Credit Card Default Prediction

Syed Muhammad Maisam
Registration Number: 2310691
sm23587@essex.ac.uk
University Of Essex

Abstract

This report explores credit card default prediction using a dataset of Taiwanese credit card clients with six months of credit history from April 2005 to September 2005. The research aims to identify relationships between multiple demographic and financial variables with default risk. Leveraging multiple machine learning algorithms to predict credit card defaults enabling the financial institutions and intermediaries to reduce risk and achieve maximum profits. Reflections include fine tuning of the models and include further possibilities of model expansion. In the end, research findings add to the data driven management strategies to mitigate risks.

I. Introduction

In the financial services, targeted and tailored based segmentation of clients on their risk profiles holds immense importance to innovate best products. Advanced Analytics and Machine Learning enables financial institutions to correctly identify their credit worthy customers. This granular insight helps them create specific and successful tailored products for their ever-changing customer base. Every organization hopes to achieve two major goals i-e profit maximization and risk minimization. And to achieve these goals they have utilized data driven approaches to their models. In doing so, credit providers gain valuable insights and offer financial advice to customers without hassle. This synergy establishes a robust risk management and a resilient financial ecosystem while maintaining proper sustainable growth.

II. Data, Research Questions and Methodology

A. Dataset

The dataset used in this study consists of Credit Card Clients in Taiwan from the April 2005 – September 2005 (6 Months), half of the year's data. The dataset has 30,000 rows (clients) and 25 columns (variables) which contributes to a person being a defaulter or not. This dataset has certain limitations as well because it is an old dataset from 2005 so there might be new variables along with these variables to predict a defaulter now. However, these variables are still the most crucial factors for credit modelers even today.

Statistics of some important variables are shown in figures below. Fig.1 shows the demographical features of the data and its distribution in different segments of the clients. Fig.2 similarly extends the split of defaulter vs non-defaulter in each segment. For example, people who are Single (Marriage = 0) have a higher rate of default than the Married people (Marriage = 1).

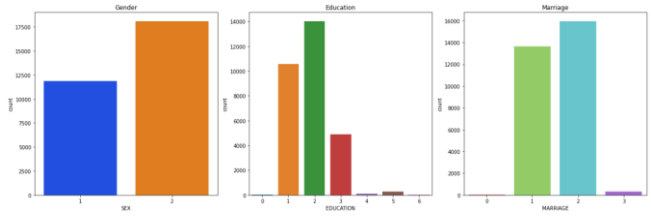

Fig. 1. Data Distribution Categorical Variables (*Gender, Education and Marriage*)

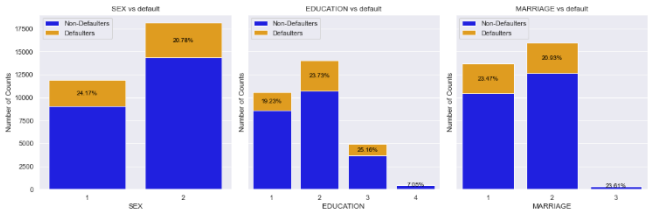

Fig. 2. Data Distribution Cat Variables against Defaulters and Non-Defaulters

Fig.3 shows the distribution of Repayments each month with their counts to show that mostly people tend to buy a loan of around 40,000 to 100,000 (NTD).

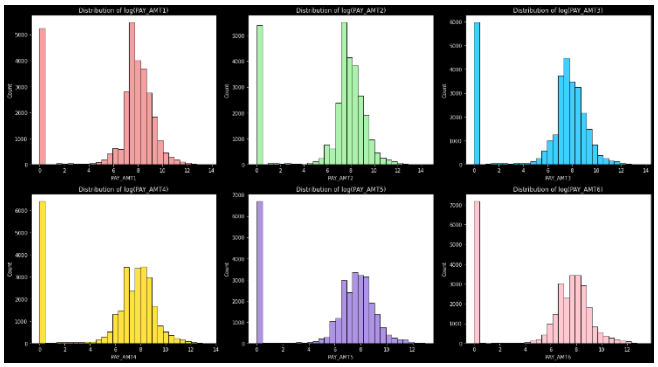

Fig. 3. Repayments Distribution for Months (*1-6*)

B. Research Questions

These Analytical questions highlight the features and the importance of data driven approach to create strong risk minimizing strategies and reduce default rates in financial markets. These questions are derived to pursue and capture a trend between credit features and default rates. The questions are as followed:

- Do demographic groups exhibit higher default rates and depict certain behaviour among customers?
- Can credit limits for each segment can justify delinquency?

