****

**Group 3 | PGP-DSE July 2019 (Pune) | 27 Nov. 2019**

**CAPSTONE PROJECT - FINAL REPORT**

**Credit Card Default Propensity Prediction**

**SUBMITTED BY:-**

**Pratik Nirgudkar - OFK1USYQEW**

**Akash Patil - AKPGO6YVX5**

**Pranil Shinde - H17W4YTGB1**

**Shivam Kulkarni - DCR8BP0LLQ**

**Akshay Sonawane - U2GDJV22MM**

**Under the Guidance of**

Mrs. Anjana Agarwal

**ACKNOWLEDGEMENT**

At the outset we wish to express our gratitude to various people, for their support, without which this project would have taken a different

We would like to express our sincere gratitude to our mentor Mrs. Anjana Agarwal for being with us always and providing invaluable guidance throughout the process. His long experience has helped us immensely in completing this project and we are very thankful to her.

The completion of this work as a part of the capstone project in Great Lakes Institute of Management gives us immense pleasure and would definitely be a milestone in our Data Science careers.

**ABSTRACT / SUMMARY**

In the last few years, credit card issuers have become one of the major consumer lending products. Credit cards issued by banks hold the majority of the market share with approximately 70% of the total outstanding balance.

In this project, the main aim is to predict the probability of a customer defaulting payment for the credit card the subsequent month, based on past information. The past information is provided in the dataset. This probability will help the collections team to prioritize follow up with customers who have a high propensity of defaulting.

The fundamental objective of the project is implementing a proactive default prevention guideline to help the bank identify and take action on customers with high probability of defaulting to improve their bottom line. The challenge is to help the bank to improve its credit card services for the mutual benefit of customers and the business itself. Creating a human-interpretable solution is emphasized in each stage of the project.

**CONTENTS**

1. Acknowledgement
2. Abstract
3. Introduction
4. Qualitative Analysis
5. Quantitative Analysis
6. Model Development
7. Business Insights
8. Conclusion

**PROBLEM STATEMENT**

Predict the probability of a customer defaulting payment for the credit card the subsequent month, based on past information. The past information is provided in the dataset. This probability will help the collections team to prioritise follow up with customers who have a high propensity of defaulting.

**PROJECT METHODOLOGY**

Qualitative and quantitative are two different approaches for researching. The problem is the main criteria to select which approach is the most appropriate and the data, on which we one does the research, also plays an important part while deciding the methods.

A Qualitative research refers to a process that produces theoretical and descriptive outcomes. A Quantitative research is the one in which the information collected is expressed and analysed in numerical form.

In this research work, both qualitative and quantitative approaches have been used in order to understand the problem and find out the best solutions for the business.

**ABOUT DATA:**

**UNDERSTANDING THE FEATURES**

The data consists of 30,000 customers and 26 columns of variables. Each sample corresponds to a single customer. The columns consist of the following variables:

* + - Default (Yes or no) as a binary response variable i.e 1 or 0
    - Balance limit (Amount of credit)
    - Sex (Male, Female)
    - Education (Graduate school, University, High school, Others)
    - Marital status (Married, Single, Others)
    - Age (Years)
    - Employer (Company name)
    - Location (Latitude, Longitude)
    - Payment status (last 6 months)
      * Indicates payment delay in months or whether payment was made duly
    - Bill amount (last 6 months)
      * States amount of bill statement
    - Payment amount (last 6 months)

#### Default

This variable indicates whether or not the customer defaulted in their credit card debt payment. For the purpose of this project, predicting default is the main focus of the data analysis. A value of 1 indicates default, and a value of 0 indicates no default.

#### Balance limit

Balance limit states the amount of given credit in US $. This is the maximum amount a customer can spend with their credit card in a single month. The amount of balance limit is dependent on the bank’s own screening processes and other unknown factors.

#### Sex

This variable can obtain a value of 1 for male and 2 for female. In this study, sex and gender are used interchangeably to intend the same thing. It is unknown whether the difference between the two definitions were taken into account when the data was collected.

