# Credit Card Default: A predictive Analysis

Adedamola Bowale
*A0353496*

***Abstract*— In the majority of financial institutions, credit risk management is an essential part of the overall business strategy. There are a variety of instruments and approaches that financial institutions can employ to reduce the risk of their loan recipients defaulting. In order to assess the risk of loan applicants, credit scoring is a typical procedure that may be performed using classic statistical techniques as well as AI and machine learning.**

**This study examines customers' payment defaults so as to accurately analyze the chance of payment default. The dataset's preprocessing phases (including EDA) have enabled us to clean and extract insights from the entire dataset. While the default class of the dataset is unbalanced, we utilized the (SMOTE) oversampling method to rectify the dataset's imbalance.**

**In terms of overall performance, the Random Forest approach outperforms other machine learning classifiers such as Logistic regression and Decision Tree Classifier, all of which were utilized in this research.**

***Keywords—credit scoring, oversampling, random forest***

## I. Introduction

In credit risk management, the recent development of AI and Machine Learning (ML) technology has led to a number of benefits and risks. The purpose of this project is to provide banking organizations and other financial institutions with a straightforward and reliable method for predicting whether or not their customers will be able to pay their credit card bills on time using Machine Learning Models.

The customer record contains crucial information that can be utilized to identify defaulters. Information such as gender, age, and level of education, combined with customer details and the credit history of the previous six months, are used to make a trustworthy and accurate prediction as to whether a customer will default payment the following month(s) or not. [1].

In order for a Machine Learning Algorithm to better grasp the dataset, a data preprocessing phase had to be carried out ahead of time. Our variables were first screened for missing values, duplicates and outliers, and then visually analyzed in this phase so that we could better understand the variable distribution in the data set. To convert string values to integers, we employed the one-hot encoding method.

While the dataset is imbalanced, an oversampling strategy was used to avoid bias when applying Models. Accuracy, Confusion Matrix, and Recall were calculated using Logistic regression, SVM, and Random Forest models. To improve accuracy, hyperparameter adjustment was also attempted.

## II. Literature Review

Credit card debt began to drop in the early '90s, which sparked a surge in personal bankruptcy after 1995. This has become a major issue for banks and policymakers. [2]

Xia et al. [3] established a credit rating methodology to define consumers as either healthy or loan defaulting. Preprocessing the data was necessary because of the presence of erratic values in the P2P lending dataset. Evaluation of the outcomes was carried out utilizing modern xgboost algorithms and keyword clustering-based methods. Classifier performance was improved by removing dominant features. As demonstrated in their tests, the Cat boost algorithm that relies on xgboost beat out other traditional models.. [3]

According to Ausubel's 1991 article [4], illogical customer behavior and unfavorable selection problems are to blame for credit card market failure. Credit card holders with bigger amounts are more likely to default, according to this study of default in this credit card space. [5]

## III. Data Exploration and Feature Selection

### *A. Strategy*

The strategy is to perform various pre-processing steps on the data, including checking for null values, converting categorical columns to numerical columns, and removing unnecessary columns, as well as engineering the features, scaling, and splitting in order to apply Machine Learning models and verify their level of accuracy using Python programming tools.

### *B. Dataset for analysis*

This dataset was downloaded from Kaggle website (https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset) [6]. Data on Taiwanese credit card users between April 2005 up until September 2005 includes default payment details, demographic details, payment information, and billing statements. There are 30,000 variables in the dataset array

### *C. Data Exploration*

The data was first shown in a way that helped people understand how the probability of defaulting on a payment change by different categories of demographic variables and which variables are the best predictors of defaulting on a payment.