- Repayments in credit profiles are always crucial. Are there any specific patterns in payment behaviour that are indicative of potential default?
- Machine Learning predictors can help formulate an approach for the Credit Risk Institutions like Credit Cards for Banks?

C. Methodology and Approach

The approach utilizes the data of credit card clients in Taiwan from year April 2005 – September 2005. By leveraging some machine learning "classification" algorithms and visualizations this report follows the basic data science pipeline to answer the analytical questions. The insights generated gave a full and comprehensive distribution of the dataset and the model features to gain most accurate results.

- Loading the data into a DataFrame. (The Data is from Kaggle)
- Perform Exploratory Data Analysis (EDA) to gain deeper understanding and underlying patterns in the Data.
- Correlation matrix to sense which features are most suitable for the Model as a whole.
- Data Cleaning, Filtering and Transformations.
- Feature Engineering to generate input for the Model.
- Standardization of the features/variables to get better results.
- Synthetic Minority Oversampling (SMOTE) to solve problems of Overfitting.
- Classification Algorithms such as KNN, Logistic Regression, and XGBoost with best hyperparameters to classify Defaulters and Non-Defaulters
- Predicting Probability of Default
- Drawing Feature Importance and other plots to make a statement.

III. Preliminary Analysis And Findings

A. Age and Consumption

The distribution in Fig.4 shows distribution of credit cards with respect to age. It gives an indication that people within the age of 20 to 30 have more inclination towards financial products such as credit cards in comparison to people above the age of 50.

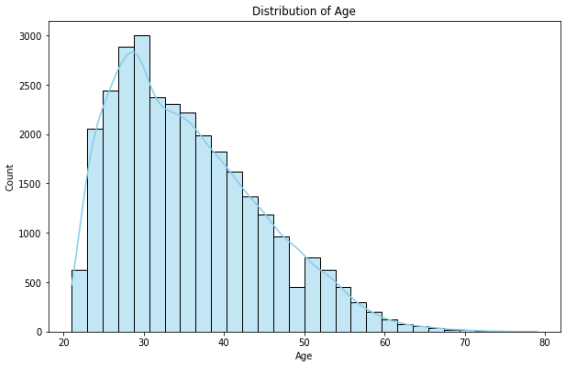

Fig. 4. Age Distribution

However, there is also a high number of defaulters in the same age bracket as compared to old people indicating caution prior to lending financial services to younger people as shown in Fig.5.

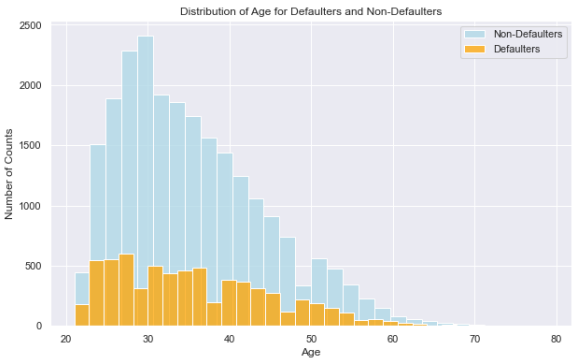

Fig. 5. Age Distribution (*With Target*)

B. Credit Limits and Delinquency

Analysing the relationship between (LIMIT BAL) and diverse financial variables, such as billing amounts (BILL_AMT1 to BILL_AMT6) and repayment amounts (PAY_AMT1 to PAY_AMT6), the plots revealed that credit limits show a pattern relating to default rates. People with higher the credit limits have lower default rates hence a lower delinquency rate. This is largely because people with higher credit limits have better socioeconomic prospects. The study provides concrete evidence showing that credit limits also help ascertain boundaries between defaulters and non-defaulters as a proxy. Thus, tailor made campaigns and products can be designed for different demographics to suit their needs and avoid defaults. With more data on the plate specific credit thresholds can be derived from these plots to illustrate the importance of studying the financial features to predict delinquency in Banks and other financial organizations.

