# 

# 

# 

# 

**CET023**

**AI and Analytics in Finance, Credit and Related Risks**

**Group Report**

Group 2

Instructor: Teoh Teik Toe

# 

| **Name** | **Matriculation Number** |
| --- | --- |
| Ng Yew Kong | K2122342D |
| Ho Faye | K2122339G |
| Lim Jun Jee Gabriel | K2122341G |
| Theng Foo Yin | K2122345F |
| Chai Si Ting | K2122348H |

# 

[**1. Introduction & Objective**](#_ddb1wyh41buv) **3**

[1.1 Introduction](#_9umbdfszffo9) 3

[1.2 Objectives](#_p0kowre8fr5q) 3

[**2. Literature Review**](#_ttbxqxajprc) **4**

[2.1 Methodology](#_c5owj993unpz) 4

[2.2 Determinants of Credit Card Defaults](#_hgh40itg1e9x) 4

[2.3 Consequence of Credit Card Defaults](#_152l325mt19g) 4

[**3. Business Solution (Problem Statement)**](#_i6cq692gyol7) **5**

[**4. Exploratory Data Analysis**](#_vi2ojh73qomj) **6**

[4.1 Dataset Information](#_yq3m8quqm0ls) 6

[4.2 Data Visualization: Default Rate](#_kk1k4k8tyjsf) 7

[4.3 Data Visualization: Credit Limit](#_samivm9jptja) 7

[4.4 Data Visualisation: Age](#_qfs5c82zj2aa) 8

[4.5 Data Visualisation: Six months Outstanding Debt](#_xsfkk7vthu5h) 9

[4.6 Data Visualisation: Gender](#_pbzu7eel4ujv) 10

[4.7 Data Visualisation: Education](#_8mfxf5v21nla) 10

[4.8 Data Cleaning and Correlation Heatmap](#_5q8n29xcimuv) 11

[**5. Feature Engineering**](#_1qmey5gpgxib) **12**

[5.1 Remapping Data](#_sqvgkpc6a6r3) 12

[5.2 Data Manipulation and Standard Scalar](#_fgv3hlv000lp) 13

[5.3 Feature Selection](#_dryebooi3g6o) 14

[5.4 Train Test Split](#_429ayzjyh5z0) 14

[5.5 SMOTE](#_xexgiiqt3yi1) 15

[**6. Technical (Model & Configuration)**](#_ep2rwtke8z4g) **16**

[6.1 Model Selection](#_7xc3bv3ao3fq) 16

[**7. Technical (Model Comparison)**](#_nv665l5kyr7x) **18**

[**7.1 Model Performance**](#_2ekp1gy7si6w) **18**

[7.2 Feature Importance](#_we5738dy3kxl) 20

[**8. Business Solution and Recommendations**](#_ghb9f6vwuqn0) **21**

[8.1 Key features influencing default rates](#_8rex2z18f5z0) 21

[8.2 Recommendations](#_wqufd78jbpld) 21

[8.2.1 More Stringent Checks](#_dpqmmf8y17gd) 21

[**9. Conclusions & Future Study**](#_27c6xmjl6pvp) **23**

[**References**](#_t4sy30mf393o) **24**

# 

# 1. Introduction & Objective

## 1.1 Introduction

With the rise of technology, globalization, and concerns over Covid-19 being easily transmissible through contaminated surfaces, contactless and cashless payment options have been a new norm for consumers in conducting daily purchases. Credit cards, being one of the widely used alternatives, evolved from vendors allowing customers to receive items on credit with the expectation that the account balances be repaid in a given timeframe. It serves as a valuable financial tool when used responsibly, providing accessibility and convenience to users, through near field communication (NFC) cards, mobile apps, and wearables.

However, individuals may succumb to credit card defaults upon successive failures to repay its principal amount. This is often tagged with high-interest rates, for an extended period of time. Especially during unsettling economic times, with defaults occurring after six months in a row of not making at least the minimum payment due by holders. Banks are, henceforth, put at risk when these numbers start to accumulate.

It is therefore vital, in the perspective of banks, to conduct an investigation through credit analysis across the population of its credit cardholders to identify the key warning signs of poor credit and the probability of an individuals' repayment abilities. In this report, Six AI modeling techniques are adopted to provide an in-depth study, targeting to improve the bank’s loaning processes. Through analyzing customers’ data sets and combining machine-learning algorithms, better predictions can raise early warning flags to capture credit defaulters in advance.

## 1.2 Objectives

To address the study on credit card defaults, this project aims to achieve the following objectives:

* In order to **reduce the risk** of banks’ exposure in large **credit card default** incidents, utilize various data classification techniques with a large data set of customer records to screen for potential credit card defaulters.
* Identify common **key traits, features, and conditions** through the pool of customers’ datasets for early detection and trigger for preventive measures.
* To avoid experiencing customer defaults with a snow-balling effect, derive accurate **predictions and recommendations** of test results to the flag for default accounts in advance.

# 2. Literature Review

## 2.1 Methodology

Using machine learning methods on a large dataset of historical customer credit lines in predicting credit defaults, and analyzing features of customer behaviors associated with credit delinquencies, is a widely discussed topic across researchers in this era.

Butaru et al (2016) have illustrated the benefits of running large data sets collected across six banks to construct Decision Trees, regularized Logistic Regression, and Random Forest models to draw conclusions on consumers’ tendencies on credit delinquencies and to derive a prediction. Meanwhile, Neema and Soibam (2017) drew a conclusion on Random Forest, amongst seven techniques in comparison, that it provides the highest predictive accuracy in credit card defaults, thanks to the non-linearity on various cost factors.