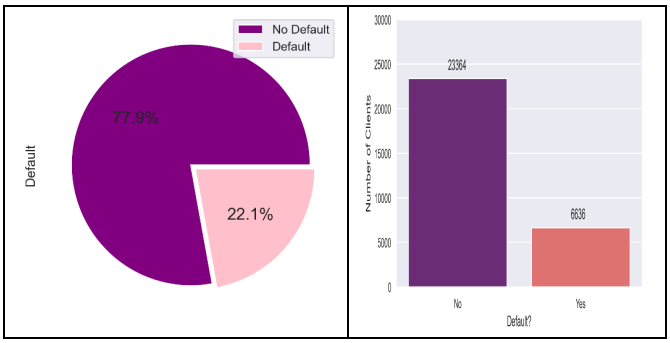


*Figure 1: Distribution of clients that defaulted*

Figure 1 depicts the distribution of non-defaulters and defaulters. Biased or imbalanced dataset

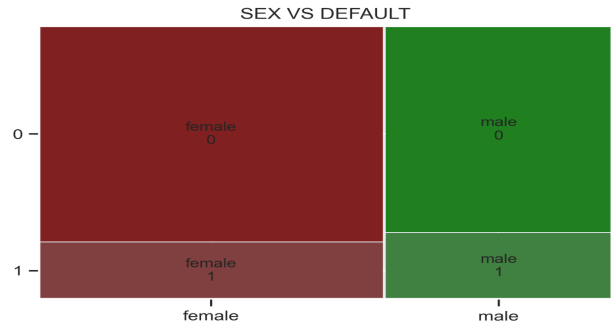


*Fig 2: Default vs Sex*

Figure 2 illustrates that the Male class defaults more frequently than the Female category, despite the fact that females are more in this research.

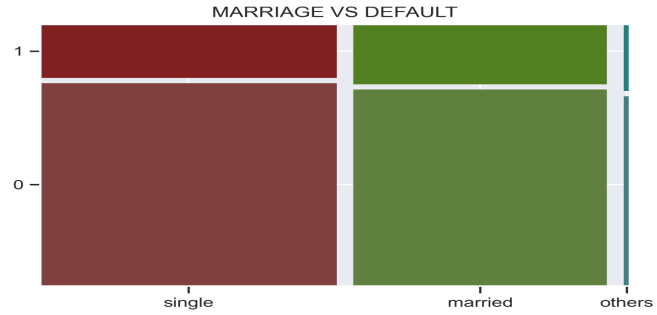


*Fig 3: Default vs Marriage*

Figure 3 indicates that married people default at a greater rate than singles, despite the bigger number of singles. The "other" category is likewise most probable to happen, but we have a limited sample size.

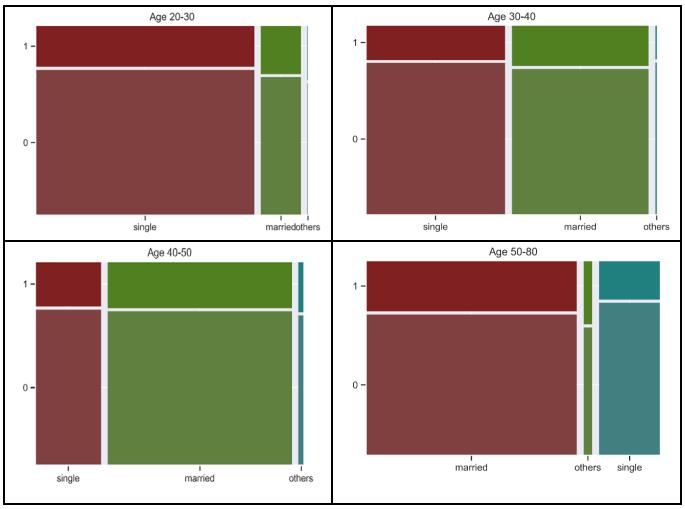


*Fig 4: Marriage vs Age bin vs Default*

Figure 4 shows that married couple's default at a higher rate than singles, regardless of age. People in their 20s and 30s are more likely to be unmarried, while those in their 40s and 80s are more likely to be married therefore more likely to default.

