Loan Repayment Prediction Using Machine Learning Algorithm

Ranjana jatansing vasave.[1]

[1]*Computer engineering,S.S.V.P.S.College of Engineering Dhule*

---***---

Abstract - A loan is money you receive from a bank or financial institution in exchange for a commitment to repay the principal amount with interest. Since lenders take the risk of a possible loan non-payer, they charge a fee to offset this risk this fee is known as the interest. The loans typically are secured or unsecured. In a secured loan, you need to pledge assurance to get the loan. If you are a non-payer of a loan that means you do not pay back the loan, the bank has a right to take possession of the asset that had been pledged as assurance. An unsecured loan doesn't ask for money assurance. If you do not pay back the unsecured loan, the lender has no right to take anything in return. Unsecured loans include credit cards, student loans, and personal loans all of which can be revolving or term loans. Loan clearance is the act of paying back the borrowed money to the lender. The clearance occurs through a series of scheduled payments, also known as EMIs, which include both principal amount and interest. In this project using a machine learning technique, we are going to predict whether the applicant will pay back the loan amount with interest or not, prior to sanction the loan

Key Words: *Decision Tree / KNN / Naïve Bayes*

1.INTRODUCTION

The loan repayment prediction system developed in this project aims to provide a comprehensive solution to the loan default problem. The project uses machine learning techniques to predict the likelihood of loan repayment based on various factors such as income, credit history, loan amount, and loan term.The loan repayment prediction system will be developed using Java and JavaFX technologies. JavaFX is a platform-independent user interface toolkit that enables developers to create rich user interfaces for desktop and mobile applications. Java is a widely used programming language that provides excellent support for developing complex applications.The project will use the loan repayment dataset provided in the Kaggle platform, which contains information on loan repayment, including demographic and financial information of borrowers. The dataset will be pre-processed using various modules such as Exploratory Data Analysis, generating train test, creating machine learning models, comparing machine learning models, and predicting results using the selected model.The loan repayment prediction system will use three machine learning algorithms: Random Forest, Logistic Regression, and SVM. The Random Forest algorithm has an accuracy of 99.2647% and will be used for final prediction with user inputs. The Logistic Regression and SVM algorithms have accuracies of 82.3004% and 82.1954%, respectively. The project aims to provide a comprehensive prediction system for loan repayment and provide meaningful and project-specific content.The next chapter will provide a detailed description of the various modules used in the loan repayment prediction system, including dataset, Exploratory Data Analysis, generating train test, creating machine learning models, comparing machine learning models, and predicting results using the selected model.

2. module of the system

There are various modules in this prediction system:
1. Dataset
2. EDA (Exploratory Data Analysis)
3. Understanding Dataset and Generating train test
4. Generating train test
5. Creating machine learning models
6. Comparing machine learning models
7. Predicting results using the selected model

2.1 Dataset: The dataset used in this loan repayment prediction system is obtained from Kaggle, a popular online platform for data science and machine learning projects. The dataset consists of various attributes related to borrowers such as loan amount, loan term, interest rate, credit score, debt-to-income ratio, and employment status. The dataset also includes information about whether the borrower has repaid the loan or not. This information is used as the target variable for the machine learning models. The dataset is pre-processed and cleaned to remove missing values and irrelevant features.

2.2 EDA (Exploratory Data Analysis): EDA is the process of analyzing and visualizing data to uncover patterns, relationships, and anomalies. In this module, the loan repayment dataset is analyzed using various statistical and visualization techniques to gain insights into the data. The EDA process involves examining the distribution of each feature, identifying correlations between features, and detecting outliers and missing values. The results of the EDA process guide the selection of appropriate machine learning algorithms and help in feature engineering.

2.3 Understanding Dataset and Generating train test: In this module, the dataset is split into two sets: training and testing. The training set is used to train the machine learning models, while the testing set is used to evaluate the performance of the models. Before splitting the dataset, it is important to understand the distribution of the target variable in the dataset. If the dataset is imbalanced, meaning there are significantly more instances of one class than the other, then stratified sampling is used to ensure that the distribution of the target variable is maintained in both the training and testing sets.

2.4 Generating train test: Once the dataset is split into training and testing sets, the next step is to prepare the data for machine learning. This involves scaling the features to ensure that they have similar ranges and applying any necessary feature transformations such as one-hot encoding or label encoding. The pre-processed training and testing data are then used to train and evaluate the performance of the machine learning models.