## 2.2 Determinants of Credit Card Defaults

Identifying key features of credit card default behaviors can flag early warning signals to banks or risk management teams within an institution. T. M. Alam et al (2020) explore multiple factors such as an individual’s payment history, approved credit limits, client’s personal information, and economic status are closely linked to the probability of defaults. Neema and Soibam (2017) further asserted the importance of placing a heavier weight on credit limit, billing and payment information, while a lighter weight other discriminant variables are less significant like personal information.

## 2.3 Consequence of Credit Card Defaults

The key takeaway of this study is to provide better insights to card issuers such that a more stringent model can be rolled out starting from the client application process. Potential applicants should be carefully assessed based on the most relative characteristics indicating a higher probability of default. H. Kim et al (2018) emphasized the avoidance of significant losses institutions would suffer, with every small improvement in modeling accuracy for high-risk defaulters.

# 

# 3. Business Solution (Problem Statement)

With poor and rising uncertainties in the global macroeconomics environment, such as entering a recession or rising unemployment rate, card issuers face an increasing rate of credit card delinquencies, especially from the lower-income earners. This personal line of credit holds one of the highest interest rates when compared to a mortgage or auto loan, as it does not require collateral to be granted with a card. Hence, it is targeted for large-scale consumer consumption.

To reduce the susceptibility of banks with a hefty write-down on outstanding balances left unpaid and funding the losses from their reserves, enhancing a classifier performance on AI modeling can provide solutions to better predict potential credit card defaults, and help identify key factors leading to a default.

# 

# 4. Exploratory Data Analysis

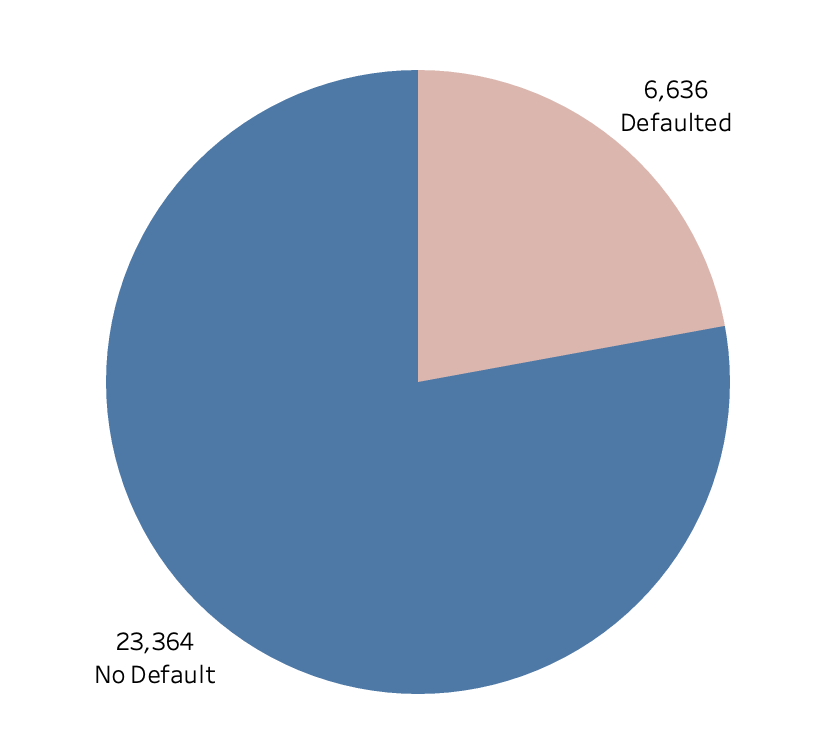
## 4.1 Dataset Information

This dataset contains 30,000 records on default payments, demographic factors, credit data, repayment records, and bill statements of credit card owners in Taiwan between April 2005 to September 2005.

| **Column ID** | **Details** |
| --- | --- |
| ID | ID assigned to each individual client |
| LIMIT\_BAL | Amount of credit limit given to the client based on their customer profile |
| SEX | Gender of client (1 = Male, 2 = Female) |
| EDUCATION | Highest level of education obtained (1 = graduate school, 2 = university, 3 = high school, 4 = others, 5 = unknown, 6 = unknown) |
| MARRIAGE | Last reported marital status (1 = Married, 2 = Single, 3 = Others) |
| AGE | Client age in years |
| PAY\_0 | Repayment status in September, 2005 (-1 = pay duly, 1 = payment delay for one month, 2 = payment delay for two months, … 8 = payment delay for eight months, 9 = payment delay for nine months and above) |
| PAY\_2 | Repayment status in August, 2005 (Same as above) |
| PAY\_3 | Repayment status in July, 2005 (Same as above) |
| PAY\_4 | Repayment status in June, 2005 (Same as above) |
| PAY\_5 | Repayment status in May, 2005 (Same as above) |
| PAY\_6 | Repayment status in April, 2005 (Same as above) |
| BILL\_AMT1 | Amount of outstanding balance in September, 2005 (NT dollar) |
| BILL\_AMT2 | Amount of outstanding balance in August, 2005 (NT dollar) |
| BILL\_AMT3 | Amount of outstanding balance in July, 2005 (NT dollar) |
| BILL\_AMT4 | Amount of outstanding balance in June, 2005 (NT dollar) |
| BILL\_AMT5 | Amount of outstanding balance in May, 2005 (NT dollar) |
| BILL\_AMT6 | Amount of outstanding balance in April, 2005 (NT dollar) |
| PAY\_AMT1 | Amount of previous bill payment in September, 2005 (NT dollar) |
| PAY\_AMT2 | Amount of previous bill payment in August, 2005 (NT dollar) |
| PAY\_AMT3 | Amount of previous bill payment in July, 2005 (NT dollar) |
| PAY\_AMT4 | Amount of previous bill payment in June, 2005 (NT dollar) |
| PAY\_AMT5 | Amount of previous bill payment in May, 2005 (NT dollar) |
| PAY\_AMT6 | Amount of previous bill payment in April, 2005 (NT dollar) |
| default.payment.next.month | Payment defaulted (1 = Yes, 0 = No) |

## 4.2 Data Visualization: Default Rate

Credit card debt historically has a higher default rate than other debt areas. This is attributed to the high-interest rate and ease of racking up debt with each swipe. 22.12% of customers in the dataset defaulted on their loans within the 7 month period on record.