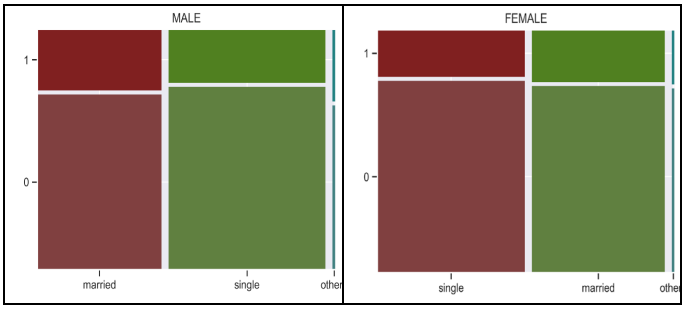


*Fig 5: Marriage vs Gender vs Default*

Singles are more common among men and women, but Figure 5 reveals that married individuals defaults way more rate than those who are unmarried.

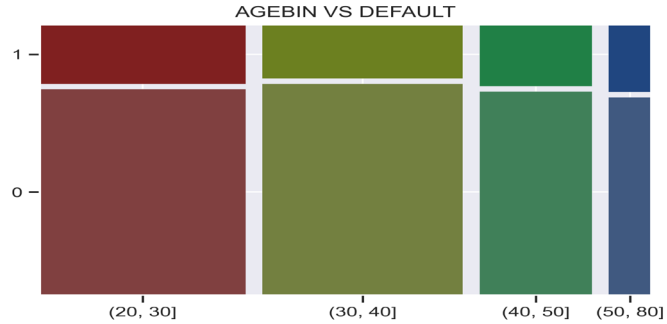


*Fig 6: Default vs Age bin*

Age 20-30 is more common than 30-40 and 40-50, with 50-80 being the least common. 50-80-year-olds default the most, followed by 40-50-year-olds, 20-30-year-olds, and 30-40-year-olds.

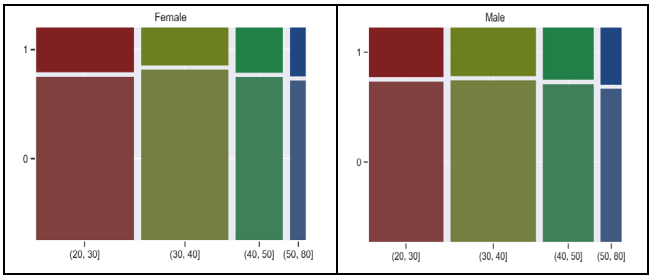


*Fig 7: Age bin vs Gender vs Default*

20-30 is more frequent than 30-40, 40-50, and 50-80 in the female class, whereas 30-40 is more common than 20-30, 40-50, and 50-80 in the male category. 50-80 age range defaults most regardless of gender, followed by 40-50, 20-30, and 30-40.

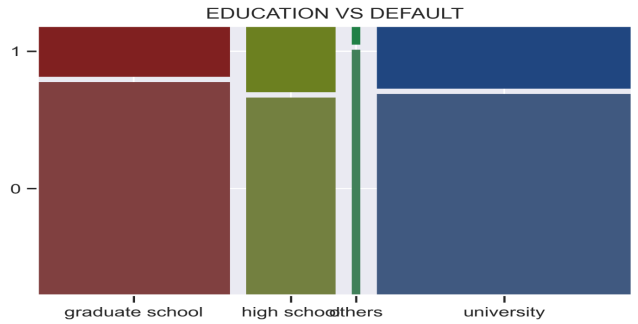


*Fig 8: Default vs Education*

Despite the fact that those with a university education are more common than those with a graduate school education or a high school education, Fig. 8 shows that those with a high school education are the ones who default the most, preceded by those with a university education and then those with a graduate school education.

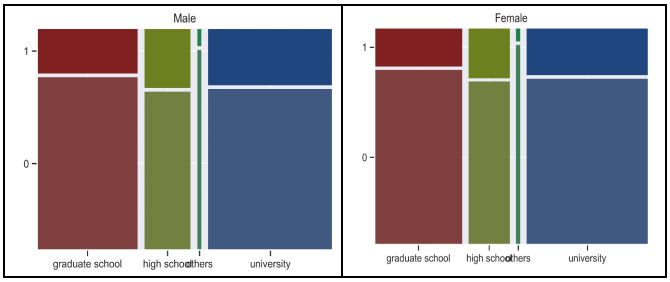


*Fig 9: Education vs Gender vs Default*

Figure 9 shows that, regardless of gender, those who attended high school default the most frequently, followed by those who attended university and, finally, those who attended graduate school, whereas the University Education grade seems to be more common.

