# CREDIT CARD CUSTOMER DEFAULT PREDICTION

**Team**: Shreya Krishna Hegde (191046024), Deeksha Dsouza (191046040)

**Guide**: Sudarsan N S Acharya

SCHOOL OF INFORMATION SCIENCE
MANIPAL
*(A constituent institution of MAHE, Manipal)*

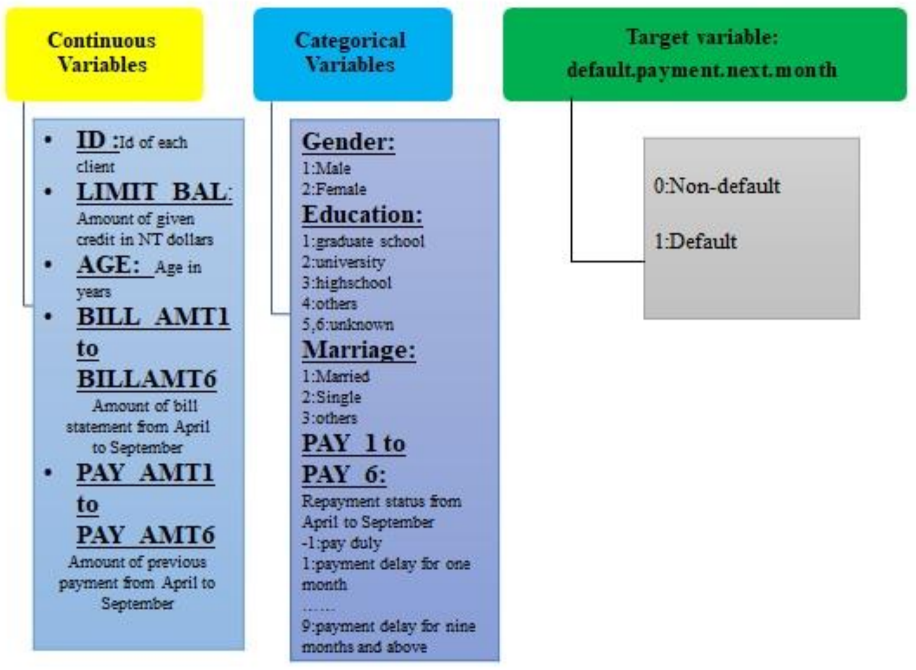
## Overview

Credit card default prediction is useful for the banks or other financial institutes to deal with defaulters. The main goal of this project is to predict the customers who will become defaulter customer for the upcoming month. In this model scores obtained from GLRM for each cluster is given as input and the customer classified as the defaulter or non-default is the output. Different correlation measures are applied to the features based on its type, the features which are highly correlated are clustered with the help of hierarchical clustering, Feature dimensionality reduction is done with the help of GLRM. The reduced features are given as the input to different binary classification algorithms and predicted the customer who is defaulter.

## Dataset

Dataset is downloaded from UCI machine learning repository.This credit card dataset is collected from credit card clients in Taiwan from April 2005 to September 2005, which contains 30,000 samples and 25 features.

All the features are listed in the image below:



Continuous Variables

- **ID :**Id of each client
- **LIMIT_BAL:** Amount of given credit in NT dollars
- **AGE:** Age in years
- **BILL_AMT1 to BILLAMT6** Amount of bill statement from April to September
- **PAY_AMT1 to PAY_AMT6** Amount of previous payment from April to September

Categorical Variables

**Gender:**
1:Male
2:Female
**Education:**
1:graduate school
2:university
3:highschool
4:others
5,6:unknown
**Marriage:**
1:Married
2:Single
3:others
**PAY_1 to PAY_6:**
Repayment status from April to September
-1:pay duly
1:payment delay for one month
……
9:payment delay for nine months and above

Target variable: default.payment.next.month

0:Non-default
1:Default

Fig1: Dataset

From the below pie chart, observe that 78% of the customers are non-default and 22% of the customers are default, which clearly says that the dataset is a imbalanced dataset. Categorical features contain the misrepresented values, those values are adjusted by created one more level for those values.