Graph 1: Overall default numbers in the dataset

## 4.3 Data Visualization: Credit Limit

To understand the distribution of credit limits - we have broken down the credit limits into $50K buckets from $0K to $1M. The lowest bucket of credit limits between 0 - 50K has the highest percentage of default at 36.07%, nearly 10% higher than the next closest bucket of 50K - 100K.

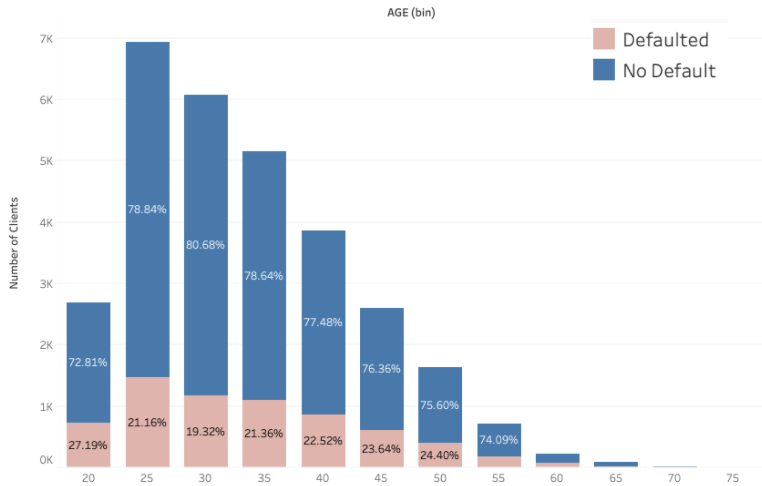
## 

Graph 2: Distribution of credit limits across the dataset with percentage defaults

## 

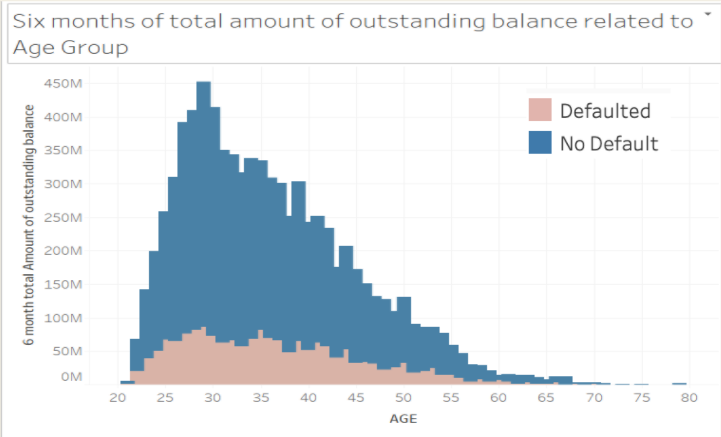
## 4.4 Data Visualisation: Age

The dataset contains clients ranging from 20 to 70 years old and have a weighted average age of customer at 35.5 years old. Out of these, 6636 customers defaulted on their payment. We can interpret from the data that the customer age group between 20 - 29 are most likely to default with an average default rate of 22.8%.



Graph 3: Distribution of clients’ age across the dataset with percentage defaults

## 4.5 Data Visualisation: Six months Outstanding Debt



Graph 4: 6 months total amount of outstanding balance with amount defaults

Above bar chart depicts clients age group between 28 to 30 tend to have higher 6 months outstanding loan than the rest of the age group. Clients spending dramatically decline after age 30 where credit card default amounts also start to gradually **decline**.

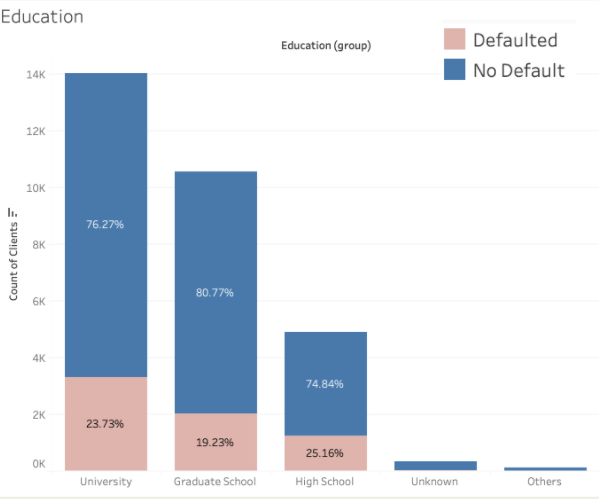
## 4.6 Data Visualisation: Gender

# 

Graph 5: Distribution of clients’ gender across the dataset with amount defaults

Females make up 60.37% of the clients in this dataset, despite that, they have a 3.37% lower default rate out of the set. This makes gender an important variable to consider when determining the overall chance of default.

