**Customer Credit Card Default** - Jesley Jacob

# LITERATURE REVIEW

# Introduction

Over the years, the delinquency rate on credit card payment is on a rise, as the result of which financial institutions like banks end up with a significant loss of money. Having a financial risk prediction model is crucial for financial institutions and this will help to provide insights about credit card holders who have higher probability to default based on their characteristics. Machine learning has been utilized in financial institutions to approach the data-driven industry in an improved manner to manage the real-time data generated customers. In the recent days, predictive analytics are allowing banks to experience a new horizon compared to the descriptive approach.

Huge financial loss would be incurred by the banks in the event of a bad loan such as a customer who defaults, thus this issue must be mitigated with risk prediction models. But at the same time, financial institutions do not want to lose the opportunity to earn a lot from those who pay back their loans on time. Hence, the assessment of whether a customer would be a credit risk or not must be done properly.

This project aims at making predictions on the credit card holders who are likely to default and the factors that are most significant with regards to correctly predicting the credit card defaulters. This would help the banking system in reducing the high delinquency rate and even making customers cautious of the factors that could affect their credit score.

To make these predictions, the Credit Card default dataset will be explored and the most outstanding features in the data would be determined, for predicting the probability of a customer to default in the next payment. Data will be preprocessed and models would be build using various machine learning classification algorithms. The performance of each model would be evaluated and the best predictive model would be selected.

**Literature Review**

The Credit card default dataset is an intensely researched topic. To acquire a better insight and understanding on the dataset and the given problem, a review of several research articles and publications focusing on the various approaches and algorithm for Credit card scoring was done.

Imbalanced class distribution has been frequently seen in the credit scoring domain. Owing to the fact that majority of the customers duly pay their credit card bills on time, a severe imbalance in the class attribute would appear which poses a major challenge. The paper [4], addresses the class imbalance problem by studying various resampling strategies such as Over-sampling and Under-sampling. Their performance was evaluated based on AUC measure, further to which the statistical significance of the differences between the average ranked performances were determined using the Nemenyi post hoc and the Friedman statistic test. The author emphasizes on the importance of tackling the imbalance problem before building the prediction model, since resampling techniques enhances the performance achieved, which in turn would be beneficial to banks and financial institutions.

In another study [5], Synthetic Minority Over-Sampling Technique (SMOTE) was used to deal with the imbalanced data, and a comparison of various data mining models using the original dataset and the dataset on with SMOTE was done. The author claims that, compared to the previous researches done in this domain, one of the highest classification accuracies of 89.01% was achieved while using Random forest along with SMOTE. The research showed how a group of simple data mining model with the right balancing technique proved better than using hybrid techniques based on complex algorithms.

To assess the accuracy for the credit card default, different Machine learning approaches are applied. *Shantanu Neema* and *Benjamin Soibam*, in their paper [6], compares the performance of seven classification algorithms: K-Nearest Neighbor, Linear Discriminant Analysis, Logistic Regression, Artificial Neural Networks, Decision Tree, Naïve Bayes Classifiers and Random Forest, based on Cost function and Mathew’s Correlation Coefficient (MCC). Resampling was done using Over-sampling, under-sampling, SMOTE and ROSE (Random oversampling). As per the research, Random forest among the different classification algorithms, resulted in the lowest cost and Artificial Neural network proved to be a better approach method when Matthew’s Correlation Coefficient was taken into consideration.

Many researches have shown the importance of feature selection, which is the process of reducing the features in the dataset by selecting only the relevant and most important variable for use during the modelling. Correlation-based Feature Selection (CFS) has been applied in the study [7] before the data analysis. Credit card default prediction was done using logistic regression, Rpart decision tree, and random forest, and their performance were evaluated. While assessing the credit risk of credit card customers, Random Forest outperformed the others, with an accuracy of 82 % and an Area under Curve of 77 %.

