research Journal of Engineering and Technology, April 2021.

[2] T.M. Luong, Harald Scheule & Nitya Wanzare, - "Impact of mortgage soft information in loan pricing on default prediction using machine learning," International Review of Finance, September 2022.

[3] Baodong Li, - "Online Loan Default Prediction Model Based on Deep Learning Neural Network," Hindawi Computational Intelligence and Neuroscience, August 2022.

[4] Weidong Chen & Yiheng Li, - "Entropy method of constructing a combined model for improving loan default prediction: A case study in China," Journal of the Operational Research Society, December 2019.

[5] T. Aditya Sai Srinivas, Somula Ramasubbareddy, and K.Govinda, - "Loan Default Prediction Using ML Techniques," Innovations in Computer Science and Engineering, March 2022.

[6] Lili Lai, - "Loan Default Prediction with ML Techniques," International Conference on Computer Communication and Network Security (CCNS), August 2020.

[7] Vishal Singh, Ayushman Yadav & N. Partheeban, - "Prediction of Modernized Loan Approval System Based on ML Approach," International Conference on Intelligent Technologies (CONIT), June 2021.

[8] Dr.C K Gomathy, Ms.Charulatha & Ms.Sowjanya, - "The Loan Prediction using ML," International Research Journal of Engineering and Technology, October 2021.