2.5 Creating machine learning models: In this module, three machine learning algorithms are implemented: Random Forest, Logistic Regression, and SVM. Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve the accuracy and reduce overfitting. Logistic Regression is a linear classification algorithm that models the probability of the target variable as a function of the features. SVM is a non-linear classification algorithm that finds the hyperplane that best separates the instances of different classes. The implementation of each algorithm involves tuning the hyperparameters to optimize the performance on the training set.

2.6 Comparing machine learning models: After training the machine learning models, their performance is compared on the testing set. The performance metrics used to evaluate the models include accuracy, precision, recall, and F1-score. The results of the performance evaluation are used to select the best performing model.

2.7 Predicting results using the selected model: Once the best performing model is selected, it is used to predict the loan repayment status of new borrowers. The user can input the borrower's information, and the model predicts whether the borrower is likely to repay the loan or not. The prediction is based on the learned patterns and relationships in the training data, and the accuracy of the prediction depends on the quality of the training data and the chosen machine learning algorithm.

APPROACH AND METHODOLOGY

Data pre-processing is considered a significant and crucial initial step in data analysis and data mining projects, as the output of this stage is inputted to the model to obtain final results, therefore, data preprocessing impacts not only the accuracy of the model but also the performance and efficiency [9]. Our model involved a thorough exploration of data, and the application of multiple preprocessing techniques prior to the classification stage. Once the data is tuned, three different classification algorithm models will be used to predict loan default, results of models will be compared with each other, and with prediction results of the same model prior to data preprocessing. Figure 1 illustrates our approach.

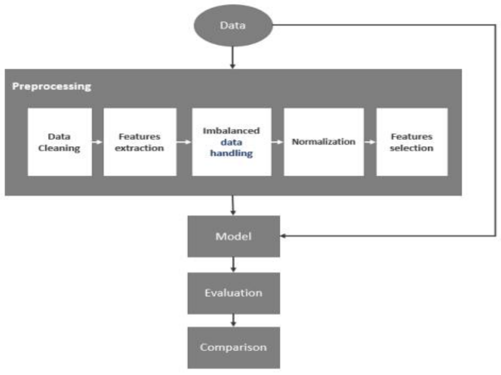

Fig. 1. Approach

Four classification models were built; Naïve Bayes, unpruned C4.5 decision tree, pruned C4.5 decision tree and Random forest. Each one of these models was run four times and evaluation measures were recorded. In the first iteration; models used unprocessed data, and the other three used processed data with three different features selection algorithms. Evaluation measures of used models using unprocessed data set are shown in table 1 below.

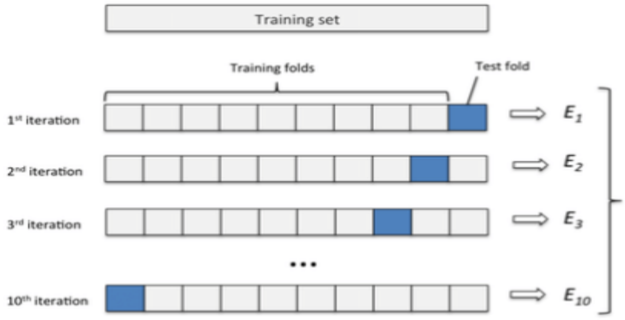

Fig. 2. Cross Validation

TABLE I.

Model	Class	Measures		
		Precision	*Recall*	*F1*
Naïve Bayes	Charged off	87.20%	64.40%	74.10%
	Fully Paid	90.90%	97.40%	94.00%
	Weighted Avg.	90.10%	90.30%	89.70%
C4.5	Charged off	93.50%	89.50%	91.50%
	Fully Paid	97.10%	98.30%	97.7%
	Weighted Avg.	96.40%	96.40%	96.40%
C4.5 (Pruned)	Charged off	99.30%	88.00%	93.30%
	Fully Paid	96.80%	99.80%	98.30%
	Weighted Avg.	97.30%	97.30%	97.20%
Random Forest	Charged off	52.30%	46.80%	49.40%
	Fully Paid	90.70%	92.40%	91.50%
	Weighted Avg.	84.90%	85.50%	85.20%