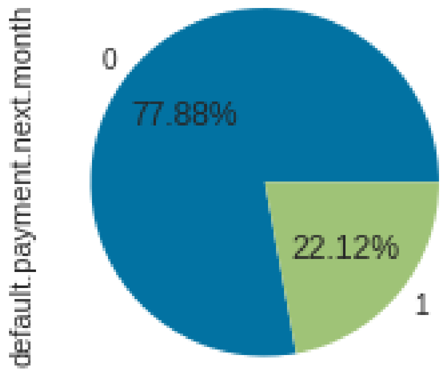
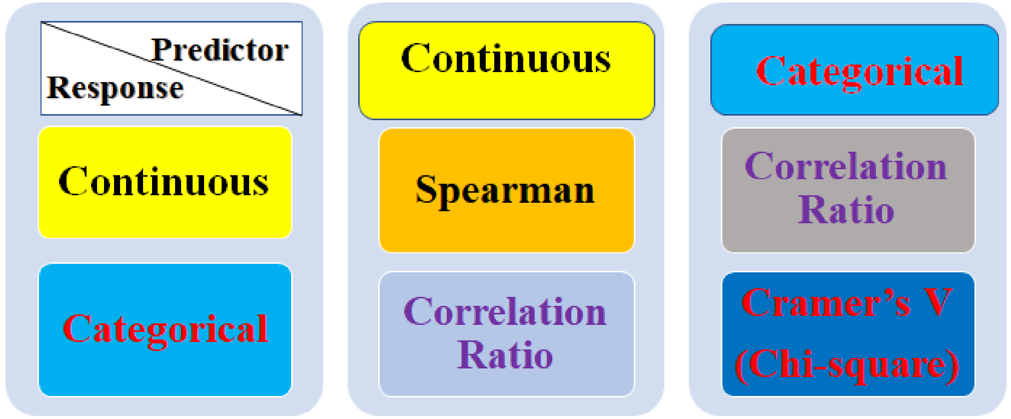


Fig2: Count of Each Class

## Clustering Of Features Using Correlation Matrix

Correlation Matrix: Various correlation techniques which are listed in the image below, helps to find the relationship between the different types of features.



| Response \ Predictor | Continuous | Categorical |
|---|---|---|
| Continuous | Spearman | Correlation Ratio |
| Categorical | Correlation Ratio | Cramer's V (Chi-square) |