Recently, neural network are widely used for predicting models in the Credit card domain and is suitable for complex, large financial data due to its better adaptability and its ability to perform well in the classification simulation. This is a more complex and sophisticated approach than a simple linear regression model. The paper [8] researched on the impact of using neural network to determine the strong predictors of probability of default comparing with linear model. The authors highlighted on the powerful processing ability of neural networks compared to that of the traditional regression model.

After reviewing several articles, it has been decided that the machine learning classification algorithms that would be used for this project are logistic regression, random forest, SVM and neural networks. The target variable being a binary output field, the project of predicting customer default is a binary classification problem. One of the most widely used algorithm for binary classification is the Logistic Regression. It also gives an idea of the important variables, which will be useful for predicting credit card default. Random forest, as noted in many studies, outperforms decision tree. Before embarking in to data analysis, the resampling of the imbalanced data is done using Over-sampling, Under-sampling and SMOTE. Feature selection will be done based on the correlation between the customer features.

**Data Description**

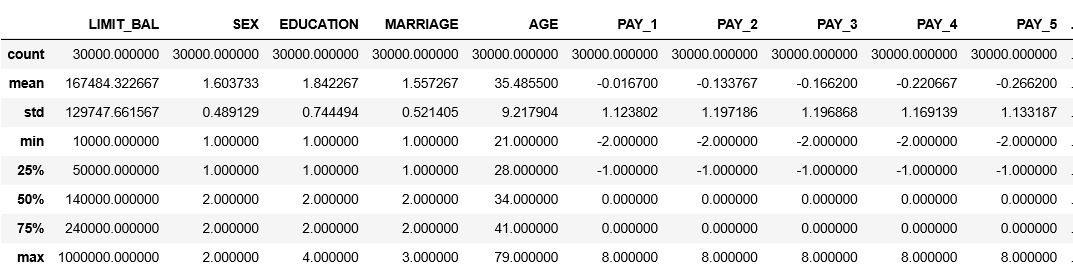
The dataset being used for this project is that of Taiwanese credit card holders from October 2005 used by [Yeh & Lien (2009)](https://pdfs.semanticscholar.org/1cac/ac4f0ea9fdff3cd88c151c94115a9fddcf33.pdf). This dataset was posted and made available to the on the UCI Machine Learning Repository, Center for machine learning and intelligent Systems.[(https://archive.ics.uci.edu/ml/datasets/default%20of%20credit%20card%20clients](https://archive.ics.uci.edu/ml/datasets/default%20of%20credit%20card%20clients)) .

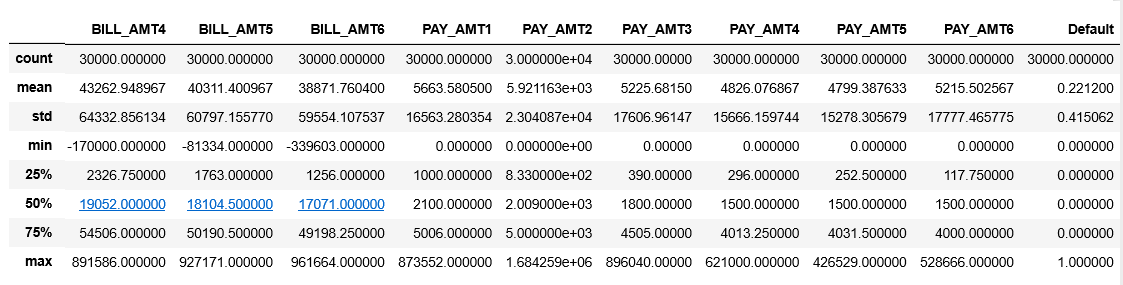
It contains **30000 observations** each of which corresponds to an individual credit card holder; **and 24 variables**. The 24 variables in this dataset are of type **numerical**, and consists of the customers demographic information (gender, education level, marriage status, and age) and financial variables of 6-months’ worth of payment data from April 2005 to September 2005 (amount of given credit, monthly amount of previous payments, monthly amount of bill statements and monthly repayment statuses). The breakdown of each attribute is given below.