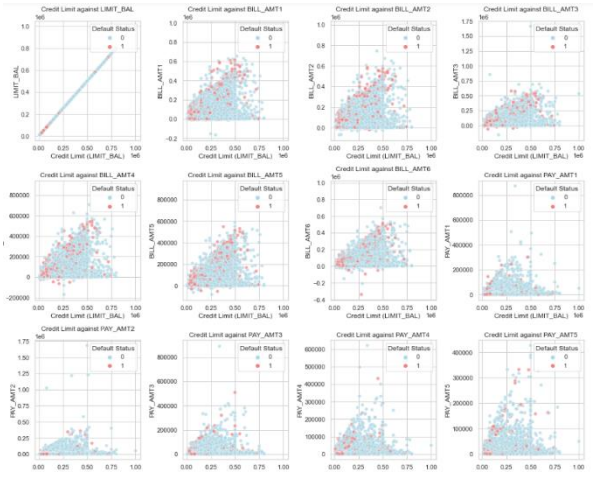

Fig. 6. Scatter Plot of Credit Limit against Financial Variables

C. *Repayment as a Tool*

Correlation Matrix exploits hidden relationships between features. Positive correlations among repayment status variables (PAY_1 to PAY_6) suggest that individuals tend to exhibit consistent repayment behaviour over multiple months which can also be deduced by the fact that there is a higher number of non-defaulters and a lower number of defaulters in the dataset. The correlation analysis between the repayment status variables (PAY_1 to PAY_6) and the target variable, "default," reveals interesting patterns in their associations. Early repayment exhibits behaviour which shows individuals with good early repayments have a higher tendency to remain non-defaulters. People that take repayments to longer periods like 4th to 6th month showed increase signs of defaults.

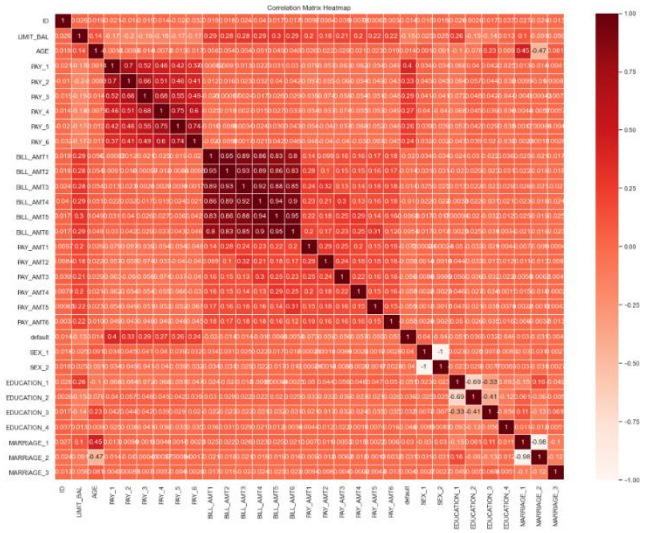

Fig. 7. Correlation Matrix

IV. CLASSIFICATION MODELS TO PREDICT DEFAULTS

Multiple supervised machine learning models such as KNN, Logistic Regression and XGBoost were developed after data preprocessing, feature engineering and standardization to predict defaults. The best model was selected based on Accuracy. The KNN model Accuracy was 76.7% whereas Logistic Regression model Accuracy was 77.9%. However, XG Boost had the highest accuracy amongst all the models.

XGBoost classifier was used to determine the best predicting model to detect default. This was crucial to analyse which model performs well in what environment and generates the best possible probability of default and feature importance. The results in the Table I below shows different runs of the Model using Grid Search Cross Validation technique to generate the best performing XGBoost model. Accuracy being the best metric in this case was realised giving the highest value of 84.9% at Fold-4 shown in Table I. Though the accuracies are not far apart for the other Folds. An average accuracy of 83% is observed across all models.

	Classification Report for XGBoost Model			
	Accuracy	***Precision***	***Recall***	***F1-Score***
Fold1	84	0.81	0.82	0.80
Fold2	83	0.81	0.82	0.80
Fold3	84	0.79	0.81	0.78
Fold4	85	0.82	0.83	0.81
Fold5	83	0.80	0.82	0.79

TABLE I. CLASSIFICATION REPORT

The confusion matrix for each fold is also useful to analyse what is the business demand and expectation when providing credits to customers. Segmentation of customers on credit risk helps in strategic refinement of the policies and decision-making process. Consider the implications of false positives and false negatives when handing out credits without the backing of data. There are two aims of every organization or firm which are to maximize profits and minimize risks. Some businesses might suggest reducing false positives and some might suggest removing false negative and some may want both to be as low as possible. By looking at the confusion matrix recall seems to be a better metric than precision since number of true negatives is relatively higher than false positives.