Fig 3: Different correlation techniques

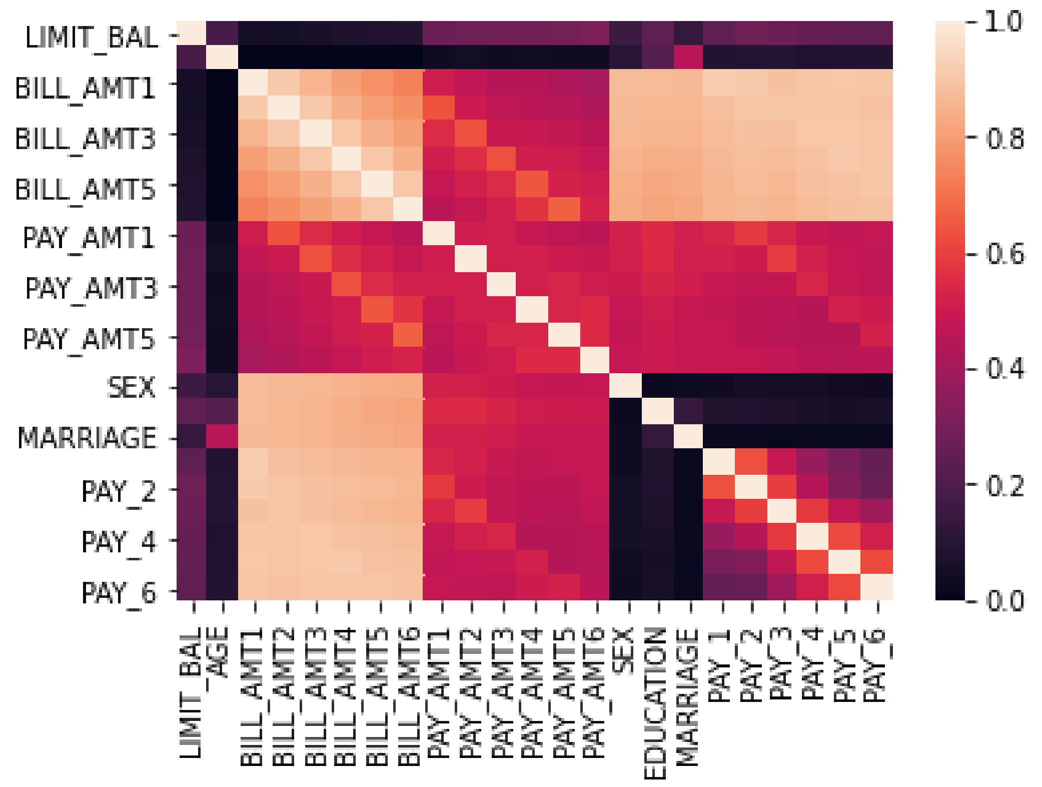


Fig 4:Correlation heatmap

**Hierarchical Clustering:** It is a bottom up approach; we can cut the dendrogram based on our interest. From the below figure we cut the dendrogram at y=2, which leads to 5 clusters.
Figure below shows the clusters based on hierarchical clustering.

