Predicting Credit Card Defaults with Machine Learning Algorithms

Hansraj Lawte[1], Laxman Pawar[2], Prathmesh Bondar[3], Gopal Malache[4], Shivshankar Chandsure[5]

[1]Student, Department of Computer Engineering, Sinhgad College of Engineering, Pune, Maharashtra, India, 411041.
[2]Assistanat Professor, Department of Computer Engineering, Sinhgad College of Engineering, Pune, Maharashtra, India, 411041.
[3]Student, Department of Computer Engineering, Sinhgad College of Engineering, Pune, Maharashtra, India, 411041.
[4]Student, Department of Computer Engineering, Sinhgad College of Engineering, Pune, Maharashtra, India, 411041.
[5]Student, Department of Computer Engineering, Sinhgad College of Engineering, Pune, Maharashtra, India, 411041.

Contributing authors: hansrajlawte@gmail.com; lbpawar.scoe@sinhgad.edu; prathmeshbondar@gmail.com; malachegopal9835@gmail.com; shivshankarchandsure001@gmail.com;

Abstract

With credit systems playing an increasingly important role in the worldwide economy, the correct anticipation of credit card defaults is now a must for not only financial institutions that want to limit their risks but also those who aim at creating sustainable financial solutions. The main objective of the research paper is to find a machine learning technique that can help banks recognize those clients that are likely to be a burden by not paying their card debits. The research dwells on one dataset that contains different customer attributes, such as demographic details, payment history, and credit usage behavior, thus allows the researchers to spot any patterns that may lead to a default. The authors choose to test the performance of Logistic Regression, Naive Bayes Classifier, and Random Forest, and evaluate them using accuracy, precision, recall, and F1-score that are the main metrics of the comparative model. The evaluation features a discussion of each of the models' predictive power and also the reasons why they are good or bad, and the fact that they can be put into practice in real-world credit risk assessment. The findings certify that the application of machine learning techniques is highly beneficial considering that they can offer reliable risk forecasts. This paper not only presents the merits of using data- based techniques in credit scoring but also leads the way for more studies that would focus on the incorporation of more updated and advanced algorithms and real-time data in financial forecasting.

Keywords: Logistic Regression, Naive Bayes, Random Forest, Credit Card, Visualization, Prediction, Accuracy

1 Introduction

Credit card defaults cause serious problems for individuals and financial institutions. Failure of card holders to make payments on ti me both creates a financial burden for the individual and poses a great risk for the institution that issues the credit card. The ability to accurately predict the structure of credit cards is crucial to reducing these risks and making informed decisions.This re search paper focuses on using machine learning techniques to predict credit card presets. Using historical data capturing var- ious asp ects of credit card usage and payment behavior, machine learning algorithms can identify patterns and patterns to detect criminals. F orecasting capability allows financial institutions to manage credit risk, improve collection strategies and adjust credit products. The main purpose of this research is to develop a more efficient credit card credit card prediction system using machine learning alg orithms. By leveraging the power of machine learning, historical patterns and characteristics associated with illegal individuals can be used to create accurate predictive models. These models can be used as decision support tools, provide insight to credit card issuers, and help man- age risk.

Machine learning algorithms can detect the relatio nship between various features and how to default them. They can detect nonlinear patterns and interactions that may not be apparen t with traditional statistical methods. Additionally, machine learning models can handle large amounts of data, making them highly c apable and pow- erful at predicting predetermined values.This research article on machine learning techniques aims to contribute to th e existing body of knowledge in credit risk assess- ment. This study aims to understand the effectiveness of credit card default casino strategies by developing and evaluating predictive models based on machine learn- ing algorithms. Additionally, the study explores d ata analysis techniques to gain a deeper understanding of the dataset, uncover risk factors, and improve understand- ing of credit card defaults.In summary, this research paper aims to predict credit card default using machine learning technology. The

research attempt s to create accurate predictive models using historic credit card data and machine learning algorithms. Through evaluation of these models and exploration of research data, this research is designed to provide a better understanding of credit card transactions that can help improve risk management and know how to make decisions in financial markets.