## 4.7 Data Visualisation: Education



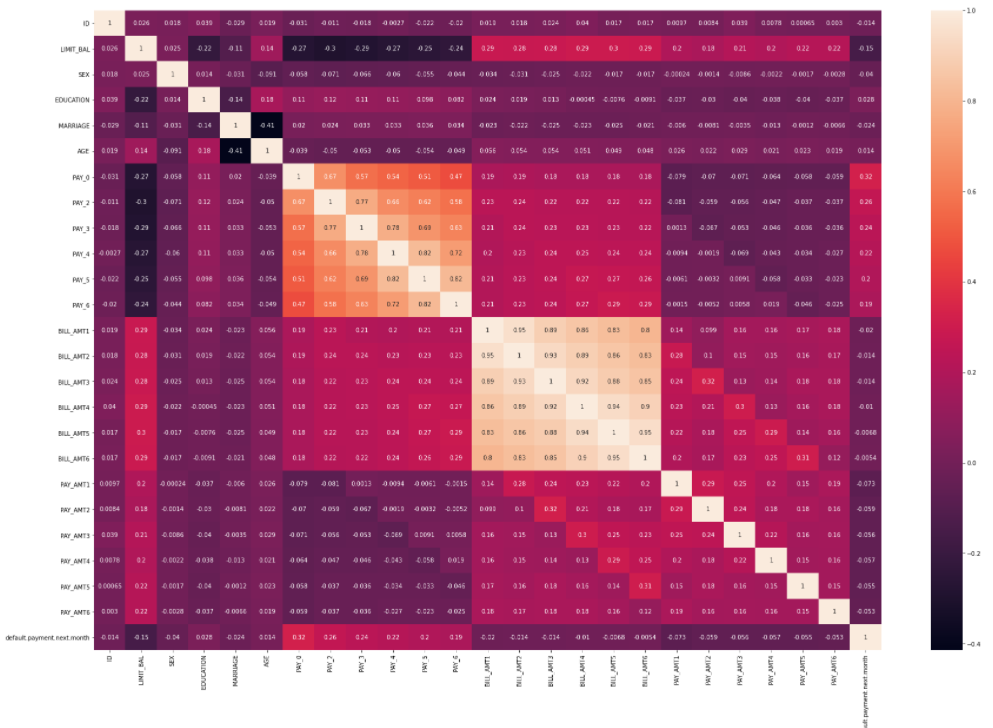
Graph 6: Distribution of clients’ educational background across the dataset with amount defaults

From this diagram, we can deduce that education has a significant contribution to the chance of default, with the highest default percentage coming from high school background (25.16%) and the lowest coming from graduate school (19.23%). There is a small percentage of unknown variables (0, 5 and 6) which are not identified in the document - we will ignore these variables in the final result.

## 4.8 Data Cleaning and Correlation Heatmap

Preparation of the dataset is carried out through screening for duplicates, missing features, dropping of null values and identifying the data type of each column to ensure uniformity. Descriptive statistics are also generated to screen for the limits and distribution of each dataset. Irrelevant features, such as ‘ID’, are dropped. Structural feature errors have been identified by remapping values in ‘EDUCATION’, ‘PAY\_0’ to ‘PAY\_6’ to remove insignificant dataset.

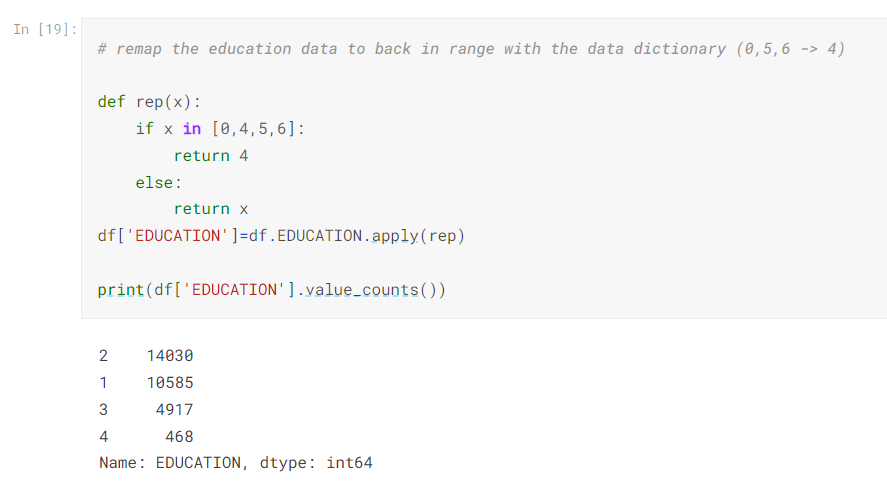
A correlation heatmap (shown below) uses a colorbar, making data easily readable, highlighting the differences and variation in the data. The color of the cell is proportional to the number of measurements that match the value. Lighter coloured cells have a high correlation coefficient. From the correlation heatmap, it is observed that PAY\_0, PAY\_2, PAY\_3, PAY\_4, PAY\_5, PAY\_6 and BILL\_AMT1, BILL\_AMT2, BILL\_AMT3, BILL\_AMT4, BILL\_AMT5, BILL\_AMT6 are highly correlated. It might be worthwhile to remove some of the correlated features to prevent overfitting. In summary, repayment status and outstanding Balance have the highest correlation within the dataset. This shows the importance of consistency of repayments when attaining the likelihood of default.



