**Customer Credit Card Default** - Jesley Jacob

# FINAL REPORT

# 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.

1. **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.

1. **Data Description**

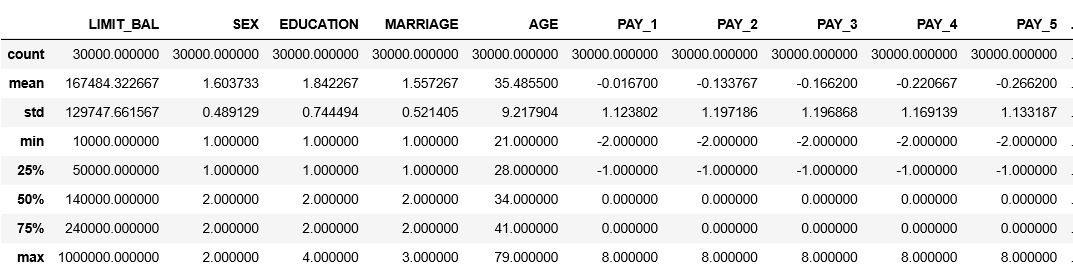
The dataset being used for this project is that of Taiwanese credit card holders from October 2005 used by [Yeh & Lien (2009)](https://pdfs.semanticscholar.org/1cac/ac4f0ea9fdff3cd88c151c94115a9fddcf33.pdf). This dataset was posted and made available to the public on the UCI Machine Learning Repository, 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)) .

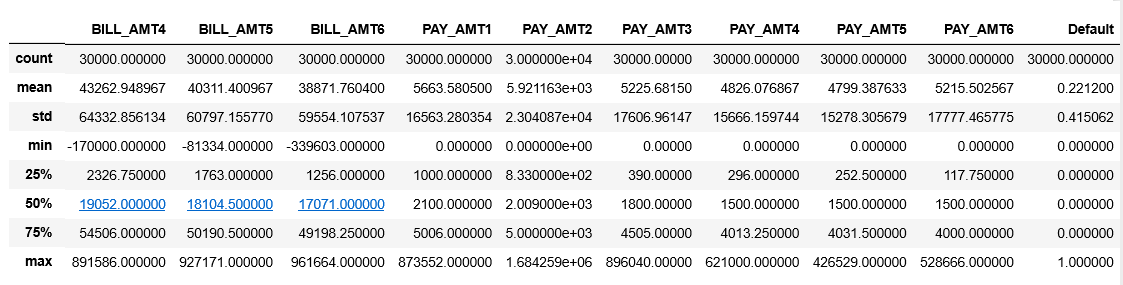
It contains **30000 observations** each of which corresponds to an individual credit card holder; **and 24 variables**. The 24 variables in this dataset are of type **numerical**, and consists of the customers demographic information (gender, education level, marriage status, and age) and financial variables 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

The approach for the project would be as outlined in the figure 4.1.

**Data Preparation**

**Exploratory Data Analysis**

**Feature Selection/Engineering**

**Train the Model**

**Test & select the Model**

**Prediction & Conclusion**

Figure 4.1

1. **Data Preprocessing and Exploratory Data Analysis**

**5.1 Data Cleaning**

For this stage, the data is loaded, after which the data cleaning and the analysis of the attributes are done. The dataset was checked for missing values and was found to not have any missing values. However, there are few columns which that needed modifications, such as the ID column has been removed since it is considered irrelevant and the attributes *default.payment.next.month and PAY\_0* were renamed. In order to make more sense of the data, attributes such as Marriage, Education, and Payment Status were revised to merge the unknown values.

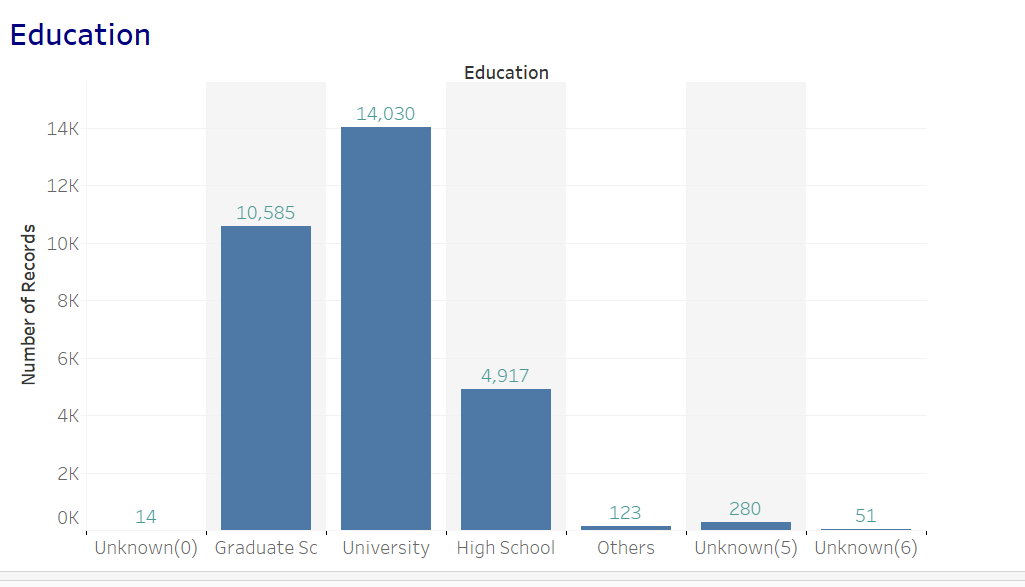


Figure 5.1.1

As shown in Fig 5.1.1, the Education variable has unknown category values (0, 5 and 6), hence those were merged to the *Others* category. Similarly the unknown value 0 in the Marriage attribute was added to the *Others* category (as in Fig 5.1.2). In the case of the Repayment status attributes, the values -2 (no consumption) and -1 (pay duly) were merged to the category level 0 (the use of revolving credit).

