Group 3 - IE7374 35456 - Machine Learning Project

Sindhu Swaroop, Niraj Sai Prasad, Nikhil Nijhawan, Saurabh Shukla

CREDIT CARD PAYMENTS DEFAULT PREDICTION USING ML

ABSTRACT

Credit cards are one of the most profitable financial services offered by banks in recent years. A credit card defaulter is a borrower who fails to make credit card payments on time. The demand for credit cards is ever-increasing, and banks have been facing alarming default rates. With the number of credit card owners rising, an enormous amount of financial data is being generated everyday. The banking sector will greatly profit by assessing an individual's credit risk and predicting the occurrence of credit card defaults, before issuing credit cards to just anyone who applies for one. Usage of a person's credit risk is not limited to issuance of credit cards, this statistic is also assessed before issuing any kind of loan. The bank needs some sort of guarantee that the loan/credit card bills will be repaid in a timely manner. In this project, we have built a machine learning model that predicts whether a certain individual will be a defaulter. This prediction is influenced by a variety of aspects which we will be exploring in subsequent sections.

Credit risk of an individual can be calculated based on a number of factors such as - age, gender, education, employment status and marital status, to name a few.

INTRODUCTION

Advancements in machine learning have made credit card default risk prediction a relatively easy task. With respect to machine learning, default risk prediction is essentially a binary classification problem with two possible outcomes - 'Defaulter' (1) and 'Not a Defaulter' (0). Binary classification can be solved using different models such as Logistic Regression, Naive-Bayes or Support Vector Machines (SVM). Below are the major steps we have followed in implementation of this project:

- Exploratory Data Analysis (EDA) - Plotting different graphs to visualize the credit limit range and distribution, distribution of other features and multivariate analysis.
- Correlation Matrix - Plotting heat map to visualize the correlation between each of the features with themselves as well as with the value to be predicted
- Data wrangling
    - Conversion of data into numeric data types to be able to perform operations on them
    - Renaming columns to make sense
    - Dropping unnecessary columns
    - One-hot/binary encoding on the nominal features of the dataset - education and marital status
    - Scaling of data - Using standardization to limit the values between [1, -1] and make their mean approximately zero.
- Train-Test split - Data has been split into two using the sklearn_model selection train_test_split function with a train size of 0.8 i.e., training set is 80% of the entire data for Logistic Regression and 3% for SVM
- Computation of results using inbuilt packages to see which model performs best - Logistic Regression, Naive Bayes, KNN or Support Vector Machine.
- Model building by writing own code
    - Custom Logistic Regression model
    - Custom Soft Margin SVM model

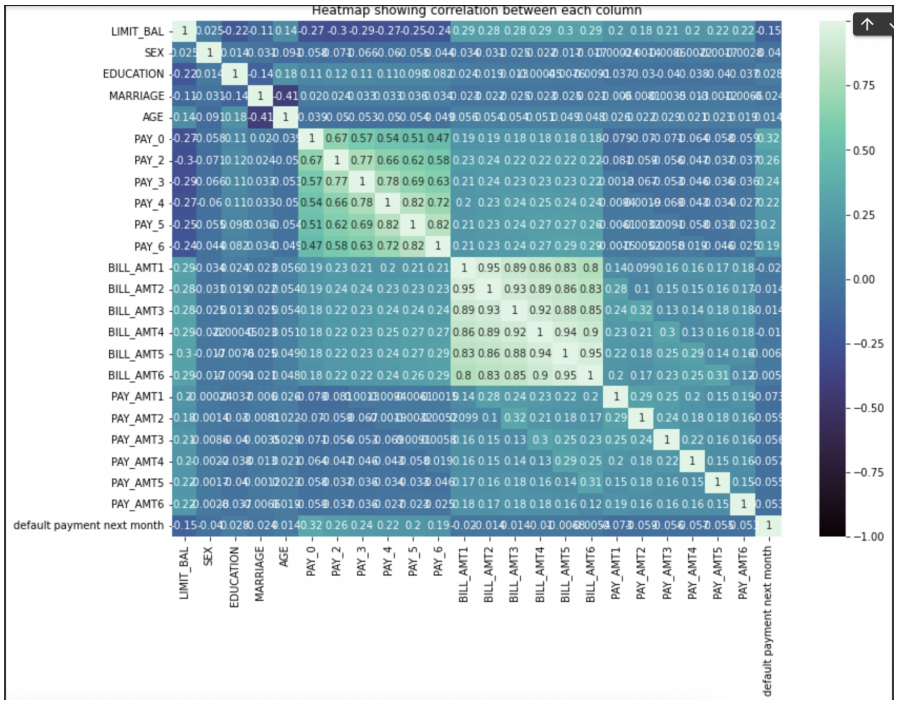
DATA DESCRIPTION

We have picked a dataset from UCI machine learning repository - https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

This dataset contains 30000 records and the following 23 variables as explanatory variables:

- C1: Amount of given credit: The amount of credit (in USD) given to the customer. This represents the credit limit available on the issued credit card.
- C2 : Gender (Male, Female, etc)
- C3 : Education: There are 6 educational classifications
- C4 : Marital status: There are 4 marital status classifications (Married, Single, etc)
- C5:Age : Age of the customer in years
- C6-C11 : History of past payment: Past month's repayment status for Apr-Sept 2005 (Paid duly, Payment delayed for one month, Payment delayed for two months, etc)
- C12-C17 : Amount of bill statement: The amount in USD of credit card bill statement for Apr-Sept 2005
- C18-C23 : Amount of previous payment: The amount paid (in USD) in the previous month for Apr-Sept 2005