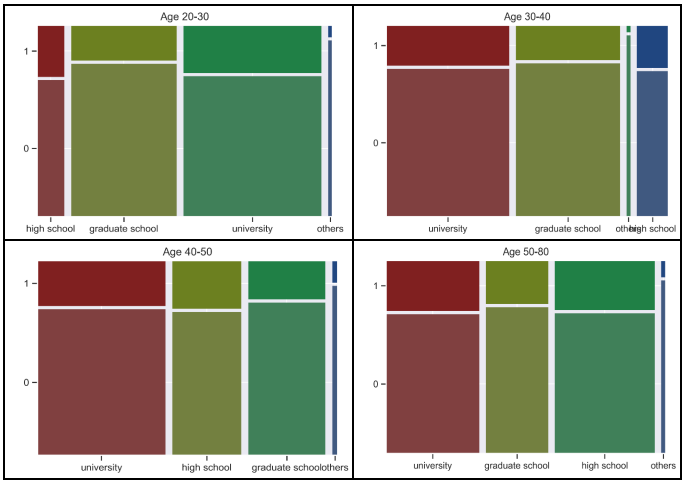


*Fig 10: Age bin vs Education Vs Default*

Figure 10 shows that regardless of age, those with a High School Education are the most likely to default on their loans, followed by those who went to university and then graduate school, even though the University Education grade is more common in all age groups except for those between 50 and 80 years old.

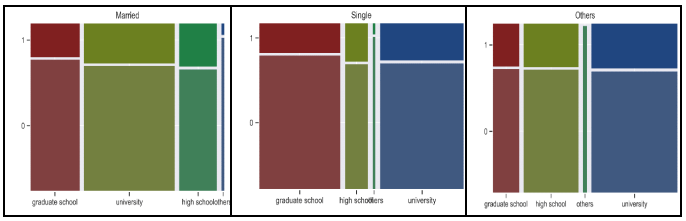


*Fig 11: Marriage vs Education Vs Marriage vs Default*

According to Fig 11, married and single people have a higher likelihood of having a university education than single people, but married and unmarried, the high school education level default rate is the highest, then by the university and ultimately graduate school.

### *D. Data Transformation*

- *Normalization*: While it was not necessary to transform binary/non-binary category data to numeric, normalization was required. Normalization provides each variable with equal weights/importance, guaranteeing that no one variable's size does not skew model performance in a particular direction. This improves the performance and training stability of the model. The data were scaled using Minmax scaling algorithms using the concept of subtracting the minimum value from the maximum value of each column and dividing by the range. Each new column has minimum and maximum values of 0 and 1, respectively.
- *Feature Selection:* Feature selection involves determining which variables are most significant for your analysis. This will train ML models. The advantages of this stage are Faster model training, Improved accuracy, Reduced over-fitting, Reduced model complexity.

  This can be accomplished by the correlation or feature significance methods. The correlation plot not only provides a variety of new perspectives on the dataset and the model, but it also helps to improve the accuracy of the model.