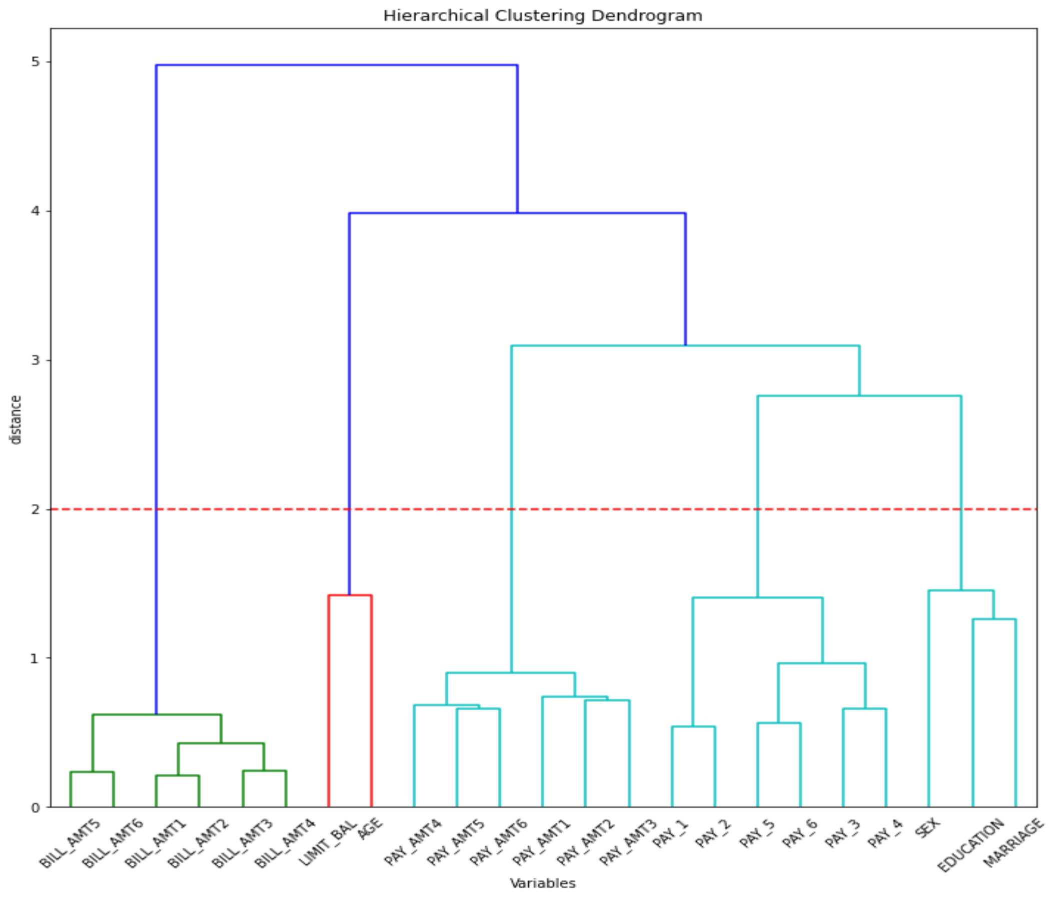


Fig 5: Hierarchical clustering

## Apply GLRM To Calculate The Cluster Scores

Generalized Low Rank Models (GLRM) is an algorithm for dimensionality reduction of a dataset.
It is like PCA, but it can handle mixed numeric, categorical, ordinal and Boolean data with an arbitrary number of missing values.

$$m\Bigg\{\Bigg[\overbrace{\qquad A \qquad}^{n}\Bigg] \approx m\Bigg\{\Bigg[\overbrace{X}^{k}\Bigg]\;\Big[\overbrace{\qquad Y \qquad}^{n}\Big]\Big\}k$$

Fig 6: GLRM matrix

We draw the dendrogram for 5 clusters. Let us say if the first cluster contains some 6 features, GLRM will combine these features and gives one feature which is called cluster score. After applying GLRM the dimensionality of the dataset is changed from (30000,23) to (30000,5)

## Splitting The Dataset Into Train Data And Test Data

Split the data in such a way that 75% of the data is used for training and 25% of the data is used for testing.

## Oversampling Of Data Using SMOT (Synthetic Minority Over-Sampling Technique )

As our dataset is a imbalanced data we are using oversampling technique called SMOT to balance the data. Before applying oversampling, proportions of class 0 and class 1 is of 78:22.

## Binary Classification Algorithms

Binary classification algorithms such as Logistic regression and Support Vector machines are applied to the reduced dimension dataset and the results of confusion matrix, precision, recall is compared. Figure below shows the confusion matrix  for Logistic regression and SVM.