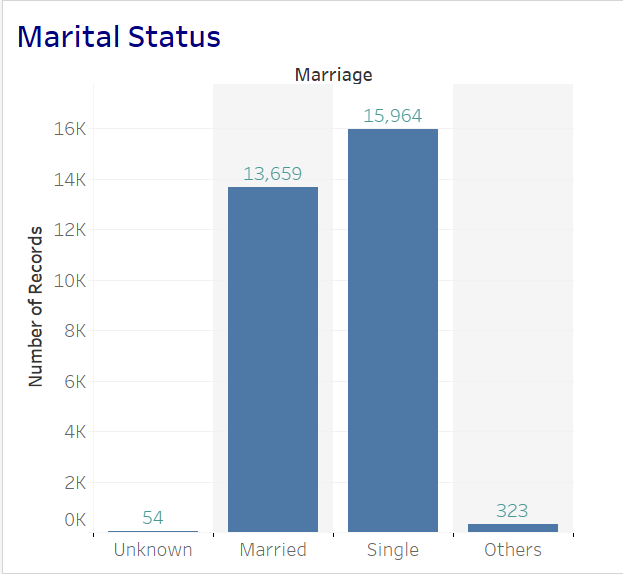


Figure 5.1.2

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 5.1.3. Unambiguously erroneous values were not found in the outliers; hence it will not be removed from our dataset.

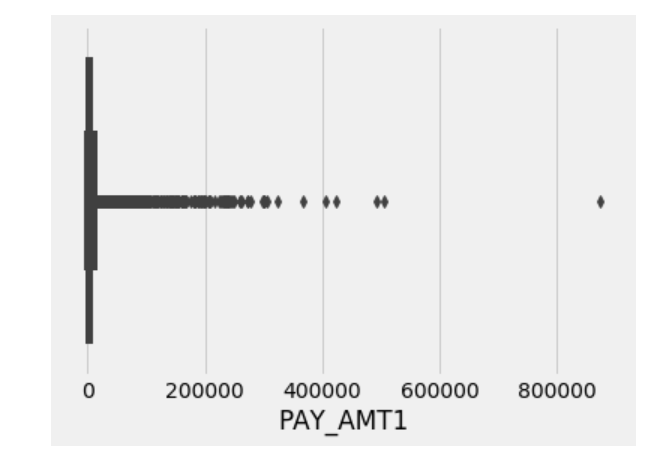


Figure 5.1.3

The data would be normalized using standardization since the features are in different scales and units as shown in the Fig 5.1.4. It is important to normalize the date since many models use some form of distance when making predictions.

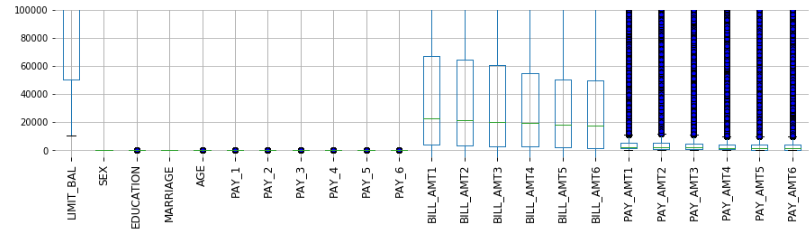


Figure 5.1.4

1. **Exploratory Data Analysis**

**6.1 Uni-variant, Bi-variant and Multi-variant Analysis**

In this step, we try to analyze the features in our dataset to find thefactors that are most significant with regards to correctly predicting the credit card defaulters. The dataset has a class imbalance, 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 6.1.1).

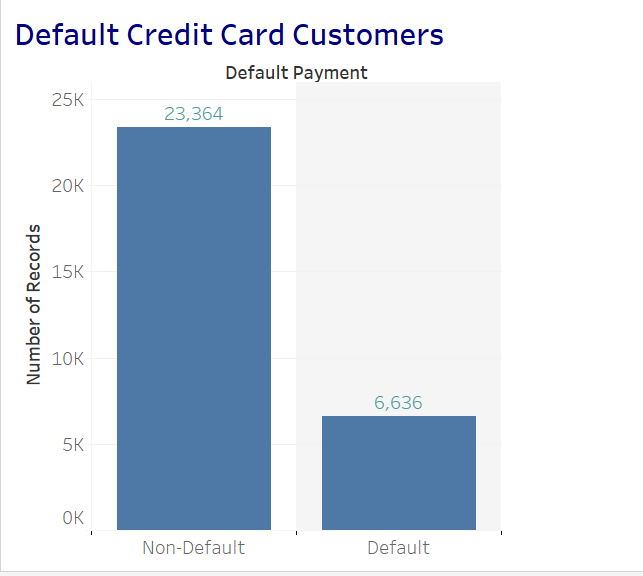


Figure 6.1.1

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.