2 Literature Survey

Many researches have studied the prediction of credit card defaults through differ- ent machine learning and data mining techniques. These researches mentioned that choosing the right algorithms, solving the data imbalance problem, and using efficient feature engineering are key factors in the improvement of prediction accuracy.

Abdulhamit Subasi and Selcuk Cankurt, (2019) used data mining techniques for the prediction of credit card defaults and noticed that classification-based models were instrumental in increasing the risk prediction performance in financial services. The results of the study described the practical performance of different models like decision trees and support vector machines in real scenarios.

Pu Xu, Zhijun Ding, and MeiQin Pan, (2017) developed an improved model in which they have adapted the RIPPER rule-based classifier for credit default prediction. Their work was only focused on the accessibility and speed of the rule-based method and was able to present RIPPER as a high performer even with less sophisticated models while still keeping the decision-making process clear and easy to understand.

Dr. E. Praynlin, Madesh S, Mohammed Thafeez H, Venu K V, and Vinod Kumar K, (2023) looked into default prediction through a variety of machine learning models. While the research was more on loan default, the techniques and findings could also be used in the prediction of credit card defaults. The authors expressed that ensemble models, which used Random Forest as their core, were able to deliver high accuracy and reliability in contrast to traditional methods.

Shreyas Khandale, Prathmesh Patil, and Rohan Patil, (2023) illustrated a very direct approach to the topic of credit card default prediction through the use of various supervised learning techniques. Their work demonstrated which techniques, namely Logistic Regression, Decision Trees, and SVMs, are to be used. One of the key issues that the study identified was applying the preprocessing techniques like scaling, encod- ing, and balancing the dataset, which are very necessary for the accurate performance of the models.

Talha Mahboob Alam, Kamran Shaukat, Muhammad Umer Sarwar, Shakir Shab- bir, Jiaming Li, and Matloob Khushi, (2020) have researched the obstacles in working with imbalanced datasets within the sphere of credit card default prediction. They conducted a comparison of the success of resampling techniques like SMOTE and undersampling with the objective of improving the model performance. Their results highlighted not only the role of recall and precision metrics, but also the need for these in addition to accuracy, especially for issues with imbalanced classes.

3 Methodology

3.0.1 Data Collection and Preprocessing

The dataset was obtained from Kaggle, which collects and maintains credit card trans- action data from a various group of cardholders. The data collection process ensures

the anonymity and privacy of the individuals involved. The dataset provides a repre- sentative sample of credit card users and covers a substantial time period, enabling the analysis of long-term payment behaviors. Data preprocessing is a crucial step in prepar- ing the dataset for analysis. It involves handling missing values, outlier detection, feature scaling, and ensuring data integrity.

3.0.2 Data Cleaning

The dataset may contain missing values, finding duplicates which need to be addressed before further analysis.

3.0.3 Feature Engineering

Feature engineering plays a important role in credit card default prediction. It contains transforming and creating new features that capture meaningful information from the existing variables. For example, derived features such as credit utilization ratio, payment to-income ratio can provide insights into cardholders' financial health and payment behavior. Feature engineering may also involve encoding categorical variables, or creating interaction terms to capture non-linear relationships.

3.0.4 Model Building

After the data preprocessing steps completed, the next step is to build predictive models for credit card default prediction. In this research, three models are involved: Logistic Regression, Naive Bayes Classifier and Random Forest.

a) Logistic Regression: Logistic regression create the relationship between input features and the probability of default. It is a widely used algorithm for binary clas- sification tasks and provides interpretability. Logistic regression can capture linear relationships between features and defaults but may struggle to capture non-linear relationships. The model is trained using the preprocessed dataset, in which the target variable is the

default.payment.next.month column.

b) Naive Bayes Classifier: The Naive Bayes classifier is based on probabilistic prin- ciples and assumes that features are conditionally independent given the class variable. It is known for its simplicity, scalability, and efficiency.

c) Random Forest : Random Forest makes the multiple decision trees and combines their outputs and reduces overfitting, which results in improves the accuracy.