We plotted a heat map to visualize the correlation between the different features:



- We observed a strong positive correlation between the bill amounts which is plausible as the bill amount of a particular credit card user will be more or less similar throughout the timeline.
- Another strong positive correlation was between the repayment status. On the same lines as bill amounts, repayment status will be similar for an individual user.
- As for the output variable - ‘Default payment next month’, it showed no strong correlation with any of the features, which tells us that the default prediction is based on a combination of all the features rather than a single feature.
- A weak negative correlation was observed between credit limit and default prediction

METHODS

We have used two classification methods for our model - Logistic Regression and Soft Margin SVM.

LOGISTIC REGRESSION:

The binary logistic regression model is a statistical model that predicts the occurrence of one out of two events by calculating the logarithm of the odds of the events by taking a linear combination of one or more independent variables or features. In binary logistic regression, there is a single binary dependent variable, coded by an indicator variable, with two values usually labeled as "0" and "1", and the independent variables can be either binary or continuous variables i.e. having either two values or can consist of any combination of real numbers. The likelihood of the value labeled "1" might fluctuate between 0 and 1, therefore the labeling. The function then converts log-odds or the logarithm of the features to a probability. The binary logistic regression is an extension of the simple linear regression and is used where the target variable is dichotomous in nature. It has a high bias and low variance which leads to underfitting.

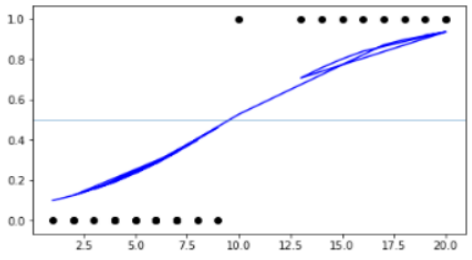
Logistic regression can be expressed mathematically in the form :

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

The left side of the equation is the log-odds, which tells us the logarithm of the ratio of the probability of success to the probability of failure for an event. To find out the p(X) we can simply calculate the inverse of the above function that yields:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

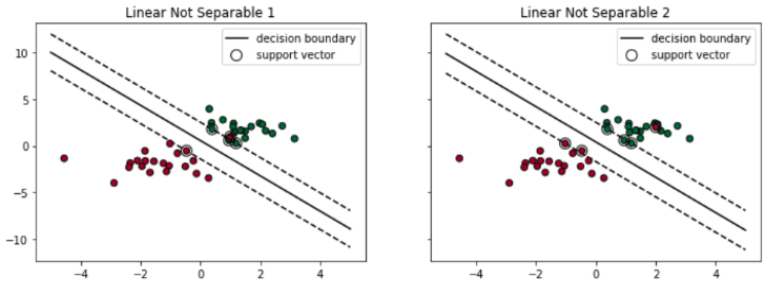
The above equation is also referred to as the sigmoid function and since it's a probability the value ranges between 0 and 1.



Sigmoid Function.

SOFT MARGIN SVM:

Support Vector Machines (SVMs) is a classifier which is discriminative in nature and it works by generating a hyperplane that separates the new samples based on the training data. In a two dimensional space, it can be considered as a line dividing the two classes. So, SVM tries to maximize the margin of the plane separating the classes or the distance between the points closest to the separating plane with the decision boundary exactly in the middle of it with a condition that the classes are correctly classified. Soft-Margin SVMs allow a certain number of points to be misclassified and balances the trade-off between maximizing the margins and the number of points that will be misclassified. The way in which this misclassification can happen is if the points are on the wrong side of both the margin and the decision boundary or if the points are on the wrong side of the decision boundary but the right side of the margin.



Degree of tolerance is an important hyper-parameter when using SVM as it tells how much penalty the algorithm is going to get. The larger the value of tolerance signifies more penalty, and hence the narrower the margins are fewer the number of support vectors. SVM is a non-linear algorithm that has low bias and high variance which leads to overfitting. The bias-variance tradeoff can be changed by modifying the C value hence allowing more number of misclassifications in the training data, increasing bias and decreasing variance.

Generally, Kernel tricks are used in order to make use of the already existing features, perform some transformations and create new usable features. These features are an integral part for the SVM to find a non-linear decision boundary.

$$L = \frac{1}{2}\|w\|^2 + C(\#\ of\ mistakes)$$

equation 1

The objective of Soft-Margin SVM is the minimization of the below given equation :

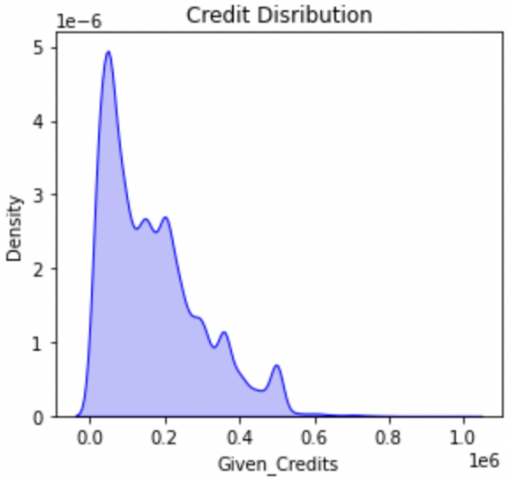
$$L = \frac{1}{2}\|\vec{w}\|^2 + C\sum_i \xi_i + \sum_i \lambda_i(y_i(\vec{w}\ .\ \vec{x}_i + b) - 1 + \xi_i)$$