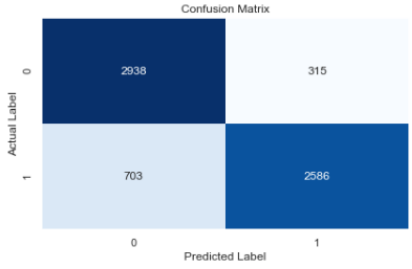

Fig. 8. Confusion Matrix Model

V. CONCLUSION AND FURTHER WORK

A. *Probability of Default and Feature Importance*

The graph of probabilities with the ROC curve can provide valuable insights into the effectiveness of your credit risk model and how well it distinguishes between defaulters and non-defaulters at different probability thresholds.

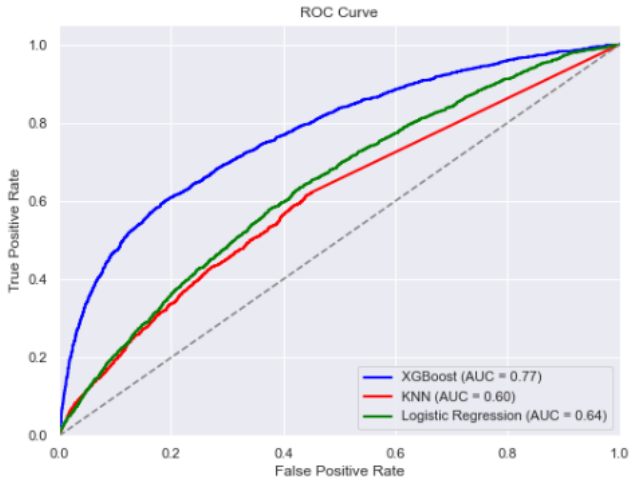

Fig. 9. ROC-Curve showing Probability of Default

The ROC curve and associated probability thresholds can guide the segmentation of customers based on their credit risk profiles. By selecting an appropriate threshold, customers can be categorized into different risk segments, allowing for targeted product development and marketing strategies. The graph helps identify regions on the curve where the true positive rate is higher, indicating a better ability to identify high-risk customers. These segments can inform strategies for developing specialized financial products or services tailored to customers with higher default risk.

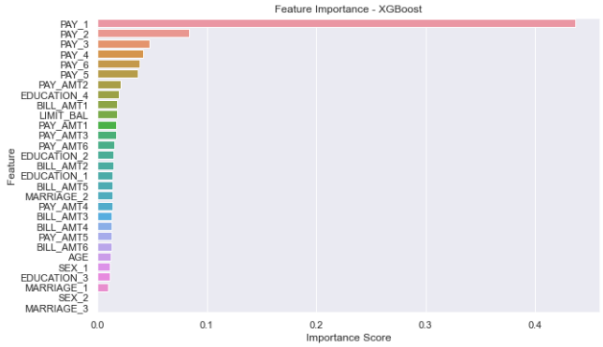

Fig. 10. Feature Importance

Fig. 10 shows that most significant features in the XGBoost model are Repayments especially in the first repayment month. While these are the most important one's others does have a low impact on the model's overall performance. This concludes the statement that demographic features with substantial data can determine a person's delinquent nature but mostly relies on financial stability and quick response when returning Credit Card Loans.

B. *Critical Reflections*

XGBoost is a very flexible model which can be more fine-tuned to reduce Type-I and Type-II errors in classifications. There was also a slight imbalance in the dataset since the dataset is mostly comprised of non-defaulters. However, SMOTE was used to tackle issues of under sampling to a certain acceptable extent.

C. *Further work*

Conducting Temporal Analysis using the most recent data to identify relationships between time and these features will also a field unexplored in this study. How default patterns are changing with time and how flexible are time related features for Risk Models to incorporate?
Working with Industry professionals to grasp new trends that are emerging. And how powerful are the tools developed that can be utilized to make this process easier and faster?

VI. REFERENCES

[1] [1] Machine learning based consumer credit risk prediction, https://ouci.dntb.gov.ua/en/works/4Lg0qgW7/ (accessed Dec. 24, 2023).

[2] Analysis of financial credit risk using machine learning - researchgate, https://www.researchgate.net/publication/318959365_Analysis_of_Financial_Credit_Risk_Using_Machine_Learning (accessed Dec. 24, 2023).

[3] Credit Risk Analysis Using Machine Learning classifiers | IEEE ..., https://ieeexplore.ieee.org/document/8389769 (accessed Dec. 24, 2023).

[4] J. C. K. Chow, "Analysis of financial credit risk using machine learning," arXiv.org, https://arxiv.org/abs/1802.05326 (accessed Dec. 24, 2023).