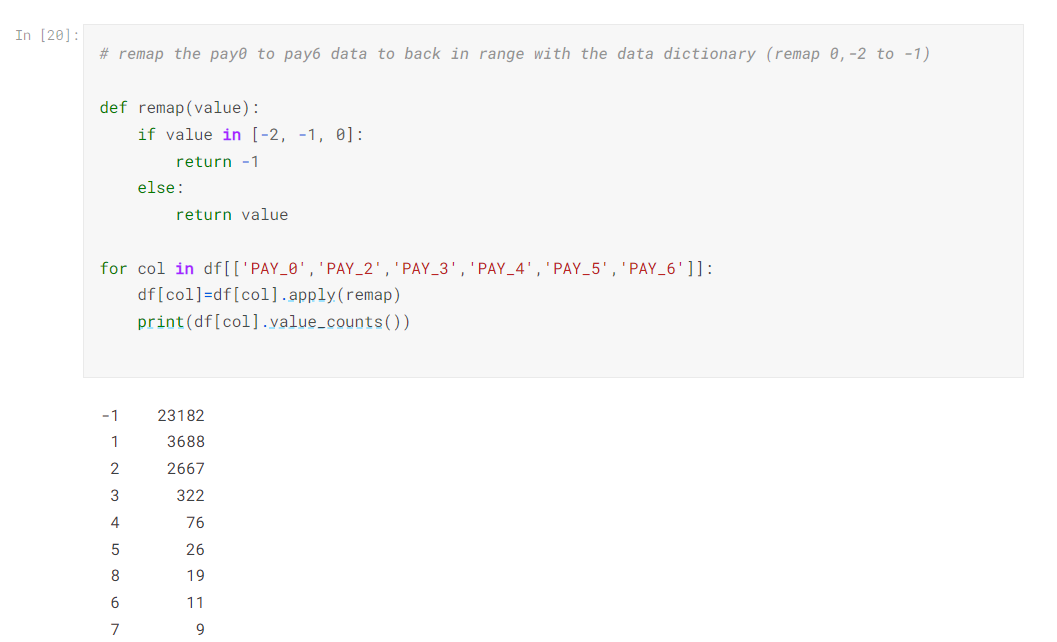
# 5. Feature Engineering

## 5.1 Remapping Data

To prepare our dataset for modeling, First starting off by remapping unknown variables from the original dataset to properly classify values (Refer to Chapter 4.7 and 4.8 for detail). The EDUCATION columns have excessive unknown variables in 0, 4, 5, and 6. It will be remapped all into 4 with a count of clients of 468.

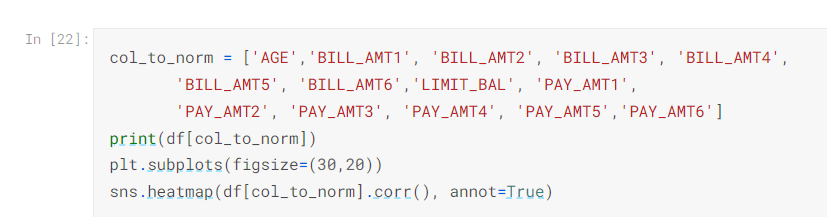


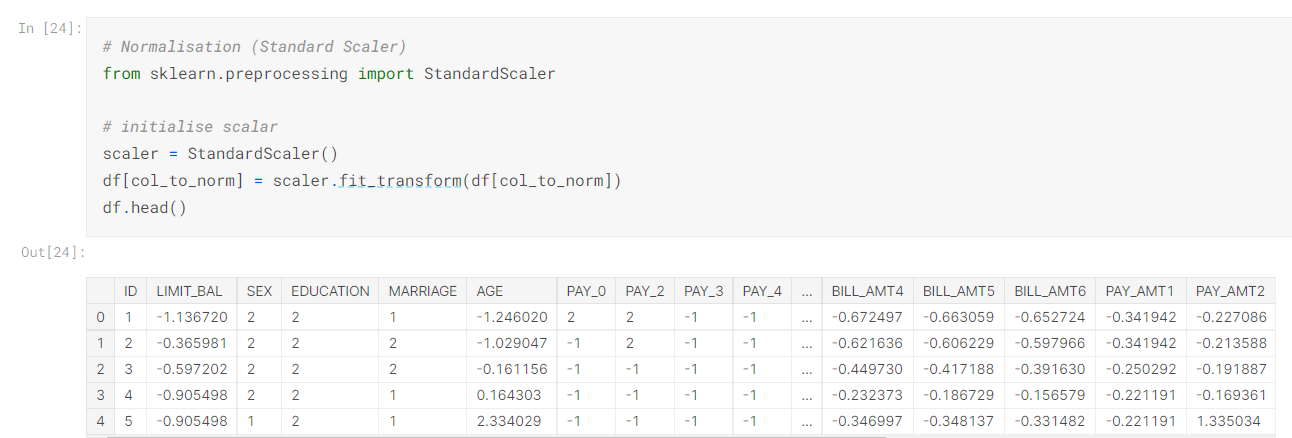
The PAY columns have excessive unknown variables in -2, -1, and 0. It will be remapped all into -1 with a count of clients of 23182..



## 5.2 Data Manipulation and Standard Scalar

After altering the data in order to make it more simplified and organized shown as below picture In [22], Standard Scalar transforms the data by scaling to unit variance and removing the mean using column summary statistics on the samples before input into train set. Refer to In[24].