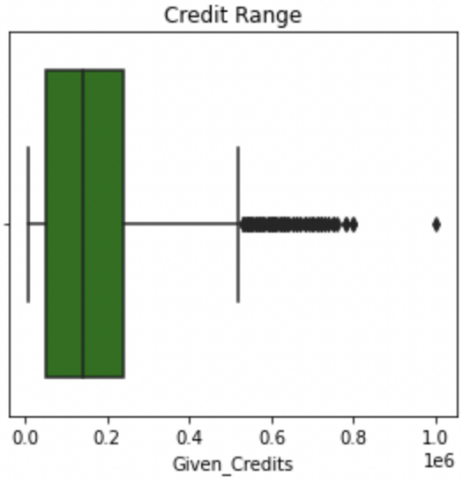
EXPLORATORY DATA ANALYSIS (EDA)

1. CREDIT LIMIT DISTRIBUTION AND RANGE:

From the below Credit Distribution graph, it can be observed that the credit limit of $50000 has the highest density followed by $20000. The credit limit with lowest density is $1000000.

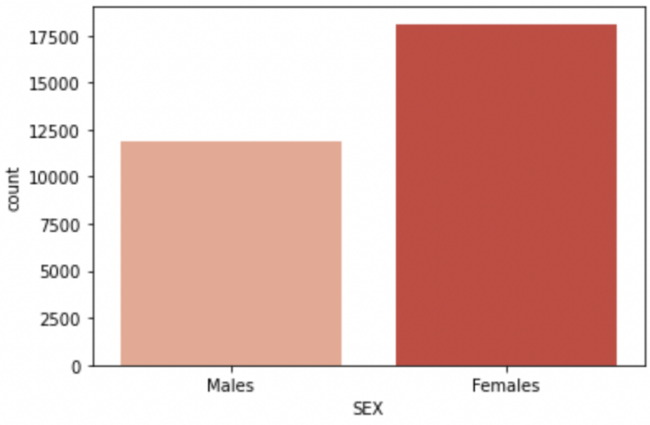
The Credit Range graph gives us the minimum - $10000, maximum - $1000000 and mean - ~$167484 credit limit values. The lower quartile is $50000 and upper quartile is $240000.





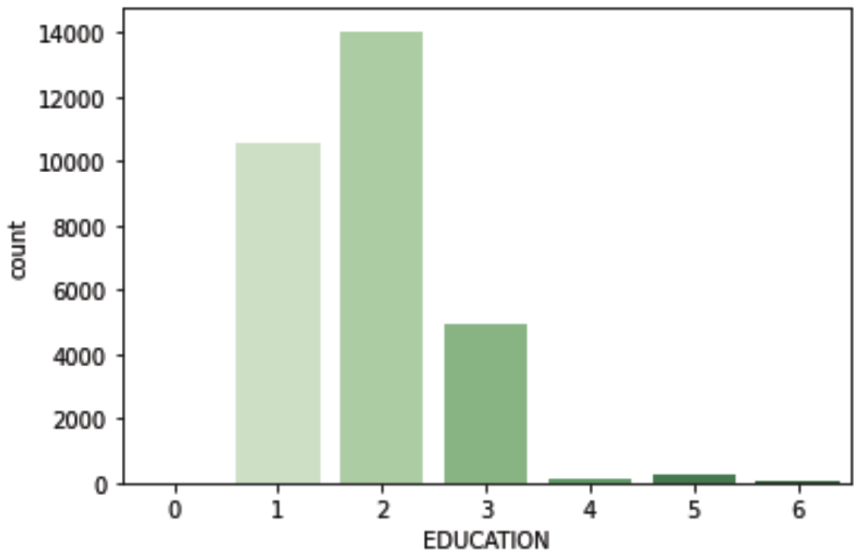
2. GENDER DISTRIBUTION:

From the below bar graph, we can see that this dataset has more females (60%) than (40%).



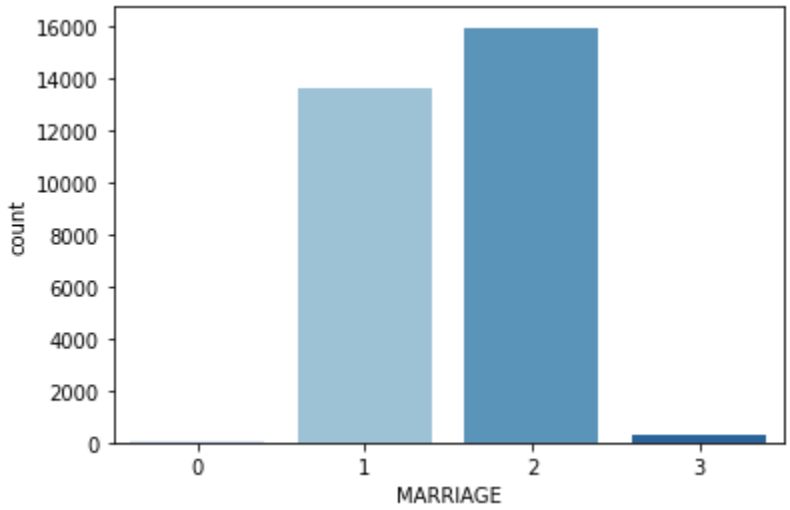
3. **EDUCATION:**

From the graph below, we can see that most of the people in the dataset have university level education(46%) followed by graduate level(35%) and high school level(16%).