As per the data,there is a higher number of female customers. However the percentage of a male customers who default is around 24%, which is higher than their counterpart whose percentage is around 20%.

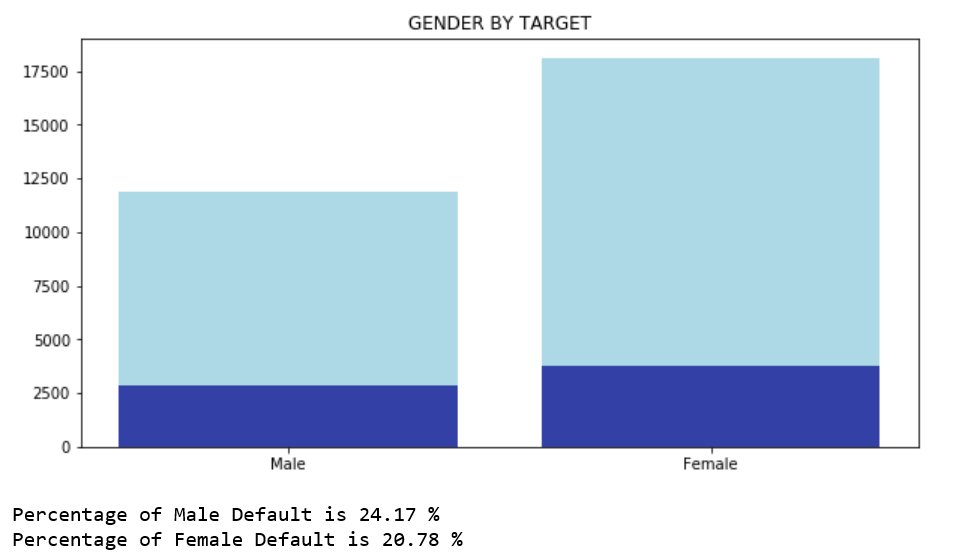
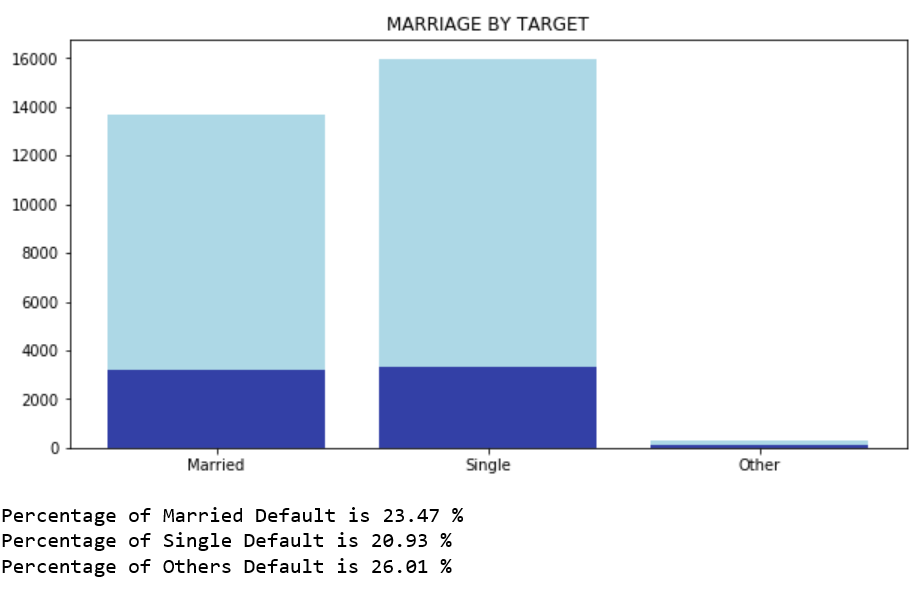
 

Figure 6.1.2 Figure 6.1.3

In the Marriage category, there are more customers who are Single, and they have a slightly lower chance of defaulting than people with the *Married* or *Other* status. There were few unknown values in the Education category, which were all merged to the *Others* value. Higher the education level, lower the chances of defaulting on the payment looks like.