|  |  |
| --- | --- |
| **Name** | **Description** |
| ID | ID of each client |
| LIMIT\_BAL | Amount of given credit in NT dollars (includes individual and family/supplementary credit) |
| SEX | Gender (1=male, 2=female) |
| EDUCATION | (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown) |
| MARRIAGE | Marital status (1=married, 2=single, 3=others) |
| AGE | Age in years |
| PAY\_0 | Repayment status in September, 2005 (-2=no consumption, -1=pay duly, 0=the use of revolving credit, 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 (scale same as above) |
| PAY\_3 | Repayment status in July, 2005 (scale same as above) |
| PAY\_4 | Repayment status in June, 2005 (scale same as above) |
| PAY\_5 | Repayment status in May, 2005 (scale same as above) |
| PAY\_6 | Repayment status in April, 2005 (scale same as above) |
| BILL\_AMT1 | Amount of bill statement in September, 2005 (NT dollar) |
| BILL\_AMT2 | Amount of bill statement in August, 2005 (NT dollar) |
| BILL\_AMT3 | Amount of bill statement in July, 2005 (NT dollar) |
| BILL\_AMT4 | Amount of bill statement in June, 2005 (NT dollar) |
| BILL\_AMT5 | Amount of bill statement in May, 2005 (NT dollar) |
| BILL\_AMT6 | Amount of bill statement in April, 2005 (NT dollar) |
| PAY\_AMT1 | Amount of previous payment in September, 2005 (NT dollar) |
| PAY\_AMT2 | Amount of previous payment in August, 2005 (NT dollar) |
| PAY\_AMT3 | Amount of previous payment in July, 2005 (NT dollar) |
| PAY\_AMT4 | Amount of previous payment in June, 2005 (NT dollar) |
| PAY\_AMT5 | Amount of previous payment in May, 2005 (NT dollar) |
| PAY\_AMT6 | Amount of previous payment in April, 2005 (NT dollar) |
| default.payment.next.month | Default payment (1=yes, 0=no) |

**Descriptive Statistics:**

The descriptive statitics were done on the dataset for each attribute in order to get better insights from this data.





# Approach

**Data Preparation**

**Exploratory Data Analysis**

**Feature Selection/Engineering**

**Train the Model**

**Test & select the Model**

**Prediction & Conclusion**

Figure 1

**Step 1 : Data Preprocessing**

In this step, the data is loaded, after which the data cleaning and the analysis of the attributes are done.

1. **Missing Values:**

The dataset was checked for missing values and was found to not have any missing values.

1. **Data Cleaning:**
2. The *ID* field was removed from the dataset
3. The field *default.payment.next.month* was renamed to *Default*
4. The field *PAY\_0* was renamed to *PAY\_1*
5. The unknown values (0, 5, and 6) in the *Education* field as shown in the fig 2.1, were changed to 4 (Others).
6. The unknown values (0) in the *Marriage* field as shown in the fig 2.2, were changed to 3 (Others).