3. CONCLUSIONS

This paper used Naïve Bayes, Decision tree (unpruned and pruned) and Random Forest classifiers to build loan default prediction models. This paper also applied several data preprocessing techniques, and compared between three features selection algorithms: Information Gain, Genetic Algorithm and Particle Swarm Optimization. Applying preprocessing techniques definitely improved the prediction of the minority class. Improvement varied between the different classifiers. Using features selection algorithms improved model as well, though improvement variation between the three used algorithms was not remarkable. It can be concluded that the data preprocessing stage is an important stage when building a classification model, as it has a valuable impact on model accuracy. Applying features selection algorithms is very significant as well when having a large dataset, not only it enhances accuracy but it also improves performance. Future work would involve other classifiers and features selection algorithms, as well as using datasets from different banks to investigate if our findings apply on datasets with different natures.

REFERENCES

[1] Gaurav Akrani., Kaylan City Life (20-Apr-2011), Available: http://kalyan-city.blogspot.com/2011/04/functions-of-banks-importantbanking.html. [Accessed: 1- Jan- 2019]

[2] Businessmodelinnovationmatters (24-Apr-2012), Available: https://businessmodelinnovationmatters.wordpress.com/2012/03/24/und erstanding-the-business-model-of-a-bank/.[Accessed: 1- Jan- 2019]

[3] E. Angelini, A. Roli, and G. di Tollo, "A neural network approach for credit risk evaluation", The Quarterly Review of Economics and Finance, vol. 48, pp. 733–755, 2008.

[4] Chun F. Hsu and H. F. Hung, "Classification Methods of Credit Rating - A Comparative Analysis on SVM, MDA and RST", International Conference on Computational Intelligence and Software Engineering, pp. 1–4, 2009.

[5] Amira Hassan and Ajith Abraham, "Modeling Consumer Loan Default Prediction Using Ensemble Neural Networks", International Coference on Computing, Electrical and Electronic Engineering (ICCEEE), pp. 719 – 724, 2013.

[6] M.V. Jagannatha Reddy and B. Kavitha, "Neural Networks for Prediction of Loan Default Using Attribute Relevance Analysis", International Conference on Signal Acquisition and Processing, pp. 274 – 277, 2010.

[7] Yu Jin and Yudan Zhu, "A Data-Driven Approach to Predict Default Risk of Loan for Online Peer-to-Peer (P2P) Lending", Fifth International Conference on Communication Systems and Network Technologies, pp. 609 – 613, 2015.

[8] Archana Gahlaut, Tushar and Prince Kumar Singh, "Prediction analysis of risky credit using Data mining classification models", 28th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-7, 2017.

[9] Li Xiang-wei and Qi Yian-fang, "A Data Preprocessing Algorithm for Classification Model Based On Rough Sets", 2012 International

[10] Kalyan Netti and Y Radhika, "A novel method for minimizing loss of accuracy in Naive Bayes classifier", IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-4, 2015.

[11] Z. Xiaoliang et al., "Research and Application ofthe improved Algorithm C4.5 on Decision Tree", International Conference on Test and Measurement, pp. 184 – 187, 2009.

[12] Afsaneh Mahanipour and Hossein Nezamabadi-pour, "Improved PSObased feature construction algorithm using Feature Selection Methods", 2nd Conference on Swarm Intelligence and Evolutionary Computation (CSIEC), pp. 1-5, 2017.

[13] Raul Eulogio, ORACLE + Data Science (2017, Aug, 12), Available: https://www.datascience.com/resources/notebooks/random-forest-intro. [Accessed: 2- Jan- 2019]

[14] M. Bentlemsan et al., "Random Forest and Filter Bank Common Spatial Patterns for EEG-Based Motor Imagery Classification", th International Conference on Intelligent Systems, Modelling and Simulation, pp. 235 – 238, 2014.

[15] Chioka (2013, Aug, 30), Available: http://www.chioka.in/classimbalance-problem/. [Accessed: 2- Jan- 2019]

[16] Hong Zhang, Yong-gong Ren and Xue Yang, "Research on Text Feature Selection Algorithm Based on Information Gain and Feature Relation Tree", 10th Web Information System and Application Conference, pp. 446 – 449, 2013.

[17] Ho-duck Kim et al., "Genetic Algorithm Based Feature Selection Method Development for Pattern Recognition", SICE-ICASE International Joint Conference, pp. 1020 – 1025, 2006.