3.0.5 Model Evaluation

After training the machine learning models, model evaluation is the very important step it need to be evaluated to assess their performance in predicting credit card defaults. An evaluation metrics such as accuracy, precision, recall, and F1 score can be used to measure the models effectiveness.

3.0.6 Model Comparison and Selection

The performance of the Logistic Regression, Naive Bayes, Random Forest models is compared based on their evaluation metrics. The model with the highest accuracy or

the most suitable evaluation metric for the specific research goal is selected as the primary model for credit card default prediction.

3.0.7 Model Deployment and Interpretation

Once the primary model is selected, it can be deployed to predict credit card defaults on new, unseen data. The model can be used to provide insights and make informed decisions related to credit risk management. The coefficients or feature importance val- ues from the selected model can be interpreted to understand the relative importance of different features in predicting credit card defaults. Furthermore, the sensitivity of credit card default prediction models necessitates ensuring the privacy and security of the dataset. Steps should be taken to anonymize and protect sensitive information to comply with data protection regulations and ethical guidelines.

a) By carefully collecting and preprocessing the data, addressing missing values, detecting outliers, integrating relevant external data, conducting feature engineer- ing, and considering specific challenges in credit card default prediction, the dataset becomes suitable for building accurate and robust predictive models.

b) Data Sources

In addition to the comprehensive dataset used for credit card default prediction, this research paper also leverages additional point data sources to enhance the pre- dictive capabilities of the models. Point data sources refer to specific types of data sets or information obtained from external sources that provide valuable insights into creditworthiness and default prediction. The integration of these point data sources complements the existing data set and provides a more holistic view of the financial profiles of individuals.

There are various types of point data sources that can be utilized in credit card default prediction research. Some of the common types include:

- Credit Bureau Data
  Credit bureau data is a crucial source of information for assessing creditworthiness. It includes credit scores, credit histories, payment delinquency records, and other credit- related information maintained by credit bureaus. By incorporating credit bureau data, the models can capture the historical payment behavior and overall creditworthiness of the cardholders.
- Economic Indicators
  Economic indicators provide insights into the broader economic conditions and can impact credit card defaults. Examples of economic indicators include GDP growth rates, inflation rates, unemployment rates, interest rates, and consumer confidence indices. These indicators can provide contextual information about the economic environment and its potential influence on credit card defaults.
- Industry-Specific Data
  Industry-specific data can be valuable for predicting credit card defaults, especially when considering the stability and risk factors associated with different indus- tries. This type of data includes industry performance metrics, regulatory changes,

  market trends, and financial indicators specific to certain sectors. Incorporating industry-specific data allows for a more nuanced assessment of creditworthiness within different sectors.

- Demographic Data
  Demographic data encompasses information related to individuals' characteristics, such as age, gender, marital status, education level, and employment status. Demo- graphic factors can provide valuable insights into credit card usage patterns and payment behaviors. Incorporating demographic data helps in capturing the socio- economic context of the cardholders and can contribute to more accurate default predictions.
- Publicly Available Financial Data
  Publicly available financial data sources, such as financial statements of companies or public records of bankruptcies, can be leveraged to gain insights into the financial health and stability of individuals and

businesses. These data sources can provide valuable indicators of creditworthiness and default risk. Integrating and analyzing these types of point data sources alongside the primary dataset contributes to a more comprehensive understanding of credit card default risks. The combined analysis enables a more accurate assessment of creditworthiness, improved predictive models, and better-informed decision-making in credit risk management.

- **3.0.8 Project Analysis**

  - Dataset Table
    The dataset is printed in a tabular format using data.head() function.

  - Dataset Description
    The dataset used for this research paper consists of credit card transaction and payment information for a sample of credit card users.
    The dataset includes the following columns:
    article [utf8]inputenc enumitem

    – **ID**: Unique identifier for each credit card user.
    – **LIMIT BAL**: The credit limit assigned to the user's credit card account.
    – **SEX**: Gender of the user

      * 1 for male
      * 2 for female

    – **EDUCATION**: Education level of the user

      * 1: Graduate school
      * 2: University
      * 3: High school
      * 4: Others

    – **MARRIAGE**: Marital status of the user

      * 1: Married
      * 2: Single
      * 3: Others

    – **AGE**: Age of the user in years.
    – **PAY 0 to PAY 6**: Payment status for the past six months (one column per month), with values:

      * -2: No consumption
      * -1: Paid in full
      * 0: Use of revolving credit
      * 1: Payment delay for one month
      * 2: Payment delay for two months

    – **BILL AMT1 to BILL AMT6**: Amount of bill statement for the past six months.
    – **PAY AMT1 to PAY AMT6**: Amount of previous payment made for the past six months.
    – **default.payment.next.month**: Binary indicator of whether the user defaulted on the credit card payment in the following month

      * 1: Default
      * 0: No default

  - Dataset Challenges and Considerations

When working with this type of dataset for credit card default prediction, several challenges faced:
Imbalanced Classes:
The dataset may suffer from class imbalance, where the number of non-default instances (0) significantly outweighs the number of default instances. This imbalance can affect the model's performance and lead to biased predictions. Proper handling of class imbalance is necessary to ensure accurate and reliable predictions.

Missing Values:

The dataset may contain missing values in certain columns, which need to be addressed during the preprocessing stage. Missing values can be imputed using appropriate techniques, such as mean imputation or more sophisticated methods like regression imputation, to avoid any bias in the analysis.

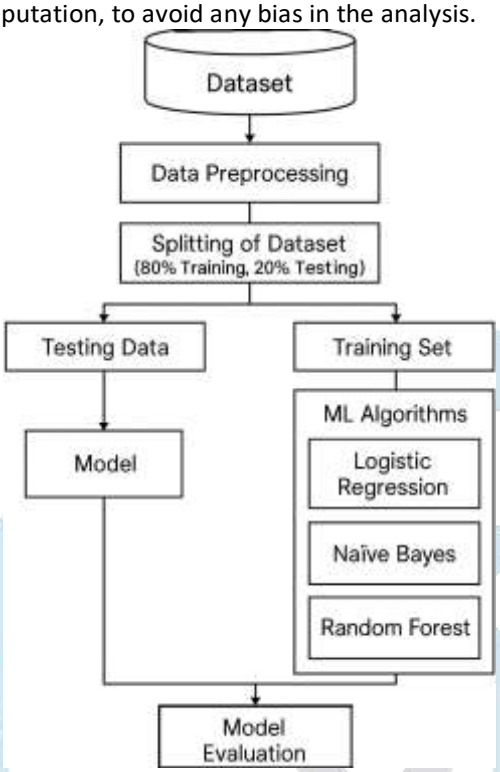

Fig. 1 System Architecture

Outliers:

Outliers in the dataset can significantly impact the model's performance. Identifica- tion and appropriate handling of outliers are essential to ensure robust and reliable predictions. Outliers can be detected using statistical methods or domain knowledge and can be treated by either removing them, transforming them, or using robust models that are less sensitive to outliers.

Feature Engineering:

The dataset offers opportunities for feature engineering to enhance the predictive power of the models. Feature engineering involves creating new features or transform- ing existing ones to capture additional information or patterns that may improve the models' performance. Techniques such as creating interaction terms, polynomial features, or aggregating features over time can be explored.

By using this dataset, we can employ various machine learning algorithms to develop predictive models for credit card default prediction. . The following steps outline the approach for model building and evaluation:

- Bar Plot Of Default Payment Counts

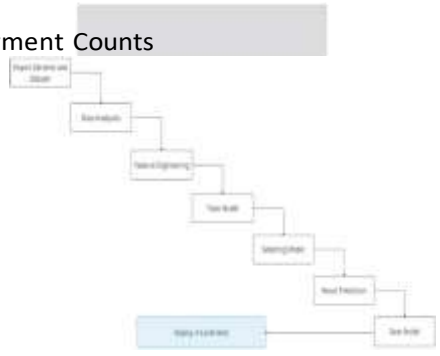

Fig. 2 Use Case

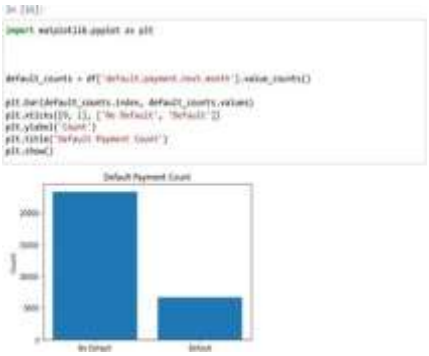

Fig. 3 Default Payment Count

A bar plot is generated to visualize the count of default payments. This plot provides a clear understanding of the distribution of default and non-default instances in the dataset. By examining the bar heights, we can identify any class imbalance issues that may affect model performance.

- Scatter plot of bill amount vs. payment amount

The scatter plot displays the relationship between bill amounts and payment amounts. By plotting bill amounts on the x-axis and payment amounts on the y-axis, we can observe the patterns and associations between these two variables. This visualization helps identify any trends or correlations between bill amounts and payment amounts. It provides insights into the payment behavior of credit card users, such as whether higher bill amounts are associated with higher payment amounts.

- Histogram of credit limits

The histogram illustrates the distribution of credit limits among credit card users. It provides a visual representation of the frequency of different credit limit ranges. The x-axis represents the credit limit ranges, and the y-axis represents the frequency or count of users falling within each range. This histogram helps in understanding the distribution of credit limits and identifying any concentration of users within specific credit limit ranges.

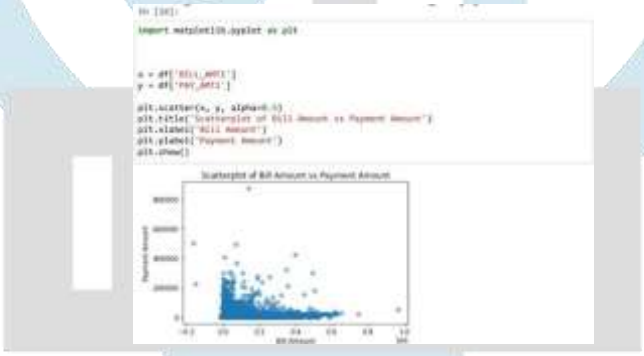

Fig. 4 Scatter Plot

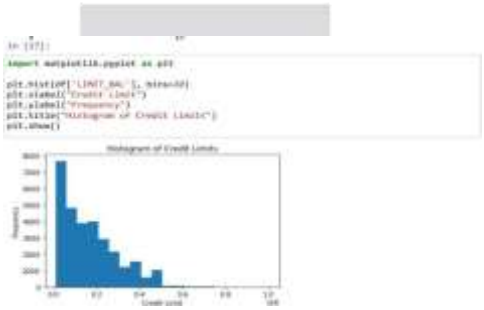

Fig. 5 Histogram of Credit Limits

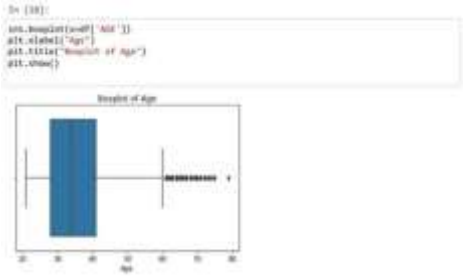

Fig. 6 Boxplot of Age

- Boxplot of age distribution:

The boxplot of the age distribution provides valuable insights into the characteristics of credit card users in terms of their ages. It visualizes the statistical measures such as the minimum, first quartile (25th percentile), median (50th percentile), third quartile (75th percentile), and maximum values of the age variable.

- Countplot of sex distribution

Fig. 7 Countplot of Sex Distribution

The countplot is an effective visualization technique used to analyze the distribution of credit card users based on their gender (male or female). It provides a clear and concise representation of the number of users in each gender category. In the count- plot, the x-axis represents the gender categories (male and female), while the y-axis represents the count of users belonging to each category. By examining the countplot, we can easily identify the gender distribution among credit card users. It allows us to compare the number of male and female users and assess any gender-based patterns or discrepancies that may exist in the dataset. The countplot helps answer questions such as whether the dataset contains an equal representation of male and female users or if there is an imbalance in the gender distribution. Analyzing the countplot can reveal insights into the demographic composition of there are any significant differences in credit card default rates between male and female users. The countplot is a simple yet powerful visualization that provides a visual overview of the gender distribution in the dataset. By including this countplot in the research paper, readers can easily grasp the gender composition of credit card users and gain insights into any gender-based patterns or discrepancies that may be relevant to credit card default prediction. This visualization adds value to the descriptive analysis of the dataset and contributes to a comprehensive understanding of the gender dynamics within the context of credit card defaults.