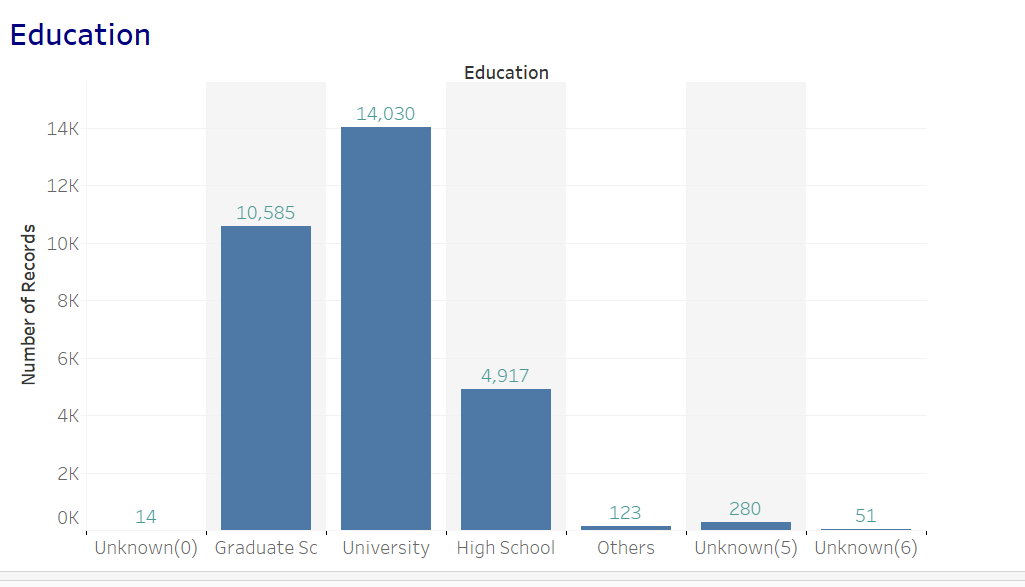


Figure 2.1

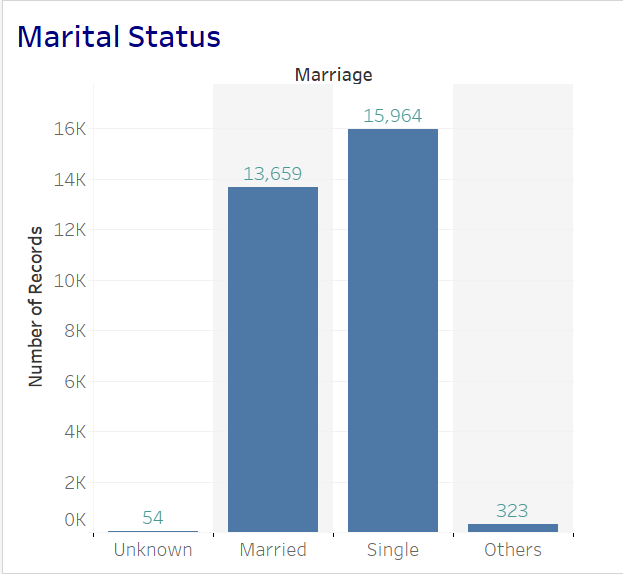


Figure 2.2

1. **Split Dataset into Training & Test sets:**

The dataset set is split into training set and test set with 70:30 ratio respectively.

1. **Data Imbalance:**

There is a class imbalance in the dataset, with customers who will default at 6636 (which comprises of 22% of the records) and customers who do not default are 23364 (as shown in the fig 3).

Machine learning algorithms will most likely to make incorrect predictions when there is a class imbalanced issue; since the minority class would be treated as noise and there would be a bias towards the majority class. Different resampling techniques such as Over-sampling, Under-sampling and SMOTE will be done on the dataset.