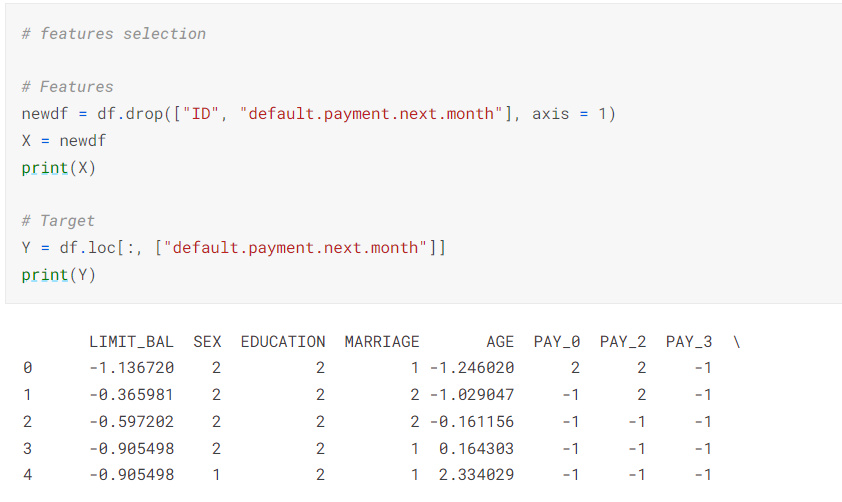




## 

## 5.3 Feature Selection

In the process of Feature selection, it is crucial to reduce the number of input variables when developing a predictive model. For this project, as depicted below, two variables have been allocated to Features X and one variable have been allocated to Target Y.



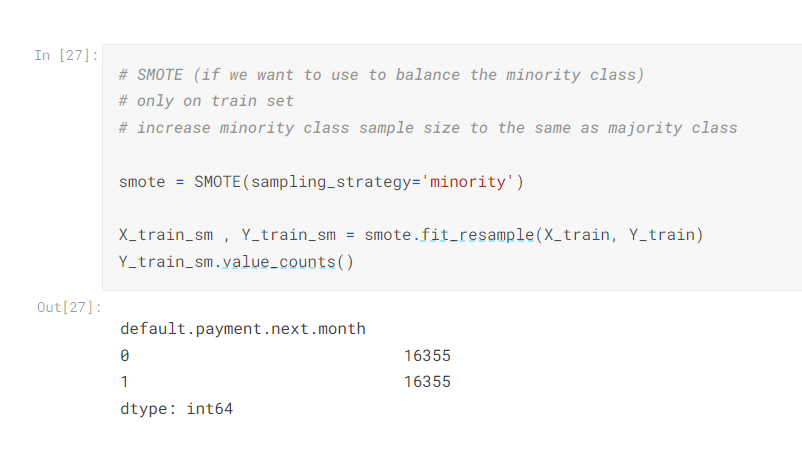
## 5.4 Train Test Split

The simplest way to split the dataset into training and testing sets is to use **the train\_test\_split()** method of the sklearn library. Assigning **30%** data points to **test** and **60%** data pointsto **train**. Therefore, we train the model using the training set and then apply the model to the test set. In this way, we can evaluate the performance of our model



## 5.5 SMOTE

As observed from the previous part on EDA, the dataset is imbalanced with the target class accounting for 22% of the whole dataset. Hence, data augmentation techniques are being used in this project to tackle the class imbalance problem. Data augmentation refers to the collection of more data through synthetic methods. There are a few commonly used data augmentation strategies such as applying random transformations, using Synthetic Minority Oversampling Technique (SMOTE) for an imbalanced dataset or using Generative Adversarial Networks (GANs). For this project, SMOTE is being used to artificially increase the minority class to be the same weightage as the majority class in the training set. SMOTE is not being used on the test set so as to preserve the data distribution of the original dataset.



# 

# 6. Technical (Model & Configuration)

## 6.1 Model Selection

In this study, **six classification algorithms** are used to classify cardholders’ data. Grid Search algorithm with 5-folds cross-validation is used to get the best estimator on our predefined parameter grid. Cross-validation is implemented as we want to ensure that our results from the models are statistically significant and not due to chance.

Our target variable is a categorical variable with two possible values (1 or 0), the cardholders who “will default” or “won’t default” their next repayment.

The goal is to assess which of the algorithms will provide the best results for predicting cardholders into the default and non-default categories. The models will be evaluated using metrics that are more robust against class imbalances, which accuracy is not good against.

**Logistic Regression**

It is a generalized linear model which uses a logistic function to predict binary results i.e if the cardholder defaults or the cardholder does not default. There must be two or more variables to come up with a logistic regression equation.

**Decision Tree**

It is a tree-like model that makes predictions based on what was answered previously. The advantage of a decision tree is that it creates a comprehensive analysis of the outcome along each branch and identifies critical decision nodes that require further analysis. Some examples of decision nodes would include the credit card bill amount and the repayment period as that would affect the probability of the credit card holder defaulting.

**Random Forest**

It is an algorithm that consists of multiple decision trees. The main difference between Random Forest and Decision Tree is that the outcome for Random Forest is not dependent on only one decision. Random Tree combines randomized decisions based on multiple decisions and makes the final call based on the majority. It is also versatile as it handles binary, categorical, and numerical features.

**XGBoost**

Extreme Gradient Boosting (XGBoost) is an additive collection of weak decision trees whereby a weak learner improves on past existing weak learners sequentially. It is a form of ensemble learning which creates new models to predict the residuals or errors of previous models and then combines them to make the ultimate prediction and minimize the error.

**Neural Network (NN)**

It is a subset of machine learning that recognizes hidden patterns and correlations in raw data, much like the brain, and then continuously learns and improves on it. NN contains neuron nodes and each neuron node connects to another node that has a certain weight and biases tagged onto it. Once the threshold has been reached, the nodes are activated which then sends data to the next node.