#### Education

The education level of a customer is represented as one of four values: 1 = Graduate school, 2 = University, 3

= High school, 4 = Other. For the purpose of analyzing customer groups, this is assumed to indicate the highest level of education completed.

#### Marital status

Referred to as “married” in the analysis, this variable can obtain three values: 1 = Married, 2 = Single, 3 = Other such as divorced or widowed.

#### Age

Age of the customer is stated in years.

#### Location

This variable is composed of two different values for each customer. One is for the latitude, and the second one is for the longitude. In order to gain benefits from this data in predictions using only the demographic variables, we applied the DBSCAN algorithm.

#### Payment status

Payment status is represented as 6 different columns, one for each month. The value of payment status for a month indicates whether repayment of credit is was delayed or paid duly. A value of -1 indicates pay duly.

#### Bill amount

Amount of bill statement in U.S. $ is recorded in this variable. It is represented in the data as 6 columns, one for each month. Data collected from 6 months, April to September.

#### Payment amount

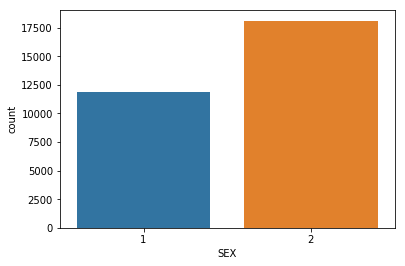
Amount of previous payment in U.S. $, stored in 6 different columns for each month, similarly to payment status and bill amount. The payment amounts correspond to the same months as payment status and bill amount. For example, the payment amount for April indicates amount paid in April.

**QUALITATIVE ANALYSIS**

The data for different features are analysed graphically, first separately and then with each other. This has helped to define how different features are behaving in themselves and with each other comparatively.

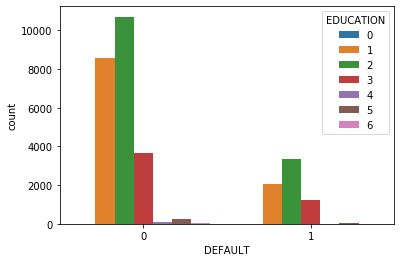
Statistical analysis has also been done to understand the relationship among the features in required places.

Graphical Analysis:



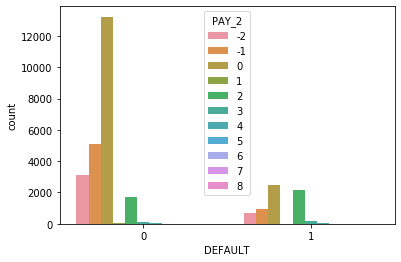
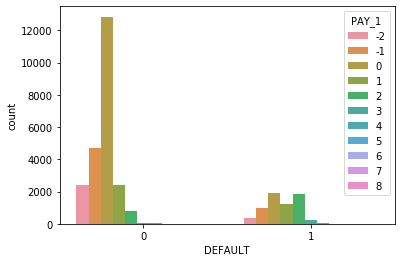
In our data, we found that male are less than female. The count of males are 11888 and the count of females are 18112.

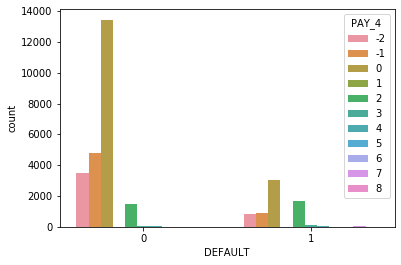
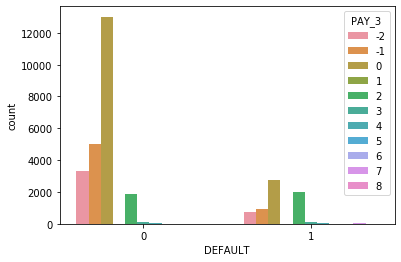
Countplot for DEFAULT and EDUCATION :

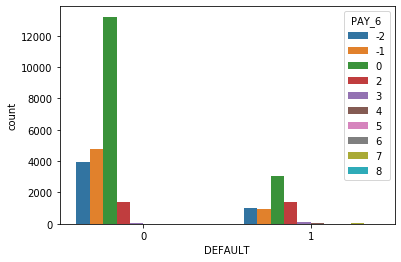
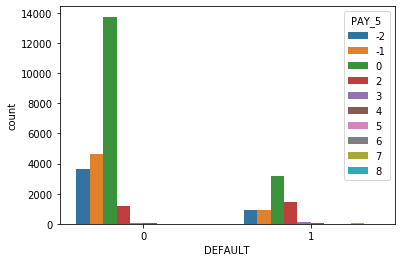


In Default feature, we have seen the dispersion of education field that 1 as graduates are more.

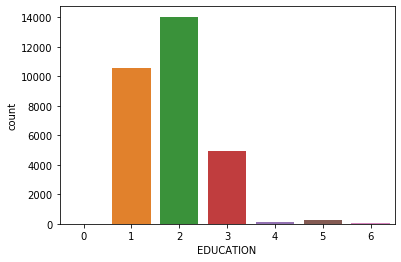
Pay\_scale vs Default :



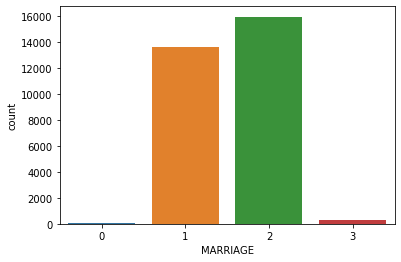




Education:

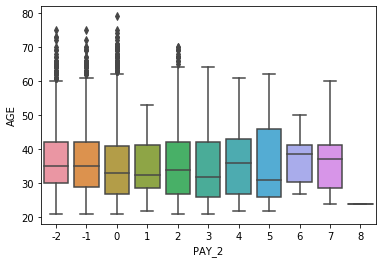
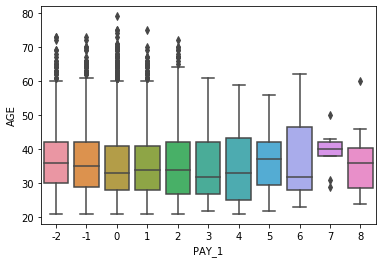


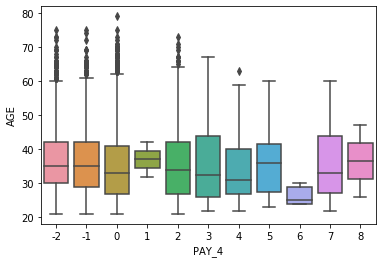
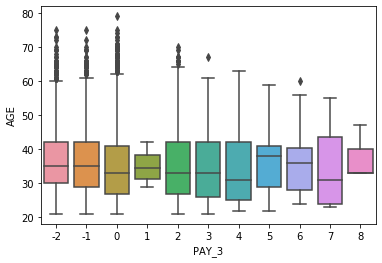
Marriage:

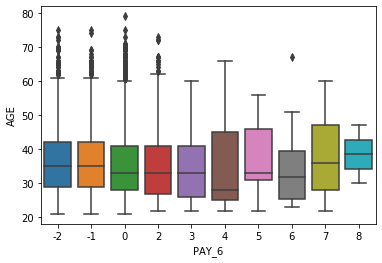
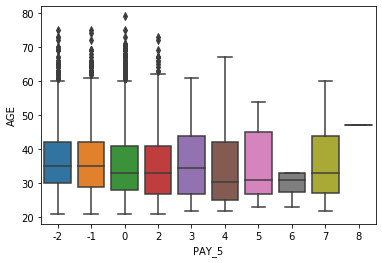


Bivariate Analysis:

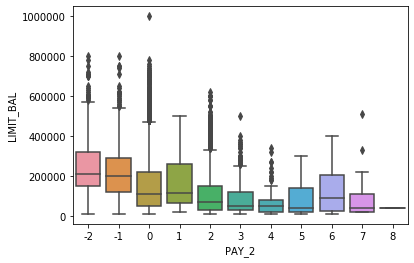
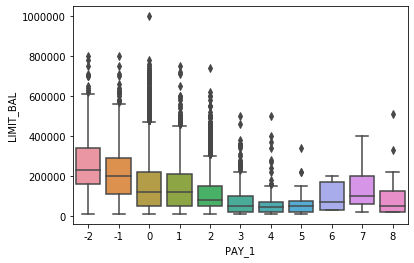
1. Comparing AGE and Past Pay:

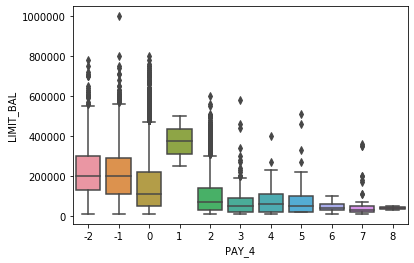
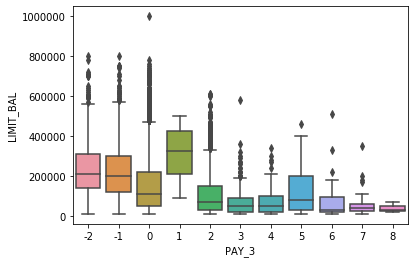


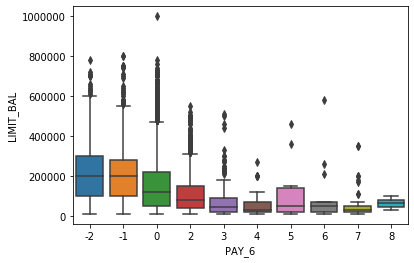
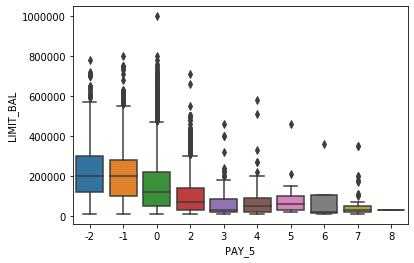
r



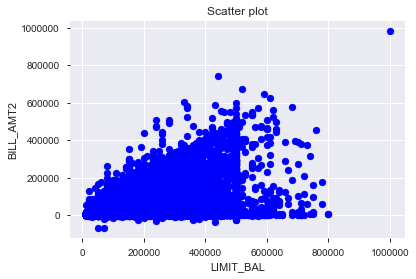
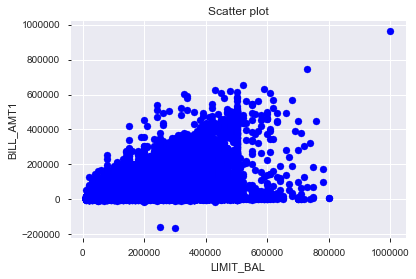
1. Comparing LIMIT BALANCE and Past PAY:

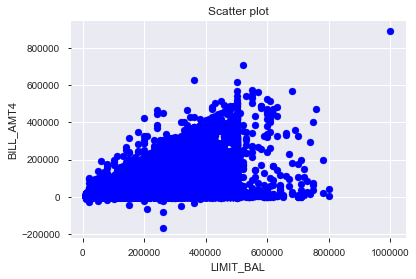
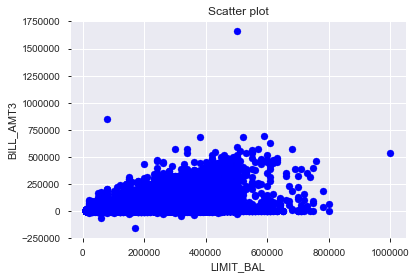