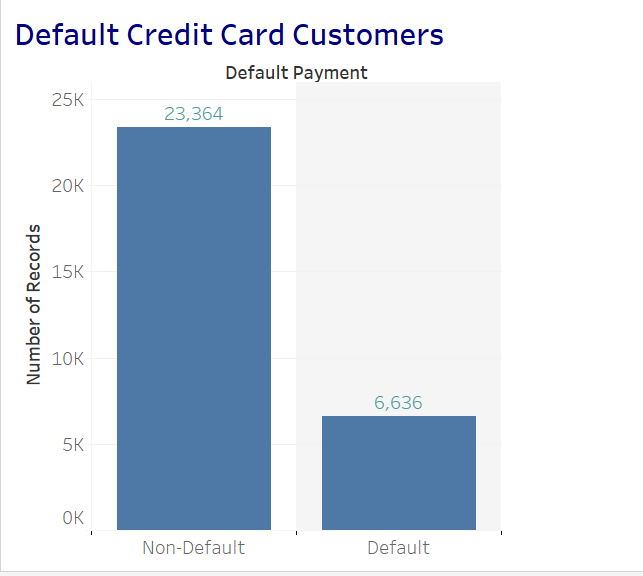


Figure 3

1. **Outliers:**

Variables that are measured in currency (credit limit, amounts billed, amounts paid) were highly skewed, having many large outliers. The distribution of Amount paid in April 2005, which is highly right skewed is shown in the box plot in fig 4. Unambiguously erroneous values were not found in the outliers. The data would be normalized using standardization.

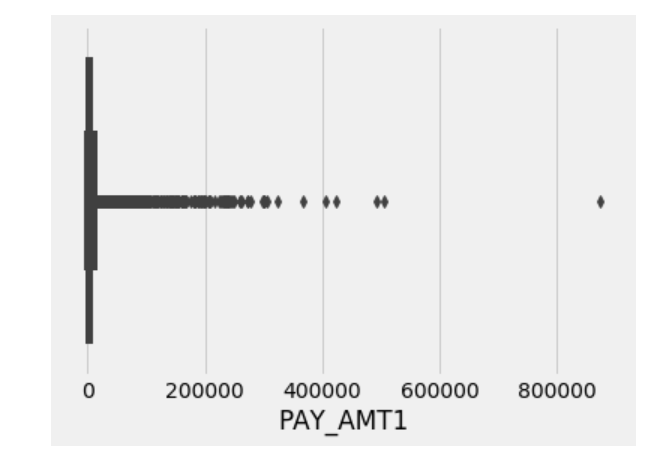


Figure 4

**Step 2 : Exploratory Data Analysis**

This stage will involve uni-variant, bi-variant and multi-variant analysis of the attributes. And data visualizations would also be used to show the relations between the various attributes.

**Step 3 : Feature Selection/Feature Engineering**

In this step, feature engineering would be done which would come up with new features given the dataset that would probably be better predictors of credit card defaulters The features that are of least importance will be removed, and this would in turn increase the performance of the models we build in the next step. One of the methods used for this would be the Correlation Matrix with heat map (given in the fig 5)