Typically, NN consists of an input layer, hidden layers and output layer. It is a black-box model where the weight and biases of each neuron layer are updated via backpropagation to minimize the loss function. The benefit of NN is the ability to do parallel processing which allows UCI to predict the possibility of a cardholder defaulting with much speed and accuracy.

**K-Nearest Neighbors (KNN)**

It is a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics) [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) method used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis). It assigns the output class of a data point by its surrounding data points’ class majority.

# 

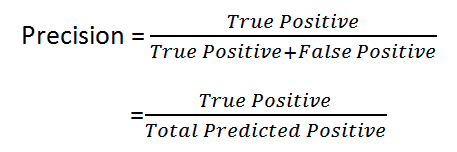
# 7. Technical (Model Comparison)

## 7.1 Model Performance

|  | **F1 Score** | **Precision** | **Recall** | **Accuracy (Train)** | **Accuracy (Test)** |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | 0.51 | 0.47 | 0.57 | 0.708 | 0.762 |
| Decision Tree | 0.46 | 0.40 | 0.53 | 0.883 | 0.720 |
| **Random Forest** | **0.52** | 0.50 | 0.54 | 0.893 | 0.779 |
| Neural Network | 0.47 | 0.39 | **0.59** | 0.796 | 0.706 |
| XGBoost | 0.49 | **0.56** | 0.43 | 0.995 | **0.799** |
| K Nearest Neighbor | 0.40 | 0.35 | 0.48 | 1.000 | 0.688 |

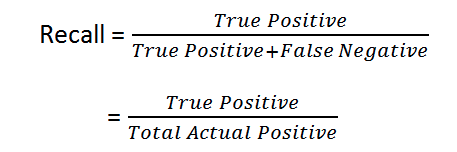
As observed from the EDA section, the dataset is imbalanced, with the target minority class being 22%. This will imply that accuracy will not be a good performance metric as a model with 99% accuracy trained on heavily imbalanced dataset (99%-1%) will be cheating the learning process by predicting the majority non-target class 100% of the time. Hence we will be looking at other metrics (F1 Score, Precision and Recall).

For Precision, it is the measure of how many actual positives are detected among the total predicted positives. ​​XGBoost is the best model having the highest Precision score of 0.56.



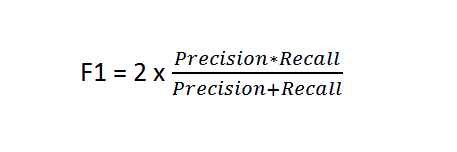
Precision is a good metric when the costs associated with False Positives are high. In our case study, the cost of having a False Positive is an intangible cost whereby a potential customer is being mislabelled as a possible defaulter on credit card loan. This might turn away valuable customers as mislabelling might lead to increased resistance or processes which affect a customer’s journey or experience negatively.

For Recall, it is the measure of how many actual positives are being labelled as true positives. Neural Network is the best model based on Recall score of 0.59.



Recall is a good metric when the costs associated with False Negatives are high. In our case study, the cost of having a False Negative is a tangible cost whereby a default gets undetected. When default gets undetected, there will be financial losses incurred to the bank. Hence it is equally important for the model to be able to make less False Negative predictions as to making less False Positive predictions. Therefore, our team has decided to strike a balance between Precision and Recall after considering the tradeoff and chose F1 Score as the overall performance metric for this case study.

F1 Score is a function of Precision and Recall and is governed by the equation below.



If we look at the F1 Score for the target class from the above table, Random Forest is the best model followed by Logistic Regression and XGBoost.

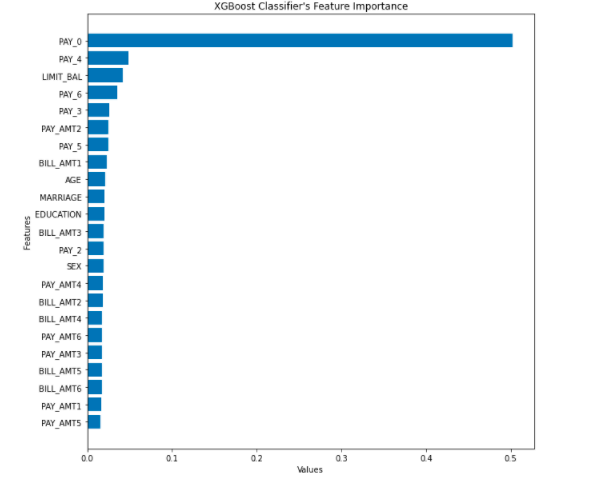
Moving on the model train-test accuracy, it is evident that there is overfitting for all the tree-based models and K-Nearest Neighbors model.

This is likely due to the tree-based models having a very deep decision tree structure (all the best estimator for the tree-based models has a max\_depth of 14). In addition, it might be due to a different data distribution in the test set as we only applied data augmentation technique (SMOTE) in the training set.

It is also observed that Logistic Regression and Neural Network models do not overfit on training set. This implies that they are more robust and are likely to generalize better to unseen dataset. Furthermore, Logistic Regression is the only model that performs better on the test set than the training set.

Finally, XGBoost model achieved the highest test accuracy of 79.9%.

## 7.2 Feature Importance



The chart above shows the feature importance in predicting the probability of the customer defaulting. This is based on XGBoost because it has the highest accuracy amongst all the other models. The most important feature in predicting the customer default is Pay\_0 which corresponds to the customer’s repayment status in September. The repayment status in June, 2005 and the amount of given credit in NT dollars are the other top 2 items. Since, Pay\_0, Pay\_4, Pay\_6 and Pay\_3 are highly correlated, we extended the analysis to look at other features such as Pay\_Amt\_2 (Amount of previous payment in July, 2005), Bill Amount\_1 (Amount of bill statement in September, 2005) and Age (Age in years) which still rank relatively high in terms of importance.