4. **MARRIAGE:**

From the graph below, we can see that most people in the dataset are single and second most are married, followed by others category.

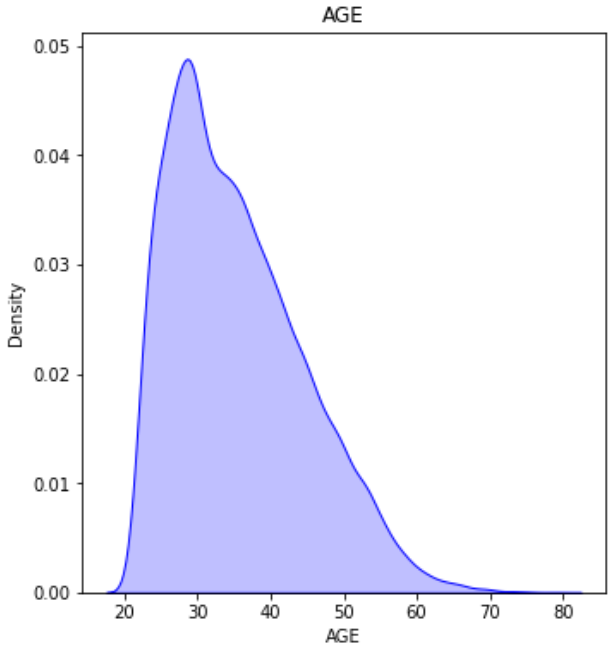


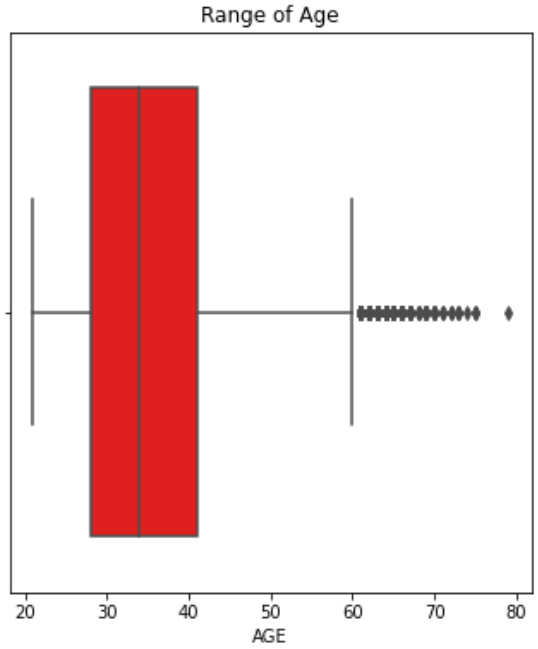
5. **AGE:**

Minimum age to get a credit account is 21 and maximum is 79 yrs old.

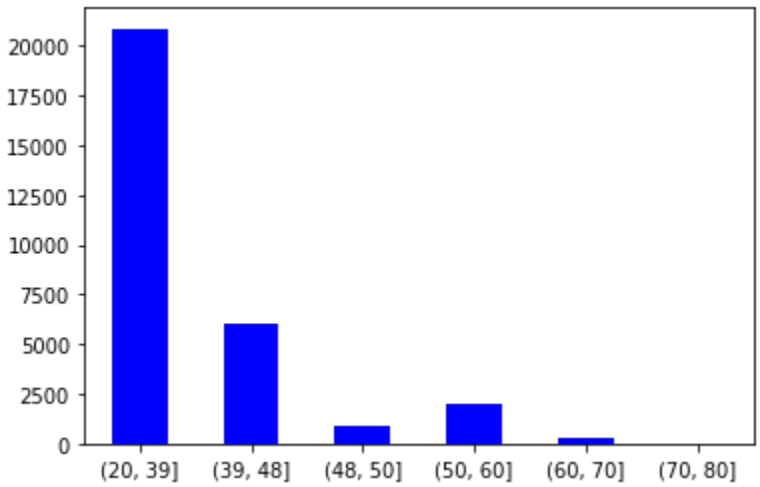
In the age distribution graph, we can see that most of the people with credit accounts are in the range of 20-40 with a peak at 29 yrs old.

In the Age range plot, most of the people with credit accounts are in range 29-41 with median age of 34 yrs old. About 70% are in the age group of 20-40 yrs.



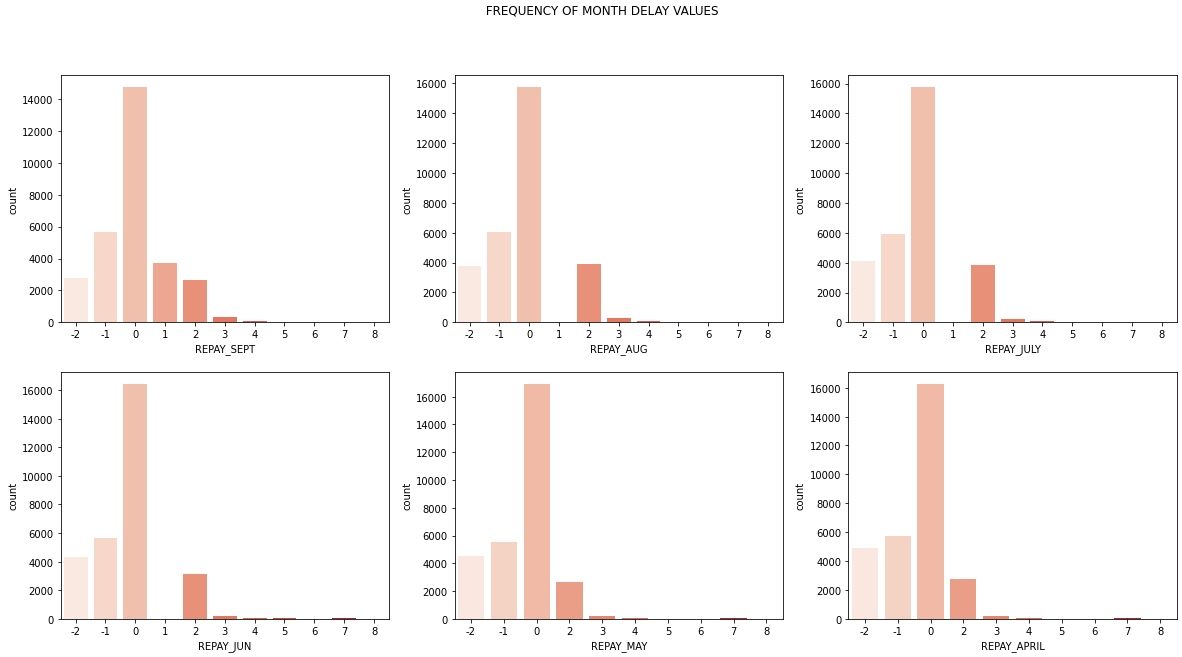


We can see that people in the 20-39 have more than 20,000 credit accounts.



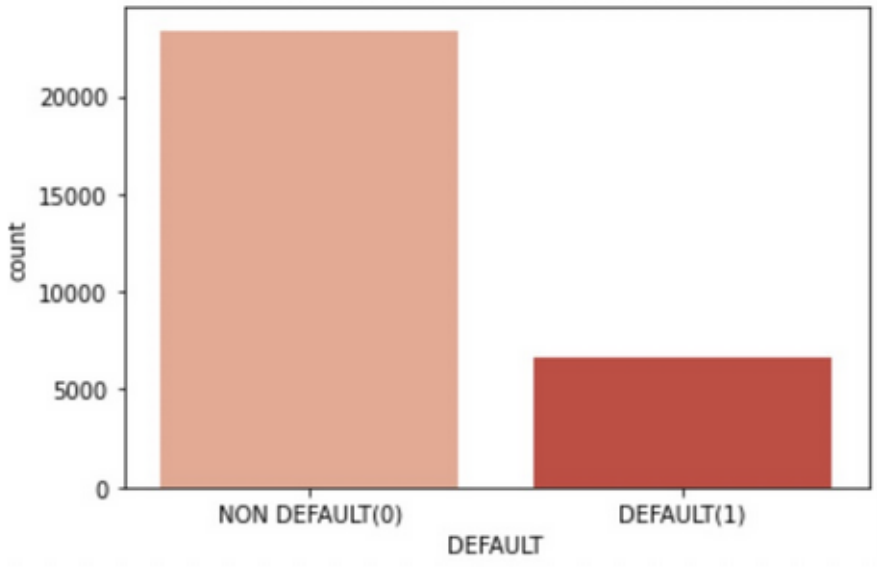
6. REPAYMENT STATUS:

In the graph below, it can be seen that most of the people in the dataset have paid their dues on time. For September, many people paid their dues a month later. For every month from April to September, many people were 2 months late to pay their dues.



7. DEFAULT:

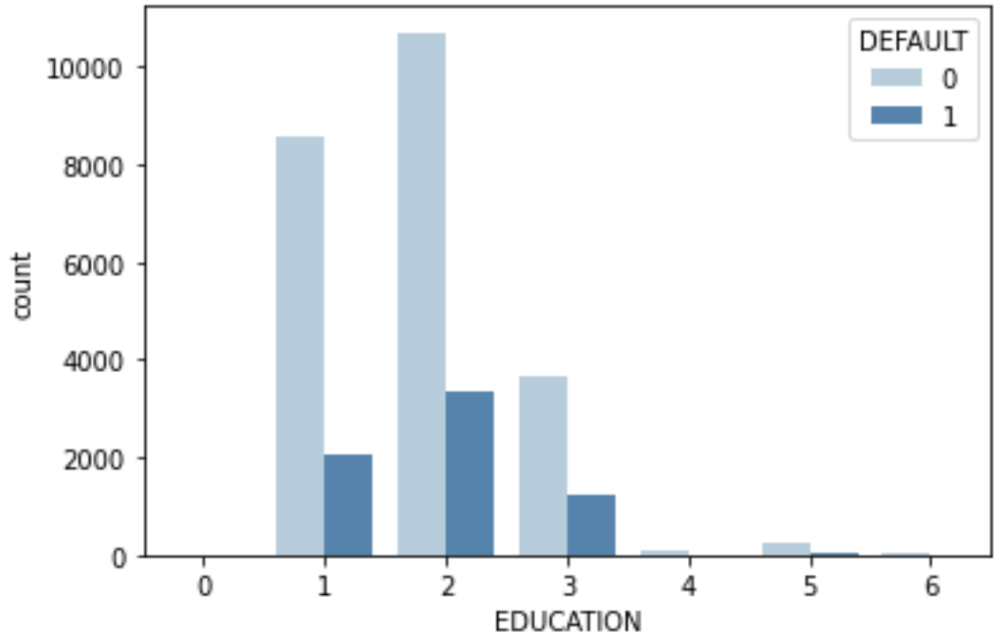
There are a significant number of people with default accounts in the dataset over non default.