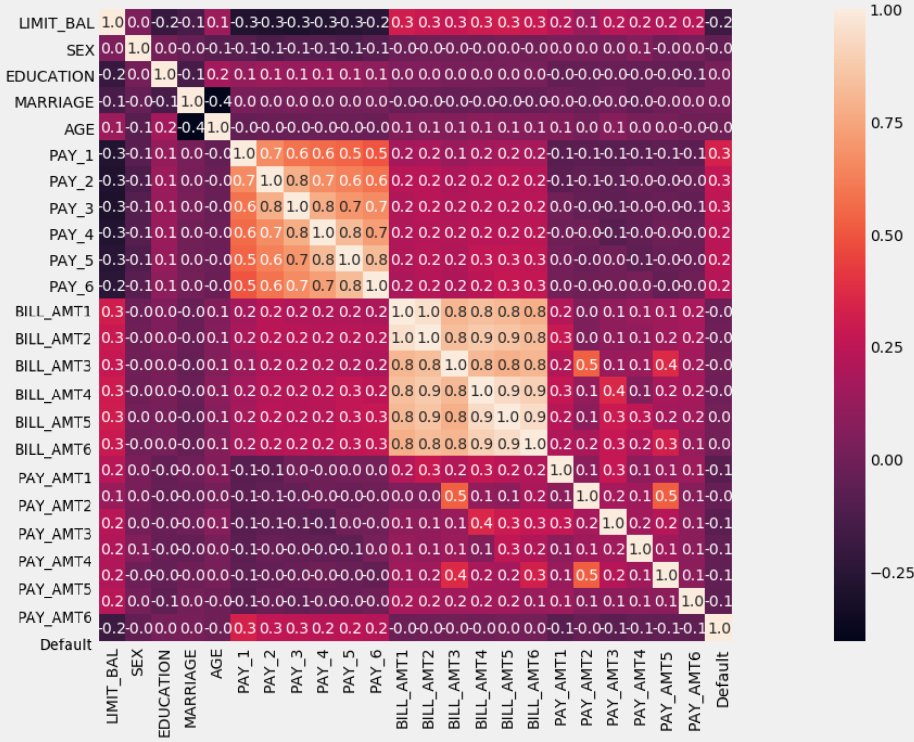


Figure 5

**Step 4: Train the Models**

In this stage, different models are build using the features selected during the previous phase on the train data. The classification algorithms such as Logistic Regression, Random Forest, SVM and Neural Networks will be applied.

**Step 5: Test and Select Model**

The models built in using the classification algorithms in the previous step, will be used on the test dataset, and their performance will be measured. Performance metrics based in which the models would be evaluated are Accuracy, F1-score, TPR, FNR and ROC. The model with the best performance would be selected as the best predictive model.

**Step 6: Prediction/Conclusion**

Finally, conclusions and inferences would be drawn based on the built model.

**REFERENCES:**

[1] [Xian Jin Seow](https://towardsdatascience.com/@xianjinseow92?source=post_page-----f4b21547a618----------------------) ‘*Catching a Welcher: Classifying a Credit Card Defaulter* ‘ <https://towardsdatascience.com/catching-a-welcher-classifying-a-credit-card-defaulter-f4b21547a618>

[2] [Jason Brownlee](https://machinelearningmastery.com/author/jasonb/) “An Introduction to Feature Selection by Jason Brownlee” <https://machinelearningmastery.com/an-introduction-to-feature-selection/>

[3] Raheel Shaikh “Feature Selection Techniques in Machine Learning with Python”

<https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>

[4] [A I Marqués](https://www-tandfonline-com.ezproxy.lib.ryerson.ca/author/Marqu%C3%A9s%2C+A+I), [V García](https://www-tandfonline-com.ezproxy.lib.ryerson.ca/author/Garc%C3%ADa%2C+V) & [J S Sánchez](https://www-tandfonline-com.ezproxy.lib.ryerson.ca/author/S%C3%A1nchez%2C+J+S) “On the suitability of resampling techniques for the class imbalance problem in credit scoring”

<https://www-tandfonline-com.ezproxy.lib.ryerson.ca/doi/full/10.1057/jors.2012.120>

[5] [Abdulhamit Subasi](https://ieeexplore-ieee-org.ezproxy.lib.ryerson.ca/author/38229988300)  & [Selcuk Cankurt](https://ieeexplore-ieee-org.ezproxy.lib.ryerson.ca/author/37085905435) “Prediction of default payment of credit card clients using Data Mining Techniques” (2017)

<https://ieeexplore-ieee-org.ezproxy.lib.ryerson.ca/document/8950597>

[6] Shantanu Neema & Benjamin Soibam “The comparison of machine learning methods to achieve most cost-effective prediction for credit card default” (2017)

<https://pdfs.semanticscholar.org/def0/9bf8f23bd163c41c22b4ceeff64784ea28b6.pdf>

[7] Credit Card Default Prediction using Machine Learning Techniques (2018) <https://ieeexplore.ieee.org/document/8776802>

[8] Bu-yun ZHANG, Shi-wei LI and Chuan-tao YIN “A Classification Approach of Neural Networks for Credit Card Default Detection” (2017)

<http://dpi-proceedings.com/index.php/dtcse/article/view/12303/11840>

[9] [I-ChengYeh](https://www.sciencedirect.com/science/article/abs/pii/S0957417407006719" \l "!) & [Che-huiLienb](https://www.sciencedirect.com/science/article/abs/pii/S0957417407006719#!) “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients”