CREDIT CARD DEFAULT PREDICTION

NAGA ADITHYA || ROHITH REDDY || PRASHANTH

1 Motivation

With technological growth and increase in potential threats to every domain, global markets are increasingly reliant on advanced analytical tools to mitigate them. Making informed decisions, assessing financial risks, promoting ethical lending practices, risk management, and protecting consumers are some important factors that determine a financial institution's success. The motivation for this project stems from the pressing need to mitigate the threat of credit card defaults, safeguarding the credit issuers' interests. In this project, we aim to develop a robust predictive model that accurately tracks credit card defaults and aid in better risk management, fostering a trustworthy relation between consumers and the institution.

2 Problem Statement

Global Financial institutions constantly fear the risk of defaults, which leads to losses to the institutions and disturbs the stability of financial health. The Capability to deal with such a complex situation remains to be a challenging task. The continuous change in behavioral and economic factors of consumers poses a complex challenge to predict the defaults accurately. The need for this project emphasizes the necessity of such machine learning based techniques to address the above challenges. Through this project we employ supervised machine learning models to create a robust predictive model to anticipate credit defaults and promote healthy financial practices.

3 Technical background

3.1 Python Libraries

3.1.1 Numpy

The cornerstone for numerical computations and data manipulation, enabling seamless handling of arrays and matrices crucial for machine learning tasks.

3.1.2 Pandas

The versatile Python library for data manipulation and analysis, streamlining tasks like loading, cleaning, and transforming data into organized Data Frames for efficient manipulation.

3.1.3 Scikit-Learn

The go-to Python library for machine learning tasks, offering a rich suite of algorithms, evaluation tools, and preprocessing methods.

3.1.4 Matplotlib

Enhancing the data analysis process with Python, it crafts static, animated, or interactive visualizations, refining data representations for insightful analysis.

3.1.5 Seaborn

Elevating data visualization with a user-friendly interface atop Matplotlib, crafting visually appealing statistical graphics effortlessly.

3.2 Machine Learning Teqniques

3.2.1 Models

This project employs supervised machine learning by training a particular model on a dataset with labels. The inputs to this model are various features of credit transactions and the output is the likelihood estimate of default. Till now we employed Logistic regression and K-Nearest Neighbors algorithms.

3.2.2 Evalution Metrics

A range of evaluation metrics are used to measure the effectiveness of our models. Accuracy, recall, precision, F1-score, and ROC curve are the evaluation metrics used.

4 Related Work

The area of credit card default prediction has many technological advancements in the field using different statistical and machine learning methods. Various models like linear regression were used initially to predict defaults, later more advanced models were developed. Sayjadah et al. studied the prediction of defaults using models like random forests, decision tree and evaluated the performance of the above models in his paper. Sijie Xu in his paper used AutoML framework, used stacking methods and evaluated the models using F1 values. His research found that the F1 values of an integrated model are better than that of a single algorithm. These studies compare the strengths and limitations of different algorithms incorporating recent advancements like feature engineering, aiding us to pick a suitable model to pick. Shamroz Qureshi, uses basic random tree and random forests following by hyperparameter tuning and performance is compred using recall. Jing Gao in his paper employs XGBoost model. Wang et al. present innovative data preprocessing methods, particularly Synthetic Minority Oversampling Technique (SMOTE), to tackle imbalanced datasets, leading to a notable enhancement in predictive accuracy. All these studies help us in picking a perfect model with all the technological advancements. We use this information from these studies to develop a robust model to predict defaults accurately.

5 Current Progress

5.1 Dataset discription and preprocessing

The dataset used for this project, sourced from Kaggle's datasets, provides a comprehensive view of credit card customers and their financial behaviors. It includes demographic information such as sex, education, marriage status, and age, which can be crucial factors in predicting credit default. Additionally, the dataset contains nineteen attributes related to credit availability, usage patterns, and repayment history over a six-month period from April to September 2005. Understanding these attributes can offer insights into customer behavior and credit risk.

One particular attribute worth noting is the repayment history, which includes information on the amount repaid by each customer and the time taken to repay. Initially described as ranging from -1 to 9, the actual data falls within the range of -2 to 8. To ensure consistency and proper interpretation of the data, a correction is applied by adding 1 to align the values with the description provided.

In preparation for model training, the dataset is split into a 70-30 ratio for training and testing, respectively. This split allows for the model to learn patterns and relationships from the majority of the data while retaining a portion for evaluation. This approach ensures that the model's performance can be accurately assessed on unseen data, helping to gauge its effectiveness in predicting credit default.

5.2 Model implementation

This project employs supervised machine learning, where a model is trained on a labeled dataset to predict the likelihood of default based on various features of credit transactions. Initially, the project used the Logistic Regression and K-Nearest Neighbors algorithms without any modifications to establish a baseline estimate of model performance. While the accuracy scores of these models

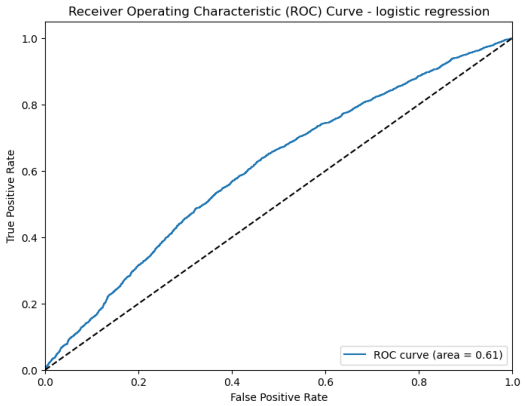
were promising, indicating a generally correct prediction of default and non-default cases, the F1 scores revealed a significant issue.