MULTIVARIATE ANALYSIS

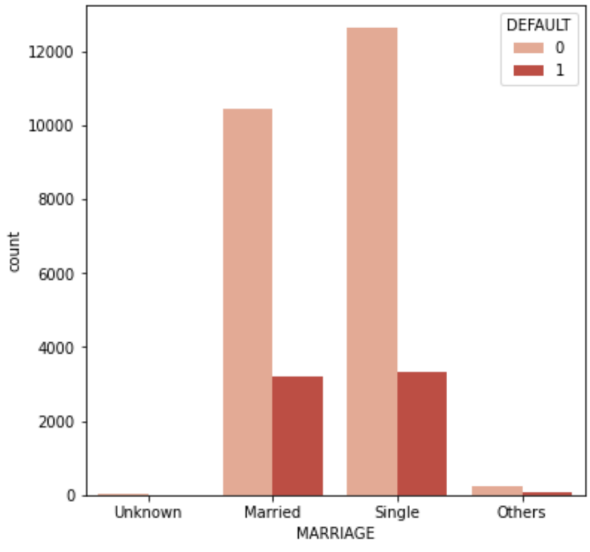
1. EDUCATION VS DEFAULT

About 20% of people with graduate level education are defaulters. About 23% of people with university level education and about 25% of people with high school level education are defaulters. More educated people have less probability of being defaulters.



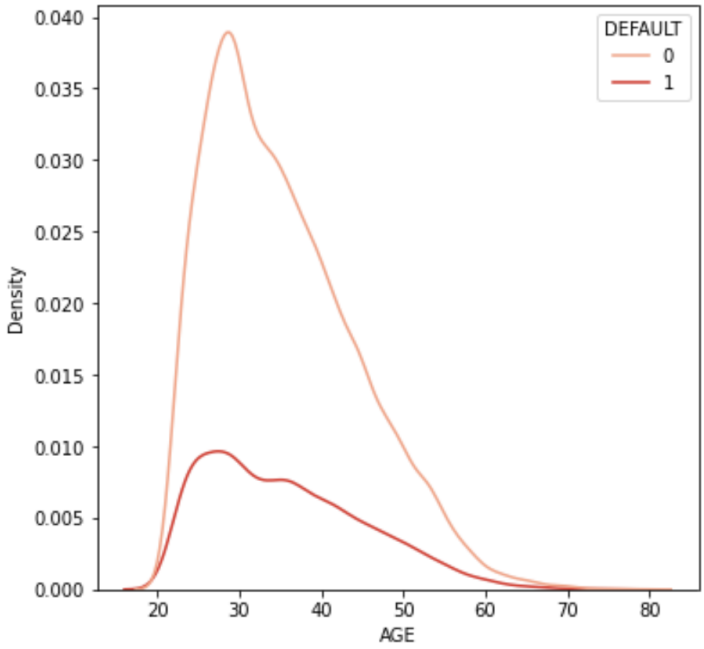
2. MARRIAGE VS DEFAULT

Although most of the people in the dataset are Single, there are about the same number of defaulters who are either married or single.