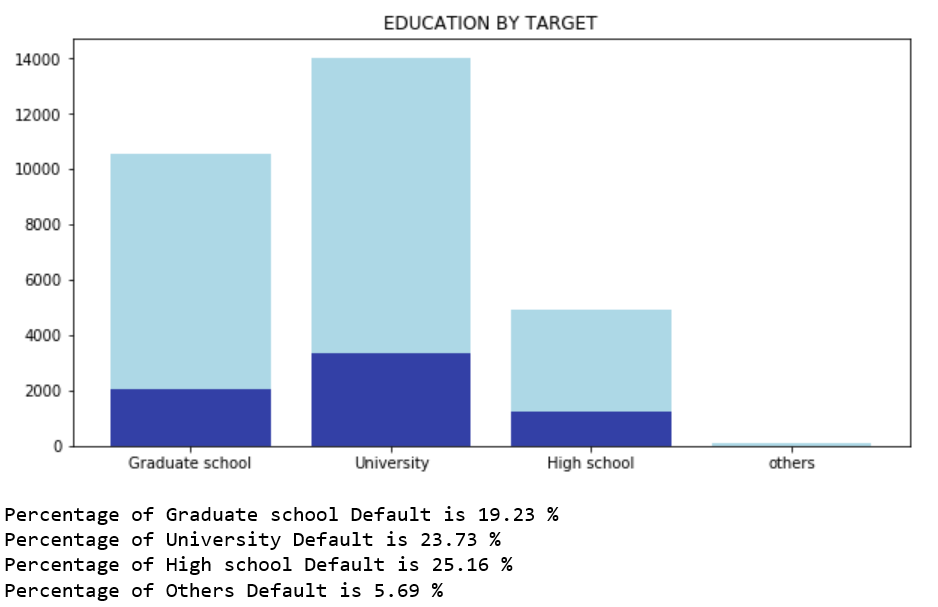


Figure 6.1.4

With respect to the age of the customers, more than 50% seems to be those of the ages between 25 and 40, and they appear to have a lower chance of defaulting on their payment

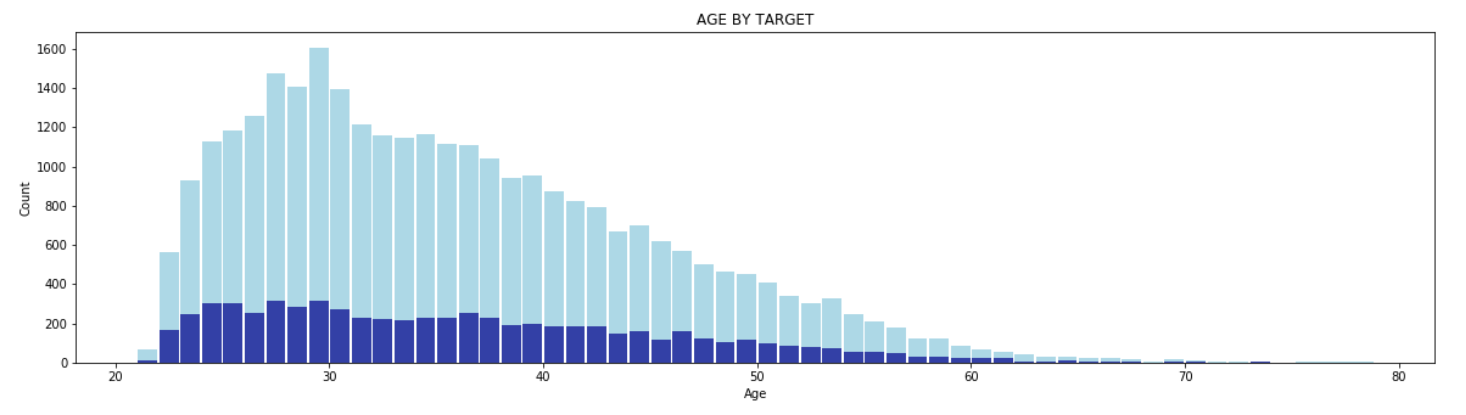


Figure 6.5

The repayment status features appear to be the highest influencers in predicting whether a customer would default his/her next payment. The values -2 and -1 were modified to be 0 for all the repayment status attributes, since they all indicate that the customer have paid the dues. As expected from the fig 6.7, those who have paid duly appears to have a very low probability of defaulting compared to those who have delayed on their payment. Especially those who have defaulted their payments for 2 or more months seems to be having more than 50% probability of defaulting the next month. From further analysis, PAY\_1(Repayment status in September) has the highest correlation with the class variable.

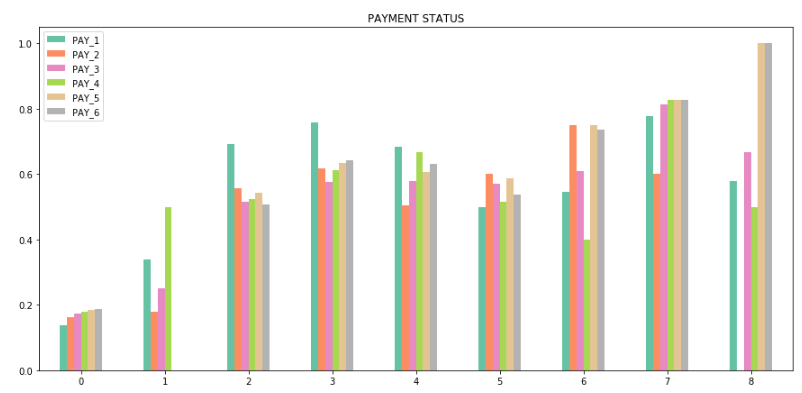


Figure 6.7

In the case of Balance Limit, those with a lower credit limit have higher probability to default compared to those who have higher credit limits (as seen in Fig 6.8)

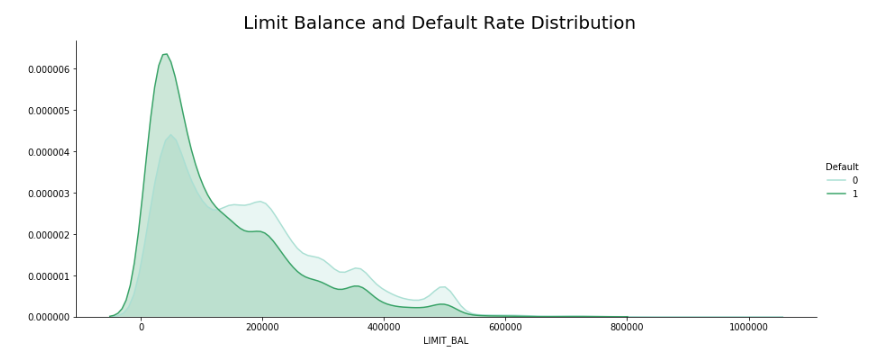


Figure 6.8

Further analysis and Data visualizations were done on the data set, such as the figure below, which shows the relation between the age, marriage, sex, education and credit balance limit.