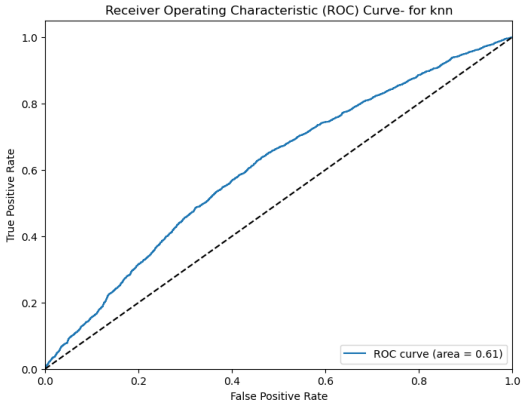
The F1 scores, which consider both precision and recall, were low. This suggests that the models were not performing well in identifying default cases. The primary reason for this discrepancy is the imbalance in the dataset, where the number of default cases is much lower than the number of non-default cases. As a result, the models were biased towards predicting non-default cases, leading to a high number of false negatives (predicting a non-default when the actual case was default).

To address this issue and improve the models' ability to predict default cases accurately, the project will explore sampling techniques. These techniques aim to balance the dataset by either oversampling the minority class (default cases) or undersampling the majority class (non-default cases) to create a more balanced training dataset. By doing so, the models can learn more effectively from the data and hopefully improve their performance in predicting default cases.

Model	F1 Score	Accuracy
Logistic Regression	0.0	0.782
K-Nearest Neighbors	0.243	0.754

Table 1: F1 Score and Accuracy Scores for Models





5.3 Feature Selection

We utilized the correlation matrix to select features with the strongest correlation to the default status, focusing on payment details (e.g., amount paid, payment delays), credit availability, and age demographics. These features were chosen for their potential to provide valuable insights into credit default likelihood, aiming to enhance the efficiency and accuracy of our predictive model.

Model	F1 Score	Accuracy
Logistic Regression	0.0	0.782
K-Nearest Neighbors	0.255	0.745

Table 2: F1 Score and Accuracy Scores for Models

5.4 Challange Faced

Feature selection did not yield the expected improvements in model accuracy or F1 scores. However, the reduced number of features has made the models more efficient.

6 Plans Ahead

The first step in improving our models is to find a more effective way to extract features that can boost performance beyond our current validation scores. Once we have optimized feature extraction, we plan to expand our analysis by introducing deep neural networks and support vector machines (SVM). By comparing the performance of these models, we aim to identify the most effective approach for predicting credit default. This comprehensive approach will allow us to leverage advanced techniques and methodologies to enhance the accuracy and reliability of our predictions.

7 References

1. Y. Sayjadah, I. A. T. Hashem, F. Alotaibi and K. A. Kasmiran, "Credit Card Default Prediction using Machine Learning Techniques," 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), Subang Jaya, Malaysia, 2018, pp. 1-4, doi: 10.1109/ICACCAF.2018.8776802. keywords: Credit cards;Banking;Decision trees;Machine learning;Predictive models;Machine learning algorithms;Credit Score;Data mining;Machine Learning;Banking
2. Sijie Xu, Peixin Lin, Wanqi Luo, Wenjun Yang, Yuntao Jia, "A study of machine learning based credit card potential default customer identification," Proc. SPIE 12599, Second International Conference on Digital Society and Intelligent Systems (DSInS 2022), 1259908 (3 April 2023)
3. W. Lee, S. Lee and J. Seok, "Credit card default prediction by using Heterogeneous Ensemble," 2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN), Paris, France, 2023, pp. 907-910, doi: 10.1109/ICUFN57995.2023.10199756. keywords: Deep learning;Machine learning algorithms;Companies;Predictive models;Credit cards;Prediction algorithms;Boosting;Gradient Boosting;TabNet;Machine Learning;Deep Learning;Ensemble Learning
4. https://nycdatascience.com/blog/student-works/data-analysis-on-credit-card-default-detection/
5. https://github.com/robertofranceschi/Default-Credit-Card-Prediction
6. https://medium.com/swlh/predicting-credit-card-defaults-with-machine-learning-fcc8da2fdafb
7. Jing Gao, Wenjun Sun, Xin Sui, "Research on Default Prediction for Credit Card Users Based on XGBoost-LSTM Model", Discrete Dynamics in Nature and Society, vol. 2021, Article ID 5080472, 13 pages, 2021. https://doi.org/10.1155/2021/5080472
8. S. Arora, S. Bindra, S. Singh, and V. K. Nassa, "Prediction of credit card defaults through data analysis and machine learning techniques," Materials Today: Proceedings, vol. 51, no. 1, pp. 110-117, 2022. DOI: 10.1016/j.matpr.2021.04.588.