Overall, the repayment status of the customer seems to be the stronger predictive factor as compared to other features such as the demographics of the customer.

# 

# 8. Business Solution and Recommendations

## 8.1 Key features influencing default rates

**8.1.1 Age**

Clients in the age group between 20 to 25 tend to have higher outstanding loans than the rest of the age group and this might be due to their spending habits whereby they spend money on entertainment with friends. As compared to the older demographics, younger clients tend to be less prudent in their spending as they have been brought up in a world where payment is swift and easy, just with the swipe of the card. The older generation are more used to paying with cold hard cash and they understand the value of hard earned money, hence they are more cautious when it comes to spending more.

**8.1.2 Credit Limit**

In terms of credit limit, most clients tend to undertake a loan credit of 50k and clients with less than 50K have the highest defaulting rate. This might be due to the laxer requirement for customers when undertaking a lower loan amount. Hence not many checks are being done to ascertain whether capability of the customer repayment.

**8.1.3 Bill Payment and Repayment Status**

Cardholders that have a history of delaying their repayments tend to have a higher probability of defaulting their bills because their bill amount would be higher due to compounded interest. As such, customers might be reluctant to repay or they might not have sufficient funds to repay the higher bill. Leading to a higher possibility of defaulting.

## 8.2 Recommendations

Based on these observations, there are three proposed solutions to reduce the loss from a customer defaulting.

### 8.2.1 More Stringent Checks

More stringent checks can be done on the customers especially on those aged 20 - 25 to determine their ability to repay the loans. Other variables can be considered such as their income level and assets to ascertain whether they have a steady flow of income to pay back the loans or assets where they can sell to pay back the loans. A mini vetting interview can be done with the customer before granting them the loan to determine whether customers have any default records in other banks or other institutions. Customers with default records in other banks and institutions should not be granted any loans until they are able to clear their loans or prove that they have the means to clear it.

**8.2.2 Timely Reminders**

To provide more timely and frequent reminders to remind the customers of their due date and repayment amount as they might miss out. The bank should also consider degrading the limit of customers with bad debt to avoid the further costs of uncollectible debts. This should be communicated across the different departments to ensure that all departments are aware of it.

**8.2.3 Offer Incentives**

To reduce the magnitude of the loss from defaulting by offering customers incentives to repay their loans early. If there are any signs that the customer is going to default on the payment, the bank could offer a debt restructuring plan to reduce the interest rate of the loan or reduce the monthly repayment amount. Some incentives would include offering interest-free repayment plans or to work together with the customer to come out with a more manageable repayment schedule.

# 9. Conclusions & Future Study

To predict customers’ default of credit card, we applied six models including logistic regression, decision trees, random forest, xgboost, neural network and K-nearest neighbor. Having taken accuracy of the train and test set, recall, precision and F1 score into consideration, we concluded that Random Forest has the best overall performance and is the most appropriate model. However if we were to look at accuracy, xgboost is the best model of all.

By analyzing the feature importance of xgboost, we found that a customer's repayment status in September is the most predictive feature for default, and payment status in June, 2005 and the amount of given credit in NT dollars are the other top 2 items that are also of great importance.

Our exploratory data analysis (EDA) complemented our analysis from the important features. Through our EDA, we found that the customer age group between 20 - 29 are most likely to default and the highest default percentage comes from high school background while the lowest comes from graduate school.

To reduce losses to the bank, the bank could take advantage of their existing credit card records to build prediction models to identify the potential credit card default. Banks should also pay more attention to the important features relating to the default and be more sharp when processing the application of credit card by including criterias for customers to meet.

As this project has a tight deadline, we were not able to cover all bases. One key scope that can be an extension of this project is to explore other techniques to increase dataset size. There are many other techniques that can increase the number of training samples such as data collection, data augmentation or even data synthesis using GANs. We believe that implementing GANs is a probable route to take to further increase the dataset. GANs can create synthetic images that are similar to the given dataset which essentially increase the number of samples without changing the distribution of the dataset. This might translate to better generalization of the trained models.

Another potential exploration will be to inspect the outliers. GIven there are many data points that are out of range when compared to the data dictionary, these erroneous entries can be handled in a more holistic manner. Currently, such erroneous data are being remapped back in range and often to “unknown or others” categories. A better methodology will be considering if the other columns of outlier data fall in line with other data points and taking the nearest clustering strategy to impute the wrongly-entered variable. However such methods require extensive EDA, which will require more time to clean such a big dataset.

# 

# References

Butaru, F., Chen Q., Clark B., Das S., Lo A. and Siddique A. (2016) Risk and risk management in the credit card industry, Journal of Banking and Finance, 72, pp. 218-239. https://doi.org/10.1016/j.jbankfin.2016.07.015

S. Neema and B. Soibam (2017) The comparison of machine learning methods to achieve most cost-effective prediction for credit card default, Journal of Management Science and Business Intelligence, pp. 36-41, Aug. 2017. http://www.ibii-us.org/Journals/JMSBI/

T. M. Alam and M. J. Awan, ‘‘Domain analysis of information extraction techniques,’’ Int. J. Multidisciplinary Sci. Eng., vol. 9, pp. 1–9, Jul. 2018.

H. Kim, H. Cho, and D. Ryu, ‘‘An empirical study on credit card loan delinquency,’’ Econ. Syst., vol. 42, no. 3, pp. 437–449, Sep. 2018. <https://doi.org/10.1016/j.ecosys.2017.11.003>