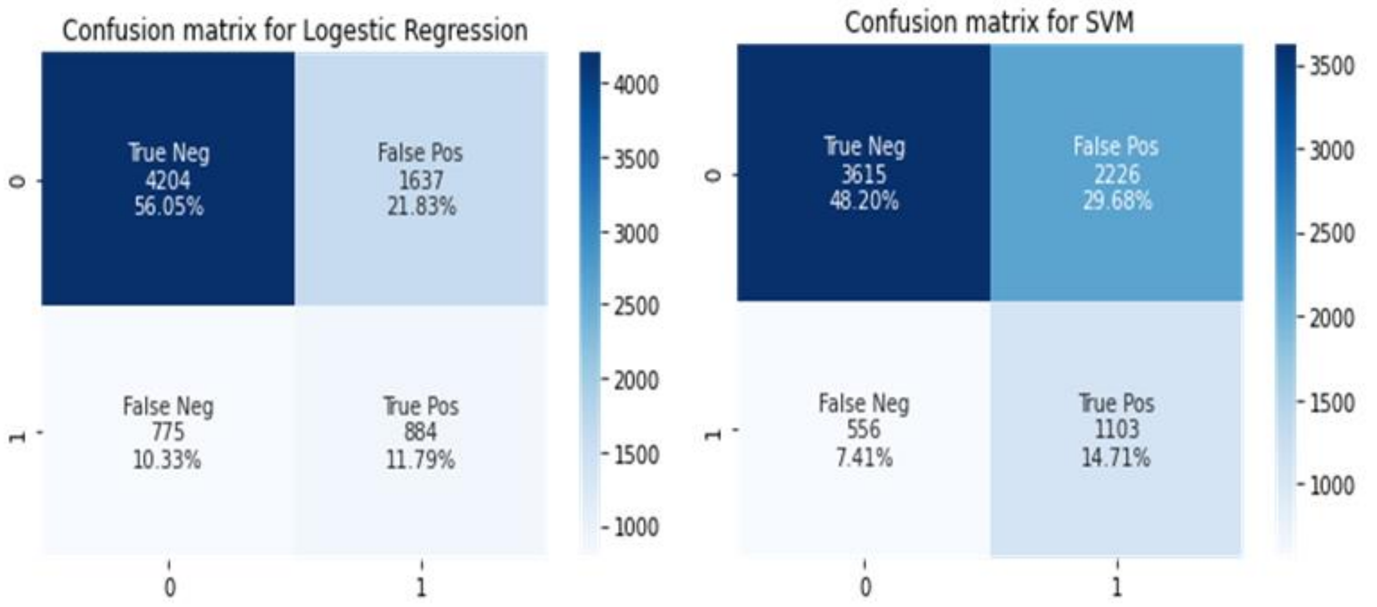
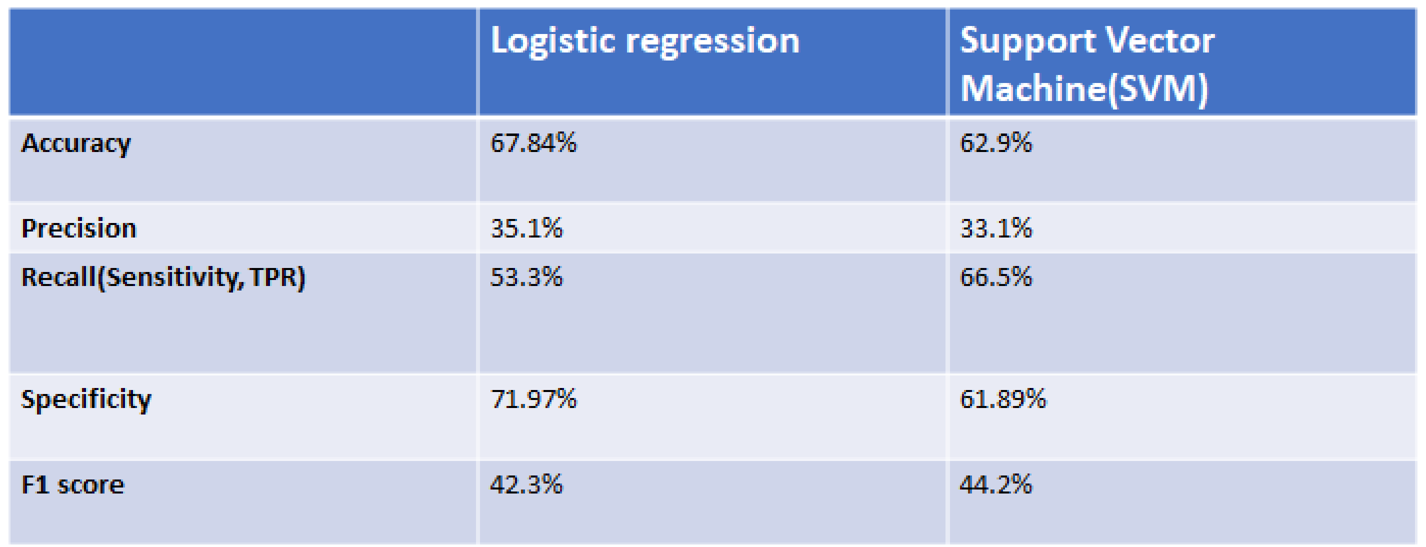