These visualizations provide valuable insights into the dataset and contribute to the exploratory analysis of credit card default prediction. They help uncover patterns, relationships, and distributions of various variables, enabling researchers and practi- tioners to make informed decisions and gain a deeper understanding of the dataset. By including these visualizations in the research paper, readers can visualize and interpret the data more effectively.

4 Results

The Random Forest model achieved higher accuracy compared to Logistic Regression and Naive Bayes, it gives the better performs in predicting credit card defaults.

5 Conclusion

The discovery of credit card defaults is an important field of research. This is because fraud among financial institutions is increasing. This problem opens the door to using artificial intelligence to create systems that can detect fraud. Creating an AI-based

system to detect defaults requires data to train the system. Real life data is dirty with missing results, noisy data, and outliers. These issues can negatively impact the accuracy of the system. To overcome these problems, a classification based on logistic regression has been proposed. The Random Forest model is significantly better than Logistic Regression model and Naive Bayes in detecting credit card defaults.

6 Acknowledgements

The authors thank Prof. Laxman Pawar for providing excellent guidance in carry- ing out the work. We also thank Prof. Pravin Kamade and Prof. Archana Dirgule, professors of Sinhgad College of Engineering for providing their valuable feedback.

7 References

1. Abdulhamit Subasi and Selcuk Cankurt. *Prediction of default payment of credit card clients using Data Mining Techniques*. International Engineering Conference (IEC), IEEE, pp. 115–120, 2019.
2. Pu Xu, Zhijun Ding, and MeiQin Pan. *An improved credit card users default prediction model based on RIPPER*. In: 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 1785–1789, 2017. https://doi.org/10.1109/FSKD.2017.8393037
3. Dr. E. Praynlin, Madesh S, Mohammed Thafeez H, Venu K V, and Vinod Kumar K. *Loan Default Prediction Using Machine Learning Techniques*. International Research Journal of Engineering and Technology (IRJET), pp. 64–67, 2023.

4. Shreyas Khandale, Prathmesh Patil, and Rohan Patil. *Predicting Credit Card Defaults with Machine Learning*. IJRASET, 2023.
5. Talha Mahboob Alam, Kamran Shaukat, Muhammad Umer Sarwar, Shakir Shab- bir, Jiaming Li, and Matloob Khushi. *An Investigation of Credit Card Default Prediction in the Imbalanced Datasets*. IEEE, pp. 201173–201198, 2020.
6. Baesens, B., Roesch, D., Scheule, H., Stepanova, M. (2017). Credit risk analytics: Measurement techniques, applications, and examples in SAS. John Wiley Sons
7. Hasan, M., Ng, A., Wu, Q. (2017). Credit risk prediction using machine learning techniques: A literature review. Intelligent Systems in Accounting, Finance and Management, 24(2), 59-82.
8. Lin, C., Lee, C., Chen, C. (2019). Credit scoring using a hybrid machine learning approach based on rough set theory and weighted k-nearest neighbors. IEEE Access, 7, 18406-18415.
9. Naeem, M., Khan, S., Khiyal, M. S. (2020). Machine learning techniques for credit card fraud detection: A systematic review. Journal of Ambient Intelligence and Humanized Computing, 11(11), 5425-5450.
10. Shraddha R. Nikam and Ashwini S. Kadam, - "Prediction for Loan Approval using ML Algorithm," International Research Journal of Engineering and Technology, April. 2021.
11. Vishal Singh, Ayushman Yadav N. Partheeban, - "Prediction of Modernized Loan Approval System Based on ML Approach," International Conference on Intelligent Technologies (CONIT), June 2021.