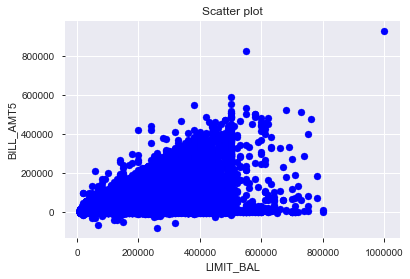
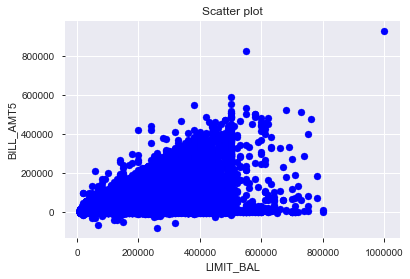




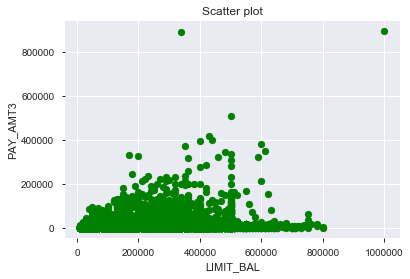
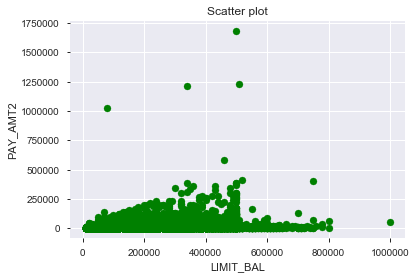
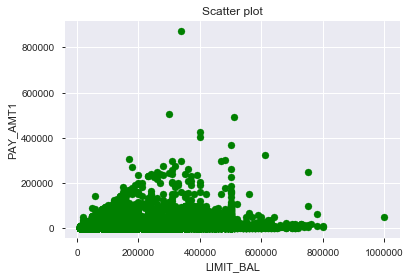
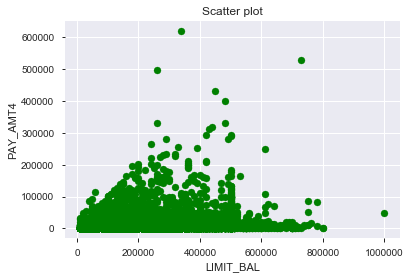
1. Comparing BILL AMOUNT within LIMIT BALANCE:

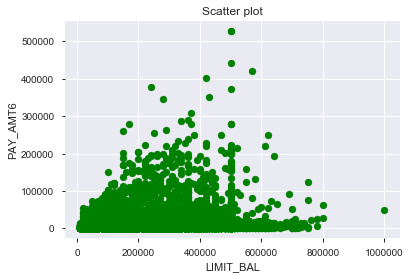
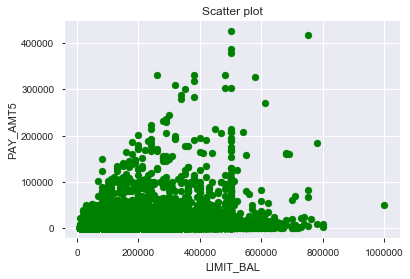




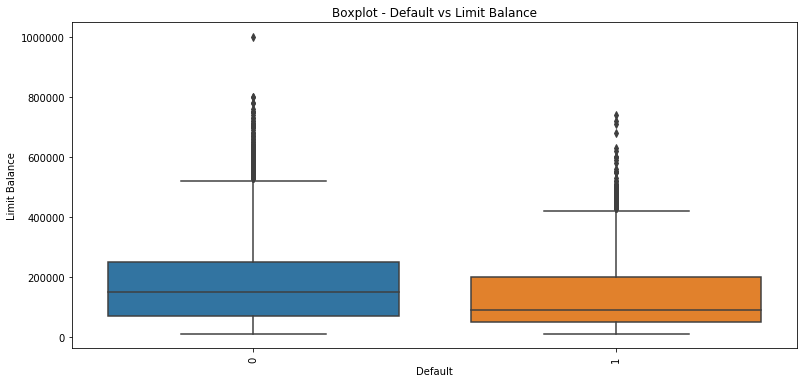


1. Comparing PAY AMT with LIMIT BALANCE:



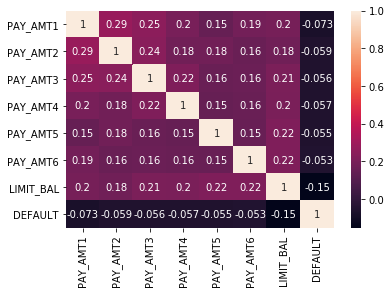
Default vs Limit balance:



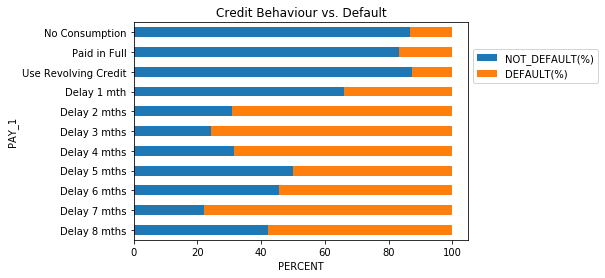
Correlation Between Default and Bill Amount :



Correlation Between Default and Pay Amount:



Credit Behaviour vs Default:



Credit behaviour, which shows their delay status, is the most important indicator for Default. When payment is delayed more than 2 months, the chances of default goes higher than 50%.

**QUANTITATIVE ANALYSIS**

The statistical analysis of the categorical features of the data is given below.

| **COUNT** | **MEAN** |  | **STD** | **MIN** | **25%** | **50%** | **75%** | **MAX** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **BILL\_AMT1** | 30000.0 |  | 51223.330900 | 73635.860576 | -165580.0 | 3558.75 | 22381.5 | 67091.00 | 964511.0 |
| **BILL\_AMT2** | 30000.0 |  | 49179.075167 | 71173.768783 | -69777.0 | 2984.75 | 21200.0 | 64006.25 | 983931.0 |
| **BILL\_AMT3** | 30000.0 |  | 47013.154800 | 69349.387427 | -157264.0 | 2666.25 | 20088.5 | 60164.75 | 1664089.0 |
| **BILL\_AMT4** | 30000.0 |  | 43262.948967 | 64332.856134 | -170000.0 | 2326.75 | 19052.0 | 54506.00 | 891586.0 |
| **BILL\_AMT5** | 30000.0 |  | 40311.400967 | 60797.155770 | -81334.0 | 1763.00 | 18104.5 | 50190.50 | 927171.0 |
| **BILL\_AMT6** | 30000.0 |  | 38871.760400 | 59554.107537 | -339603.0 | 1256.00 | 17071.0 | 49198.25 | 961664.0 |
| **PAY\_AMT1** | 30000.0 |  | 5663.580500 | 16563.280354 | 0.0 | 1000.00 | 2100.0 | 5006.00 | 873552.0 |
| **PAY\_AMT2** | 30000.0 |  | 5921.163500 | 23040.870402 | 0.0 | 833.00 | 2009.0 | 5000.00 | 1684259.0 |
| **PAY\_AMT3** | 30000.0 |  | 5225.681500 | 17606.961470 | 0.0 | 390.00 | 1800.0 | 4505.00 | 896040.0 |
| **PAY\_AMT4** | 30000.0 |  | 4826.076867 | 15666.159744 | 0.0 | 296.00 | 1500.0 | 4013.25 | 621000.0 |
| **PAY\_AMT5** | 30000.0 |  | 4799.387633 | 15278.305679 | 0.0 | 252.50 | 1500.0 | 4031.50 | 426529.0 |
| **PAY\_AMT6** | 30000.0 |  | 5215.502567 | 17777.465775 | 0.0 | 117.75 | 1500.0 | 4000.00 | 528666.0 |
| **LIMIT\_BAL** | 30000.0 |  | 167484.322667 | 129747.661567 | 10000.0 | 50000.00 | 140000.0 | 240000.00 | 1000000.0 |

**DATA COLUMN TREATMENT**

EDUCATION:

* Has category 5 and 6 that are unlabelled, and the category 0 is un-documented.
* The 0 (undocumented), 5 and 6 (label unknown) in EDUCATION can also be put in a 'Other' category (thus 4).