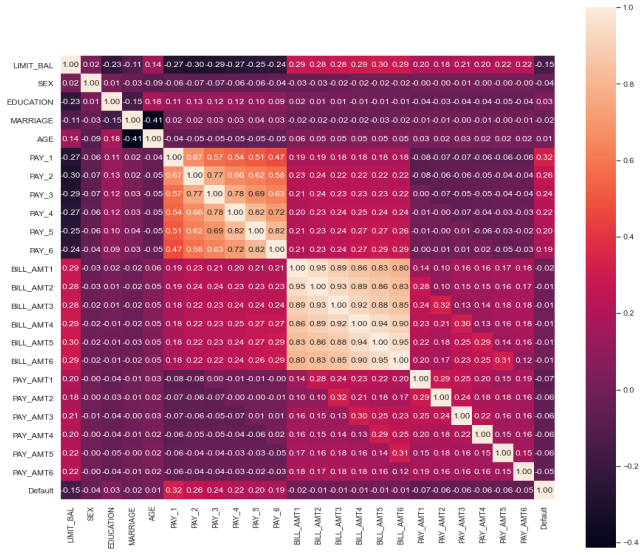


*Fig 12: Correlation Plot*

## IV. Experiment

### A. Test and Train split

*Test and Train split*: This is a model validation process that simulates the performance of a model on new/unseen data. The dataset was then divided into two parts: 70 percent for training and 30 percent for testing.

### B. Oversampling

Because the dataset is not balanced, it was essential to carry out the oversampling method. An alternative method to balancing an uneven dataset is called undersampling; however, this method can result in the loss of data. The oversampling process known as SMOTE was performed on the training dataset

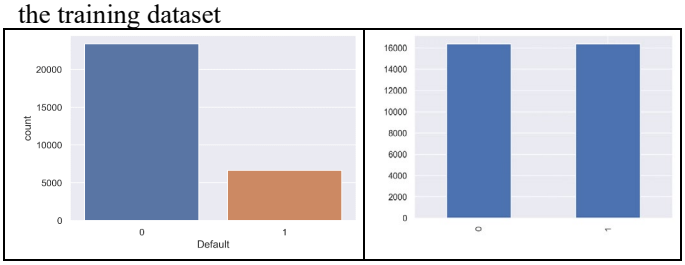


*Figure 13: Imbalance/Oversampling*

### C. Experiment Approach

Use data mining techniques to predict if a customer would default on his credit card, and then identify which model delivers the most accurate predictions for the diagnostics based on this information. We would use the following algorithm to train and test our dataset for this assessment:

- Logistic Regression
- SVM
- Random Forest

## V. Result

Confusion matrix is the foundation of our results in the sense that all other metrics are derived from its output. Metrics including F1 Score, Precision, Recall, and Accuracy, among others. Random forest has the best accuracy with 79%, followed by SVM with 76% and Logistic Regression with 68%. Results are categorized below.

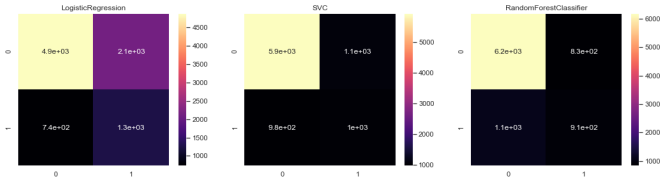


*Figure 14: Confusion Matrix*

| Models | Metrics | precision | recall | f1-score | support |
|---|---|---|---|---|---|
| Logistic Regression | 0 | 0.87 | 0.69 | 0.77 | 7000 |
| | 1 | 0.37 | 0.63 | 0.47 | 2000 |
| | accuracy | | | 0.68 | 9000 |
| | macro avg | 0.62 | 0.66 | 0.62 | 9000 |
| | weighted avg | 0.76 | 0.68 | 0.7 | 9000 |
| SVM | 0 | 0.86 | 0.82 | 0.84 | 7000 |
| | 1 | 0.46 | 0.53 | 0.49 | 2000 |
| | accuracy | | | 0.76 | 9000 |
| | macro avg | 0.66 | 0.67 | 0.67 | 9000 |
| | weighted avg | 0.77 | 0.76 | 0.76 | 9000 |
| Random Forest | 0 | 0.85 | 0.88 | 0.87 | 7000 |
| | 1 | 0.53 | 0.45 | 0.49 | 2000 |
| | accuracy | | | 0.79 | 9000 |
| | macro avg | 0.69 | 0.67 | 0.68 | 9000 |
| | weighted avg | 0.78 | 0.79 | 0.78 | 9000 |