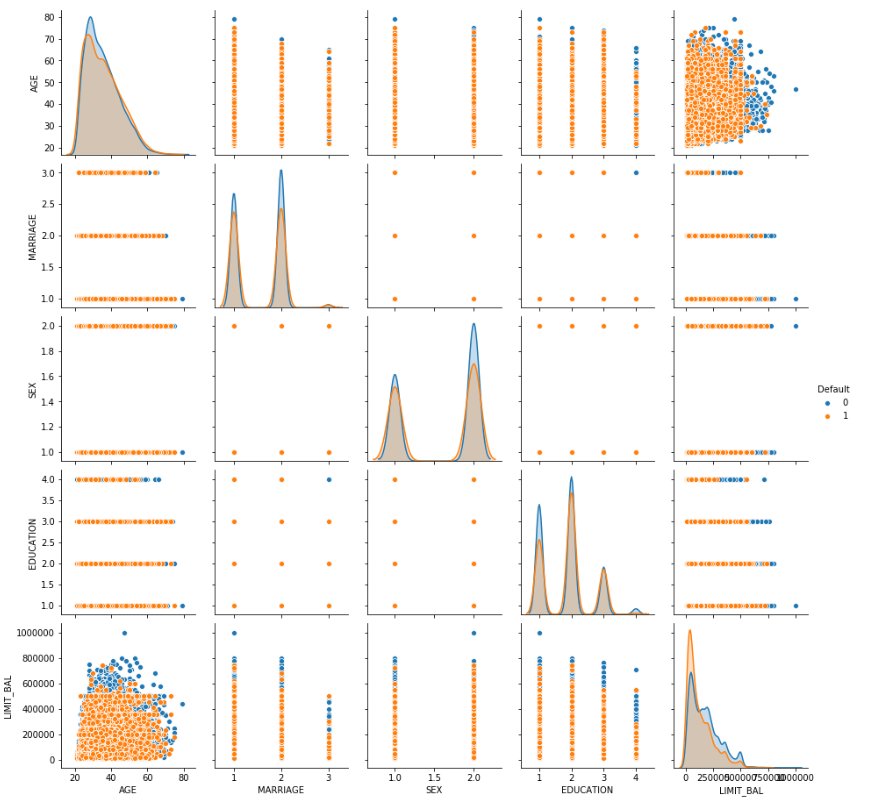


Figure 6.9

* 1. **Correlation**

A heatmap with the correlation matrix of all features in the dataset was created to get an understanding of the relationship among the various features.

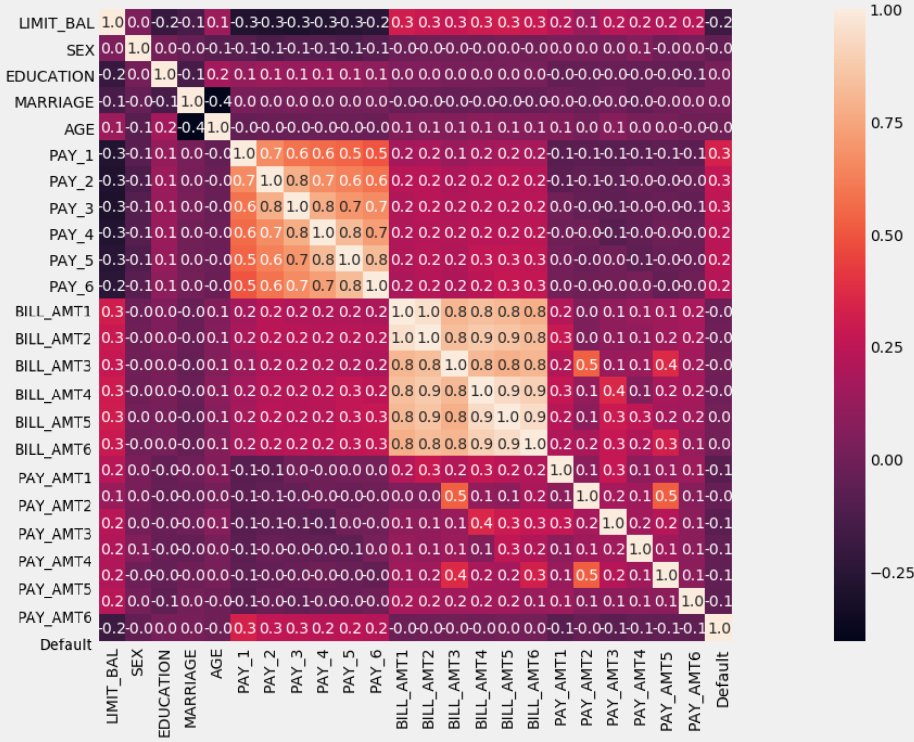


Figure 6.2.1

As we can see clearly from Figure 6.2.2, the Repayment status for the month of September (PAY\_1) is the most influential feature with a high positive correlation with the class variable (Default). Where the credit limit feature seems to be having the highest negative correlation with the class variable. Hence the customers who have a lower credit limit are more probable to default.