MARRIAGE:

* Has a label 0 that is undocumented.
* The 0 in MARRIAGE can be categorized as 'Other' (thus 3).

"Other" in education can be an education lower than the high school level.

"Other" in marriage could be, for example, "divorced".

PAY\_N Variables:

The PAY\_n variables indicate the number of months of delay and indicates "pay duly" with 0. Then what is -2? And what is -1? It seems that the label has to be adjusted to 0 for pay duly.

ID:

ID feature is not correlated with our target variable (Default) as the ID has been taken into consideration while surveying.

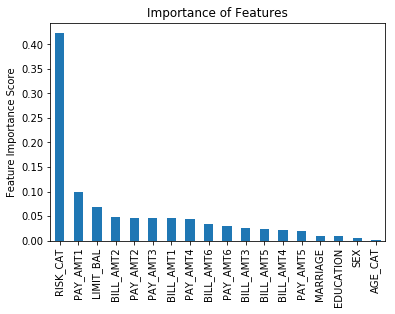
**Normalization:**

In the dataset we found that features had extreme values i.e our data was skewed so we had to perform normalization technique.

**Quantile transformation:**

This method transforms the features to follow a normal distribution. Therefore, for a given feature, this transformation tends to spread out the most frequent values. It also reduces the impact of outliers: this is therefore a robust pre-processing scheme.

**Feature Selection**



We have used Random Forest for feature importance and we have selected the features as given below:

* RISK\_CAT
* LIMIT\_BAL
* PAY\_AMT1
* PAY\_AMT2
* PAY\_AMT3
* PAY\_AMT4
* PAY\_AMT5
* PAY\_AMT6
* BILL\_AMT1
* BILL\_AMT2
* BILL\_AMT3
* BILL\_AMT4
* BILL\_AMT5
* BILL\_AMT6

**Feature Extraction**

1. Risk Category:

We have added all the Pay Status feature from April to September

in which we have discretized on the scale of

* -20 to -10 as ‘Low’
* -10 to 0 as ‘Medium’
* 0 to 40 as ‘High

Then we have mapped this into a feature named RISK\_CAT.

1. Age Category:

We have distributed the age scale on the basis of three category i.e

* 20 to 40 as ‘Young’
* 40 to 60 as ‘Middle’
* 60 to 81 as ‘Senior’

Then we have mapped this into a feature named AGE\_CAT.

**MODEL DEVELOPMENT**

After the treatment of the data, we applied multiple classification algorithms for making model. This list includes Logistic Regression, Random Forest. Apart from these ensemble technique that we used are Gradient boosting technique and

The performance metrics used to justify the models were F1-Score

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **FN** | **F1 Score** | **Recall** | **AUC** |
| Logistic (Base model) | 0.79 | 1339 | 0.27 | 0.18 | 0.75 |
| Logistic (After Changing Threshold) | 0.70 | 483 | 0.50 | 0.70 | 0.75 |
| Random Forest (Basic model) | 0.78 | 1208 | 0.35 | 0.26 | 0.70 |
| Random Forest (After Hyper Parameter Tuning) | 0.79 | 1184 | 0.37 | 0.27 | 0.74 |
| GBM | 0.79 | 1150 | 0.38 | 0.30 | 0.73 |
| X-G Boost (Max depth = 80, Learning Rate = 0.05, n estimator = 144) | 0.79 | 1134 | 0.39 | 0.31 | 0.74 |

**CONCLUSION**

Following the machine learning pattern, we have selected features (both predictors and target) distributions, built models and evaluated the performances of each model.  
 Different models are used, including logistic regression, Random Forest Classifier, GBM and XGBoost.   
The techniques in evaluating the performances of the models are cross validation, precision score, recall score, F1 score, ROC\_AUC and Confusion Matrix.  
After hyper-tuning the parameters in all models, Logistic regression model out performs others by adjusting the threshold.