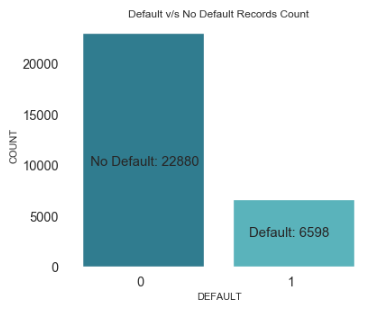
Credit Card Default and Credit Limit Prediction

## Introduction

This project aims to build a data driven and machine learning based approach for predicting the likely defaulters on credit card balance. It uses credit card data from 30,000 clients to build the classification models. The data has 25 variables including basic client demographics and 6 months of payment information (billed amount, paid amount and payment delays). Out of 30000 clients, 6636 clients (22.12%) defaulted on their credit card balance.



## Data Pre-Processing

Some of the categorical variables are in form of labels instead of numeric values. So we converted them to numeric values using LabelEncoder.

Education: ['university' 'graduate school' 'high school' 'other']

SEX: ['female' 'male']

Total number records with Education category of ‘other’ are just 468. Similarly, only 54 records with MARRIAGE category of ‘0’. As these categories don’t have significant data, we can just drop records with these categories to reduce the number of features

Also renamed the dependent variable column ‘default payment next month’ to just ‘DEFAULT’ as it is makes code more readable. The values for this column are converted into numerical values which would be needed for classification. So ‘not default’ is mapped to 0 and default is mapped to 1.

The following repayment status (PAY\_0 to PAY\_6 columns) values can be mapped to 0 (zero) as this reduces the noise in repayment columns. However, payment delay months values of 1 to 8 are more important in predicting defaulters.

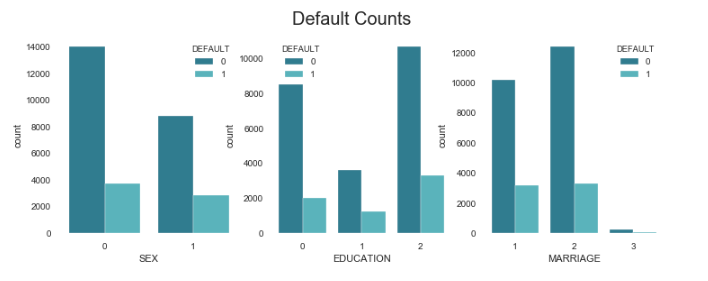
*-2: No consumption; -1: Paid in full;*

are mapped to *0: The use of revolving credit;*

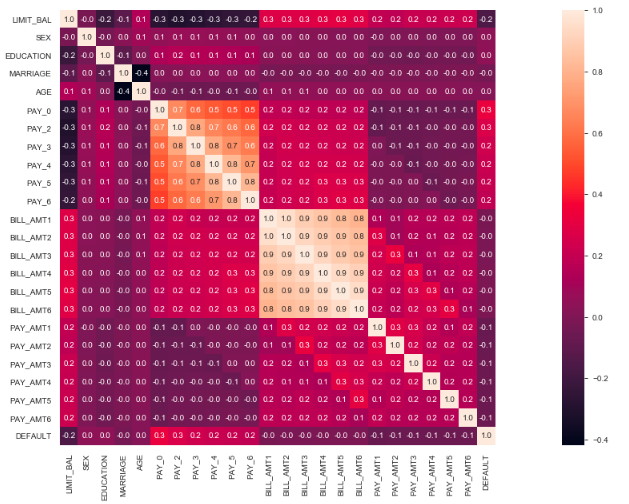
## Exploratory Data Analysis (EDA)

This section discusses the finding of exploratory data analysis and these finding become input to our feature selection process.

Grouping the Defaults data by Gender or Education or Marriage status as in picture below shows that percentage of people who default in each of these sub-categories are almost same (in range of 20% to 25%).



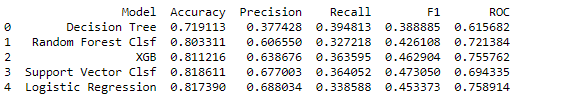
Also, the correlation matrix provides similar conclusion that Gender, Education and Marriage columns have almost no correlation to whether a person would default or not. Similarly, as seen in correlation matrix, the Bill\_Amount columns have no correlation to Default column.



Therefore, we deleted for Gender, Education, Marriage and Billing Amounts column from our dataset which we would use for our modeling exercise.

## Modeling

One of the objectives of this study is to predict if a person will default or not based on various demographic and credit card debt payment attributes (features). We evaluated various classification models to predict if a person will default. As show in data below, we performed cross validation on entire data set using 5 different classification models and printed various types of scores for the models.



Selecting relevant scoring option: Four of models have an accuracy of around 80%. This means that in 80% percent of the cases model will correctly predict if a person will default or not. However, our model should not predict a person as ‘Not Default’ if the person is ‘Default’ based on actual data. Therefore, we should pick a model with higher Recall value.

Recall = True Positive / (True Positive + False Negative)

Lower value of ‘False Negative’ means we don’t predict ‘No Default’ when person is actually a Default. This situation is not good for the business. However, high value of Recall may have a lower value of Precision. Therefore, we need to take a middle ground which is defined by F1 score.

Therefore, we pick 3 models (XGB, SVC and Logistic Regression) with highest F1 score and tried fine tuning them to select our final model.

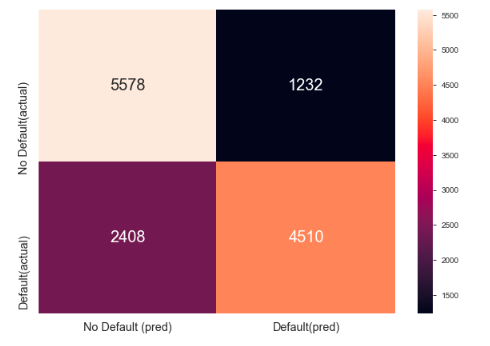
## Model Fine Tuning

Hyperparameter tuning was performed for XGB, SVC and Logistic Regression. However, it did not boost accuracy, recall, F1 or rocauc.

In order ensure that imbalance in the classes (default, no default) is not causing underperformance of our models, SMOTE technique was used for oversampling of under-represented class. This improved the scores for all the models, however, XGB performed better than others. The accuracy score decreased a little bit after SMOTE, but Recall and F1 score increased.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 | ROCAUC |
| Logistic Regression | 0.70 | 0.78 | 0.55 | 0.65 | 0.70 |
| SVC | 0.69 | 0.78 | 0.53 | 0.63 | 0.69 |
| XGB | 0.73 | 0.79 | 0.65 | 0.71 | 0.74 |

From the confusion matrix created for XGB Classification model below we understand that our model predicts 4510 (65%) defaulters correctly out of 6918 actual defaulters. However, it is still not caching 2408 (35%) of the actual defaulters.



Before SMOTE of the data, due to class imbalance, our model was not able to catch 1316 (65%) out of 2013 actual defaulters. Following is the confusion matrix before SMOTE of the data.