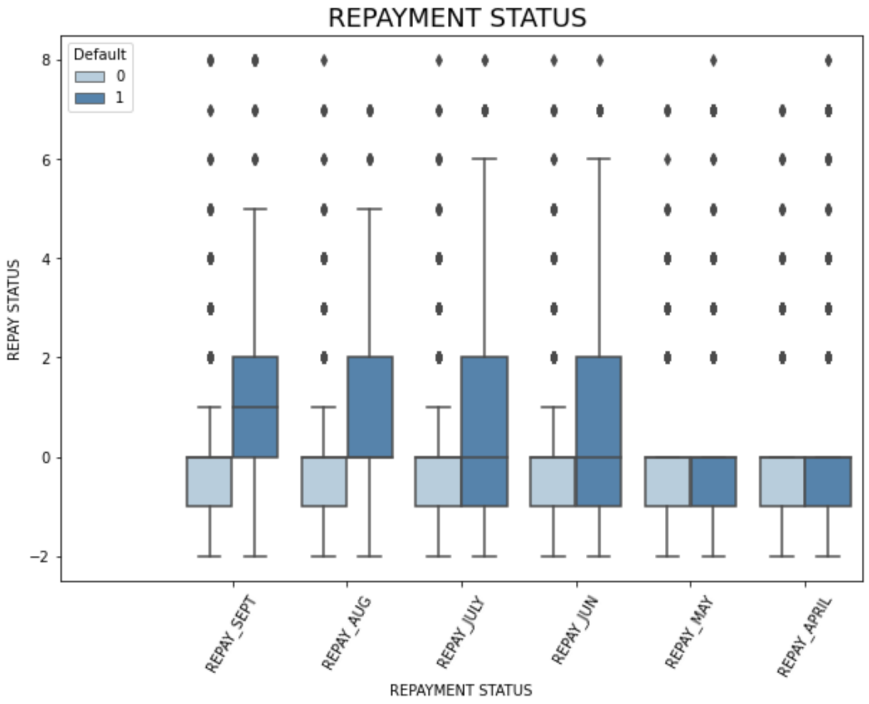
3. AGE VS DEFAULT

People who defaulted payment for the next month are majorly in the 20 to 40 years age group.



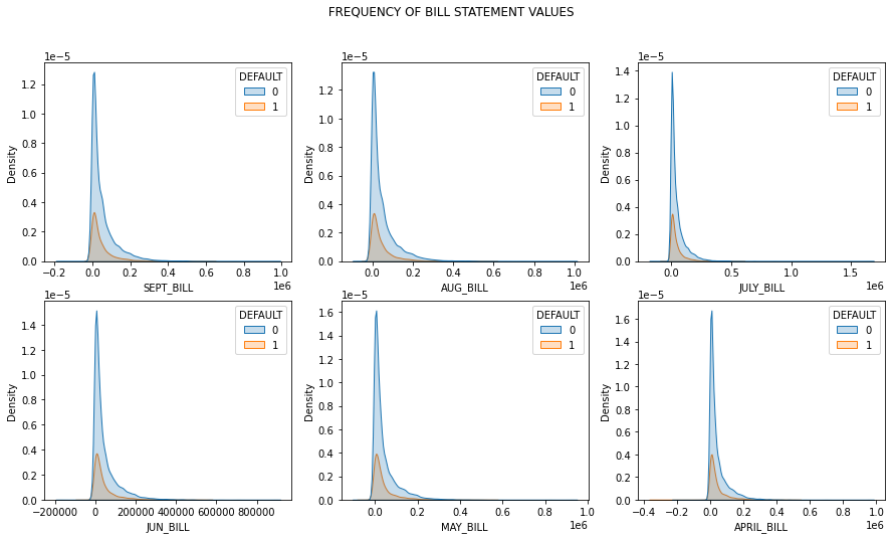
4. REPAYMENT STATUS VS DEFAULT

We can see that REPAY_SEPT and REPAY_AUG can be used to differentiate DEFAULT much better than other variables.



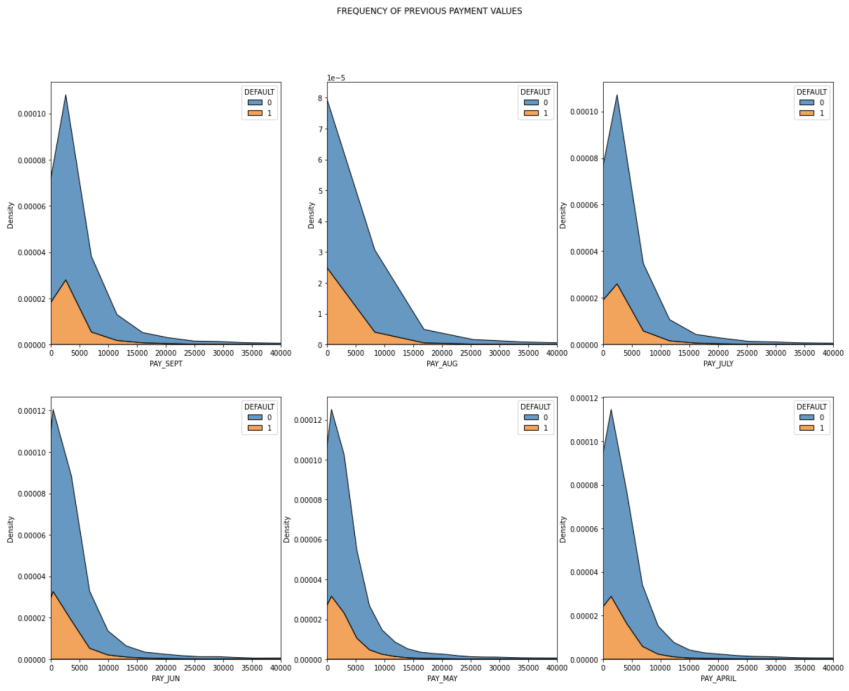
5. BILL STATEMENT VS DEFAULT

Most of the defaulters have a limit in the range of 0 to $200,000.



6. PREVIOUS PAYMENT VS DEFAULT

Most of the defaulters have previous payments in the range of 0 to $5000.