*. Table 1: Classification Report*

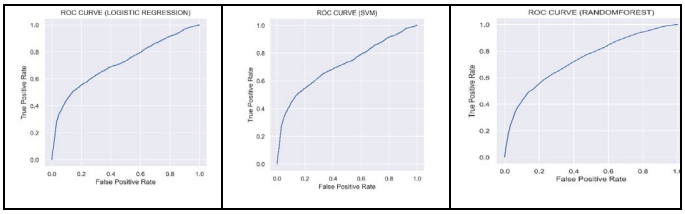


*Figure 15: ROC Curves*

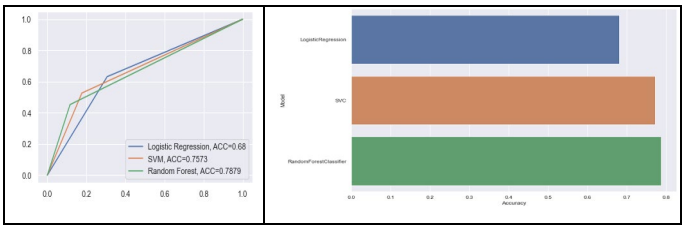


*Figure 16: Result comparison*

## VI. Ethical Concerns

### A. Data Exploratory stage Ethical concerns

Prejudice drawn from the analysis of many demographic variables/predictors is possible in this study's context Profit maximization may be a major concern for business owners who decide to limit credit approvals to married persons, male category, graduates from high school, and those between the ages of 50 and 80. However, non-members of this group may notice a pattern of rapid and excitable approvals when this isn't their primary goal.

### B. Modelling stage Ethical concerns

An overfitting model, which assumes that the bulk of credit cards really aren't defaulted, will start learning new patterns from our skewed dataset if we use a Machine Learning Algorithm, for example. SMOTE, a prominent over-sampling strategy that creates artificial minority samples by dynamically estimating across chosen minority samples and its nearest neighbors, was used to adjust for our skewed data to avoid overfitting [7]. In Fig. 13, you can see how the over-sampling worked out.

## VII. Conclusion and Recomendation

The entire process assesses the performance of numerous algorithms to calculate the potential of a credit card payment

default. Various metrics are used to evaluate the efficiency of algorithms. Selecting a model merely based on its accuracy isn't enough. Consequently, recall and model precision are essential. A high recall rate results in more defaulters being apprehended. And reasonable precision results in fewer False Positives. In order to integrate and execute the optimal model for predicting new entries, the random forest, which has the highest accuracy, would be the ideal choice.

Hyperparameter tuning was attempted in this project in order to improve accuracy for the Random Forest Model. However, the result of the accuracy was slightly reduced. It is highly recommended to keep experimenting with the parameters in order to get a higher accuracy

## REFERENCES


[1] S. Sathya Bama,, A. Maheshwaran, S. KishoreKumar, K. RaghulKumar and M. Yogeshwaran, "Identification of Default Payments of Credit Card Clients using Boosting Techniques," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 6, 2020.

[2] I. Domowitz and T. L. Eovaldi, "The Impact of the Bankruptcy Reform Act of 1978 on Consumer Bankruptcy," Journal of Law and Economics, vol. 36, no. 2, pp. 803-835, 1993.

[3] Y. Xia, L. He, Y. Li, N. Liu and Y. Ding, "Predicting loan default in peer-to-peer lending using narrative data," J. Forecasting, vol. 39, no. 2, pp. 260-280, 2020.

[4] L. M. Ausubel, "Credit Card Default, Credit Card Profits, and Bankruptcy,," American Bankruptcy Law Journal,, vol. 71, pp. 249-270, 1997.

[5] P. S. Calem, Paul and M. J. Loretta, "Consumer Behavior and the Stickiness of Credit-Card Interest Rates,," American Economic Review, vol. 85, no. 5, pp. 1327-1336., 1995.

[6] kaggle.com, "Default of Credit Card Clients Dataset | Kaggle," [Online]. Available: https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset.

[7] K. Fujiwara, Y. Huang, K. Hori, K. Nishioji, M. Kobayashi, M. Kamaguchi and M. Kano, "Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis," Frontiers in Public Health, vol. 8, p. 178, 5 2020.