Fig 7: Confusion matrix comparison

## Result

Table below shows the accuracy, Precision, Recall, Specificity and F1-score obtained for both algorithms



| | Logistic regression | Support Vector Machine(SVM) |
|---|---|---|
| **Accuracy** | 67.84% | 62.9% |
| **Precision** | 35.1% | 33.1% |
| **Recall(Sensitivity, TPR)** | 53.3% | 66.5% |
| **Specificity** | 71.97% | 61.89% |
| **F1 score** | 42.3% | 44.2% |

Table 1: comparison of classification algorithms

## Receiver Operating Characteristic Curve

The ROC curve is plotted on a graph with the True Positive Rate (Sensitivity) on the Y-axis and False Positive Rate (1 - Specificity) on the X-axis. The values of the TPR and the FPR are found for many thresholds from 0 to 1.
The main objective here is to find a point on the ROC curve where the Area under it is the maximum. This is because it is at this point, where the model could correctly distinguish between binary classes with there being minimum overlap between them. So as the AUC increases the overlap between the binary classes decreases. At that threshold point, the model can distinguish classes the best.

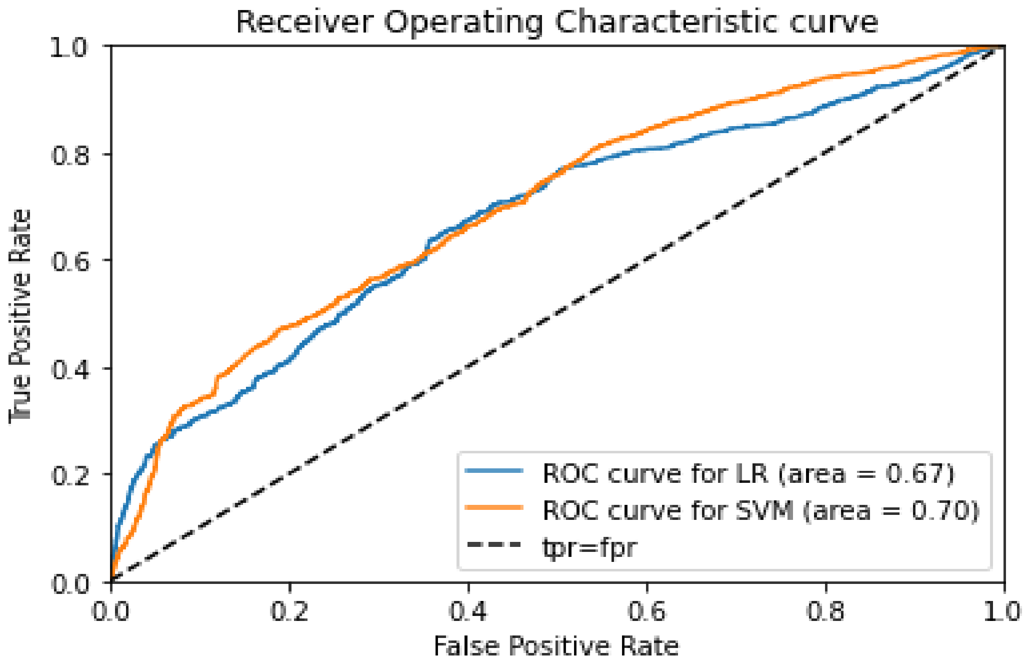


Fig 8: Receiver Operating Curve

## Summary

The confusion matrix for Logistic Regression shows that only 53% of all the defaulters (884 out of 1659) which means that misses 53% of all the defaulters. While the recall score for Support Vector Machine of 66.5% (1103 out of 1659). This means that our model can accurately catch around 67% of all the defaulters which is better than Logistic Regression.

## Future Work

Further the same dataset can be implemented using Cross Validation method and Hyper parameter tuning (Grid search CV) methods.