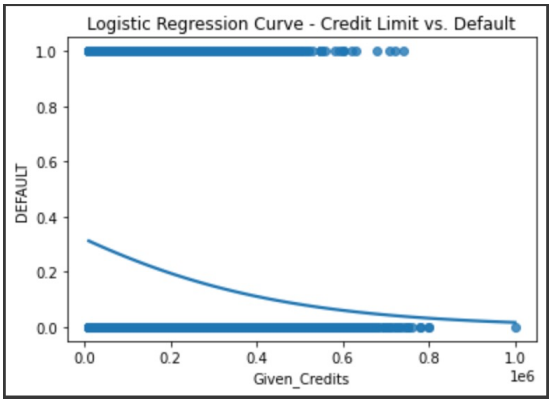
RESULTS & DISCUSSION

We have built two models for credit card default prediction - Logistic Regression and Soft Margin SVM.

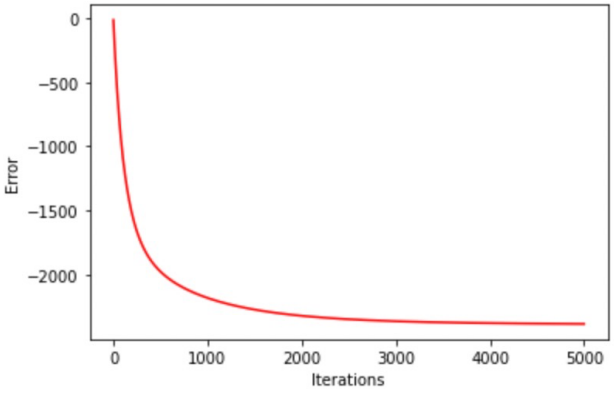
LOGISTIC REGRESSION:

We have used Logistic Regression as the baseline model for our analysis. Adding L2 Regularization (Penalty term - lambda) did not alter the accuracy significantly, hence we have used a lambda value of zero. The training size of the model was 80%. The **Accuracy** on the **training set** obtained in this model was **59.7%**, along with a Precision of 31.82% and Recall of 71.63%. The **Accuracy** on the **test set** obtained in this model was **59.08%**, along with a Precision of 31.05% and Recall of 70.84%. We can see that the Accuracy (with a **difference of ~0.6%**) , Precision and Recall are almost the same for both the training set and test set, which tells us that this model is performing well and is neither underfitting nor overfitting.

Below is the Logistic Regression Curve for Credit Limit vs Default. It can be seen from the graph that at higher credit limits, the probability of the credit card user being a defaulter is lower. At lower credit limits, there is a higher chance of the individual being a defaulter.



We also plotted the cost function (error) with the iterations. It can be seen that the model stops learning at 3000 iterations. At this point, the error function becomes constant.

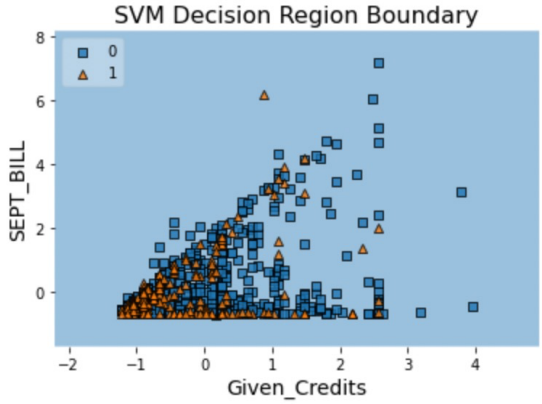


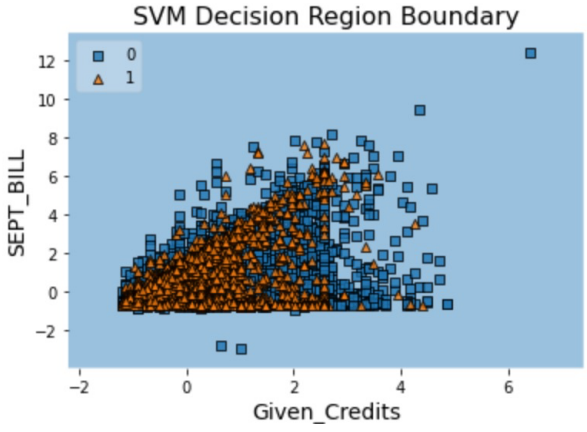
SOFT MARGIN SVM:

The training size of the model was small - 3% of the entire data. We could not use a higher training size than this as the optimize minimize function used up the entire system memory and the kernel died. Bias-Variance tradeoff for this model has been optimized by using a C value of 62 which gives the best model performance. The **Accuracy** on the **training set** obtained in this model was **60.77%**, along with a Precision of 23.45% and Recall of 34.17%. The **Accuracy** on the **test set** obtained in this model was **59.4%**, along with a Precision of 21.57% and Recall of 31.71%. We can see that the Accuracy (with a **difference of ~1.3%**), Precision and Recall are almost the same for both the training set and test set, which tells us that this model is performing well and is neither underfitting nor overfitting.

Below is the SVM Decision Region Boundary graph for the training set (top) and test set (bottom). We have used two features to plot this graph - Credit Limit and the Bill amount of September Month. It can be seen that the decision surface is somewhat similar for both sets of data. The blue points represent non-defaulters and orange represents defaulters.

There is no hyperplane that can achieve zero classification error in this case because our data is not linearly separable with respect to all X. There is no linear separator found.





To conclude, we can see an **improvement in accuracy of 1% over Logistic Regression**, however the time taken by SVM is much more and hence it proved to be too costly for this model.