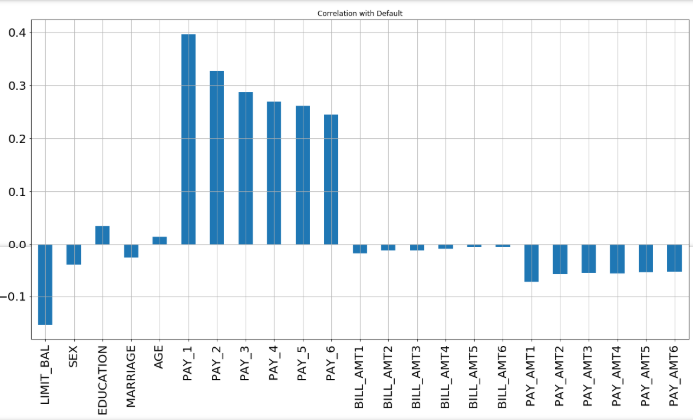
****

Figure 6.2.2

1. **Feature Engineering**

One hot encoding was performed on the categorical variables in the dataset (SEX, MARRIAGE and EDUCATION) before embarking to predictive analysis, since many algorithms in machine learning works only with numeric values and not with categorical data. This is done by converting the categorical variables into numeric values, which is done on both input and output categorical variables. This is required for both input and output variables that are categorical. The variables values are mapped to numeric values and then represented as a binary vector, where except for the index of the integer which will be marked as 1, rest would be marked as 0. In our dataset, the categorical values are already numeric, hence only the encoding to binary values of 0 and 1. An example of how the one hot encoding is done on the SEX variable in our dataset is shown the Fig 7.1.

|  |  |  |  |
| --- | --- | --- | --- |
| SEX |  | MALE | FEMALE |
| 1 |  | 1 | 0 |
| 2 |  | 0 | 1 |
| 2 |  | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |
| 1 |  | 1 | 0 |
| 2 |  | 0 | 1 |

Figure 7.1

1. **Predictive Analysis**

**8.1 Data Partitioning:**

The dataset set is split into training set and test set with 70:30 ratio respectively, which was commonly used in many of the previous literature on the same dataset. It has proved to be a good proportion for building classification models. Hence, the models would be trained on the training dataset and tested on the test dataset which was not seen by the algorithm before.

**8.2 Rescaling**

As seen earlier, we will do standardization on the data since the features are of different units and scales (Fig 5.1.4). Hence it is recommended to rescale the data in order to enhance the performance of the models since some algorithms use some form of distance to make predictions. This involves transforming the values of each variable to have mean as zero and unit-variance.

**8.3 Machine Learning Algorithms**

The data mining process involves various techniques in order to generate new information and derive possible recommendations. 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.

**Logistic Regression** is Supervised Machine Learning (ML) methods which is dedicated for Classification. For the last two decades, it has a great reputation especially in financial sector due to its ability in detecting defaulters. **Support Vector Machine (SVM)** is widely used in classification problems. It aims at finding a hyperplane in an N-dimensional space, where N is the number of features that classifies the data points distinctly.

**Random forests** are an ensemble learning method used for classification and regression consists of a large number of individual decision trees that operate as an ensemble. The accuracy of these models tends to be beat most of the other [decision trees](http://en.proft.me/2016/11/9/classification-using-decision-trees-r/). **Multilayer Perceptrons, (MLP),** one of the classic type of neural network, which consists of one or more layers of neurons. There is an input layer, to which we feed the data, and the output layers (visible layer) makes predictions. In between these layers there will be one or more hidden layers providing levels of abstraction. For our dataset, we will be using 3 hidden layers.

The scaled training set is used to train the models and along with the 5-fold cross validation, and the performance are recorded for all the algorithms, which is given in the Table 8.3.1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **IMBALANCED DATA** | | | | | |
|  | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **ROC** |
| **Logistic Regression** | 0.8188 | 0.6782 | 0.3031 | 0.419 | 0.6318 |
| **SVM** | 0.8215 | 0.6677 | 0.3428 | 0.453 | 0.648 |
| **Random Forest** | 0.817 | 0.6348 | 0.3557 | 0.4559 | 0.6497 |
| **ANN** | 0.8169 | 0.6524 | 0.3222 | 0.4313 | 0.6375 |

Table 8.3.1

We received a good accuracy overall for the models. However, in an imbalanced dataset though the accuracy values might seem good, the accuracy of default=Yes would be less than 50% accurate. So I will be looking in to the recall scores as well for this imbalanced set in order to see whether our model is actually good. The recall scores in this case, is very low, hence we can’t consider these as good models.

The data would be balanced using resampling technique and model will be built.

**8.4 Resampling**

As seen earlier, there dataset has an imbalanced class. In order to handle this we will be using various Resampling techniques such as:

**a. Random Under-sampling:** The existing instances of the majority class is reduced/eliminated in this case randomly. This technique can eliminate information or data points that could be useful for the classification algorithm hence it is not always the best approach.

**b. Random Over-sampling:** In this case, new observations of the minority class are created or duplicated randomly. This could lead to overfitting however, since existing instances of the minority class are replicated.

**c. SMOTE (Synthetic Oversampling): In** this case, a small subset of minority is chosen and synthetic examples of this subset are created to balance up the overall dataset, this approach avoid overfitting. The overall number of observations are increased by adding new information.

Models are trained using the balanced training data set, and the performances are evaluated for each sampling technique and each algorithm, and the results are shown in the Table 8.4.1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Accuracy** | **Recall** | **F1 Score** | **ROC** |
| **RANDOM UNDER-SAMPLING** | | | | |
| **Logistic Regression** | 0.7854 | 0.5582 | 0.5287 | 0.703 |
| **SVM** | 0.7767 | 0.5856 | 0.5306 | 0.7074 |
| **Random Forest** | 0.7444 | 0.627 | 0.5143 | 0.702 |
| **ANN** | 0.7687 | 0.566 | 0.5133 | 0.6952 |
|  |  |  |  |  |
| **RANDOM OVER-SAMPLING** | | | | |
| **Logistic Regression** | 0.782 | 0.5655 | 0.528 | 0.7035 |
| **SVM** | 0.7772 | 0.5737 | 0.5261 | 0.7034 |
| **Random Forest** | 0.8097 | 0.4103 | 0.4817 | 0.6649 |
| **ANN** | 0.7603 | 0.6026 | 0.5202 | 0.7031 |
|  |  |  |  |  |
| **SMOTE** | | | | |
| **Logistic Regression** | 0.7822 | 0.5691 | 0.5297 | 0.7049 |
| **SVM** | 0.7787 | 0.5769 | 0.5290 | 0.7054 |
| **Random Forest** | 0.7924 | 0.4572 | 0.4869 | 0.6708 |
| **ANN** | 0.7517 | 0.6082 | 0.5136 | 0.6996 |

Table 8.4.1

The overall performance has increased, with the recall being almost double the score we received on the imbalanced dataset. Hence we can take into account the Accuracy and the AUC-ROC values as our performance metrics. Here, the performance is similar for all the models.

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

This curve plots two parameters:

• True Positive Rate /Recall: it is defined as TPR = TP / (TP+FN)

• False Positive Rate: it is defined as FPR = FP/ (FP+TN)

In our case, the TP, TN, FP and FN can be defined as the below:

* True Positive : A defaulter is predicted as a defaulter
* True Negative: A non-defaulter is predicted as a non-defaulter
* False Positive: A Non defaulter is predicted as a defaulter
* False Negative: A defaulter is predicted as a non-defaulter

AUC - ROC curve is the area under the ROC curve. This score gives us a good idea of how well the model performs. It helps in interpreting how much a particular model is capable of distinguishing between classes, in a classification problem. Higher the AUC value better is the model at predicting the correct classes. Hence, higher the AUC, better the model is at distinguishing between customers who would make the payment the next month and those who would default.

The time taken for training and testing each of the algorithms were calculated and shown in the Figure 8.4.1a and Figure 8.4.1b. The SVM is utilizes high time for training the model. Hence, we can ignore this algorithm for the rest of the experiments with the data.

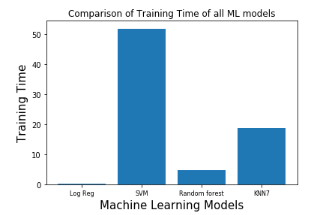


Figure 8.4.1a

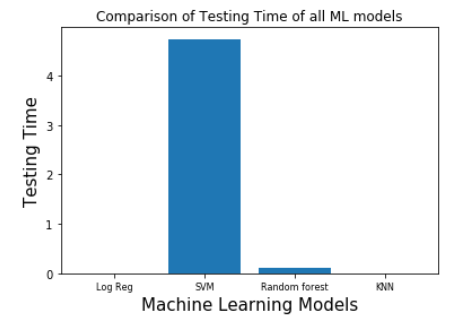


Figure 8.4.1b

In order to try and enhance the performance of my models, additional approaches such as feature selection with Recursive Feature Elimination (using cross validation) and optimization of the models using GridSearchCV was done.

**8.5 Recursive Feature Elimination**

**Recursive Feature Elimination (RFE)** is used for feature selection by recursively considering smaller and smaller sets of features. This process is to construct a model repeatedly and setting aside either the best or worst performing feature, and then repeating the process with the rest of the features, until all features in the dataset are exhausted. The goal of RFE is to select features by recursively considering smaller and smaller sets of features.

**8.6 Model Optimization using GridSearchCV**

**GridSearchCV** is a method to exhaustively search a models’ best parameters from a list of given parameters. The parameters and the model for search need to be specified for this cross-validation search method. It aims at finding the best parameter for the dataset and the target model.

The recursive feature selection and GridsearchCV was done and their performance was recorded. The results seemed is be similar to the previous model with very slight changes. Neural Networks got an accuracy of 82% along with Random Forest also having around 82% accuracy. The ROC curve of 71% was achieved for Neural Networks (as shown in Figure 8.6.1) which was oversampled with SMOTE and GridSearchCV was applied.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RECURSIVE FEATURE ELIMINATION** | | | | |
|  | Accuracy | Recall | F1 score | ROC |
| Random Forest | 0.8187 | 0.3659 | 0.4654 | 0.6545 |
| Logistic Regression | 0.7727 | 0.5778 | 0.5229 | 0.7020 |
| Logistic Regression UnderSampled | 0.7858 | 0.5587 | 0.5294 | 0.7035 |
| Logistic Regression OverSampled | 0.7738 | 0.5768 | 0.5237 | 0.7024 |
| **GRIDSEARCH** | | | | |
| **Logistic Regression** | | | | |
| Logistic Regression | 0.8187 | 0.3030 | 0.4189 | 0.6317 |
| Logistic Regression - UnderSampled | 0.7855 | 0.5587 | 0.5290 | 0.7033 |
| Logistic Regression - OverSampled | 0.7818 | 0.5675 | 0.5286 | 0.7041 |
| Logistic Regression - SMOTE | 0.7822 | 0.5690 | 0.5297 | 0.7049 |
| **Random Forest** | | | | |
| Random Forest | 0.819 | 0.3582 | 0.4604 | 0.6519 |
| Random Forest - UnderSampled | 0.7492 | 0.6128 | 0.5130 | 0.6997 |
| Random Forest - OverSampled | 0.8106 | 0.4309 | 0.4952 | 0.6729 |
| Random Forest -SMOTE | 0.7924 | 0.4701 | 0.4940 | 0.6755 |
| **Neural Networks** | | | | |
| Neural Networks | 0.8238 | 0.3793 | 0.4815 | 0.6627 |
| Neural Networks- UnderSampled | 0.7686 | 0.6015 | 0.5285 | 0.7080 |
| Neural Networks -OverSampled | 0.7717 | 0.5979 | 0.5304 | 0.7087 |
| Neural Networks - SMOTE | 0.7704 | 0.6077 | 0.5330 | 0.7114 |

Table 8.6.1

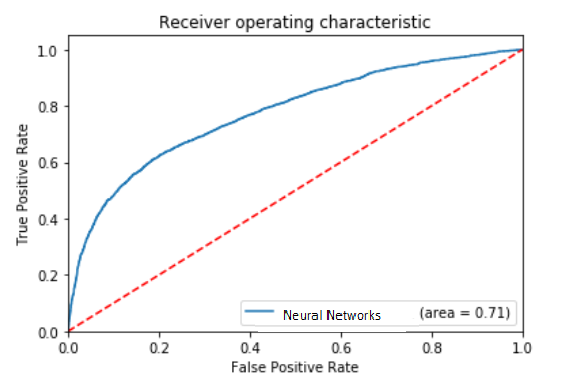


Figure 8.6.1

Further techniques were tried in order to enhance the performance of the models, such as the below, and the model performances was recorded for each of the techniques:

* Data partition with 80:20 ratio (so as to get more training data)
* Removed the outliers from the dataset
* Feature importance was evaluated for the Random Forest (as shown in Figure 8.6.2), and the models were trained after all the features of low importance were removed.
* Features for each month were removed on a monthly basis. For example, the repayment status, bill amount and repayment amount for the month April was removed, and the models were build. Likewise, the features for the consecutive month were removed, and models were build.

However, the performance didn’t improve with any of the above techniques.

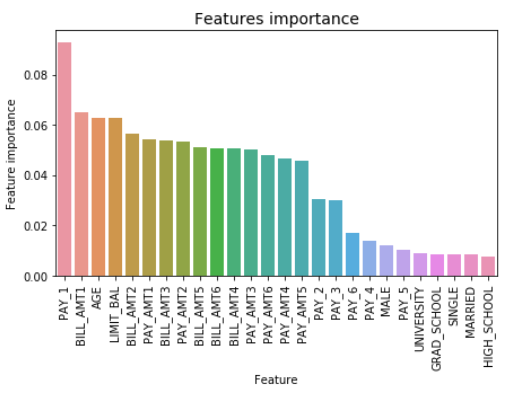


Figure 8.6.2

1. **Prediction/Conclusion**

This project was able to make 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.

The factors that are most significant with regards to correctly predicting the credit card defaulters are the repayment status history variables especially the repayment status for September (PAY\_1) and the amount of credit limit. Those who paid their bill duly were less than 20% likely to default the next month, while those who delayed payment by 2 or more months were more than 50% likely to default the next month. With respect to the credit limit, the lower the limit the higher probability for those customers to default. Female customers and those with higher education levels were less likely to default.

Using this information, I would like to recommend that the repayment status history for the customers must be monitored, especially if someone failed to pay their dues during the earlier months. And customers must be educated about how they can stay cautious from defaulting their payments.

Four machine learning methods were compared. Various techniques were utilized to enhance the performance of the models, such as standardization, resampling to treat imbalance (Random Under-Sampling, Random Over-Sampling, SMOTE), Recursive Feature elimination for feature selection and GridSearchCV for optimisation of the models. Neural Networks model gave an accuracy of 82% and 71% for AUC for the original data and with the synthetic oversampled data on which GridSearchCV was applied. Random forest also had an accuracy of almost 82%. Logistic Regression was also able to come up with similar scores overall. This project was able to match the accuracy scores that was achieved in previous works on the same dataset. As next steps, I would like to further experiment on the data using PCA or LDA and TensorFlow.

**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”

**APPENDIX:**

**GitHub**: <https://github.com/Engaged2Data/Capstone-Project>