Capstone Project –

Credit Card Default Prediction

Presented By :-
Aniket Satpute
Kaiwalya Zankar
(AlmaBetter Trainee)

Content

- Introduction
- Scope of project
- Problem Statement
- Data summery
- Process flow
- Library used
- Data Preprocessing
- EDA
- Statistical inference
- Feature Engineering
- Machine learning : Classification model
- Challenges
- Conclusion
- Q&A

Introduction

- Credit risk has traditionally been the greatest risk among all the risks that the banking and credit card industry are facing, and it is usually the one requiring the most capital.
- Despite machine learning and big data have been adopted by the banking industry, the current applications are mainly focused on credit score predicting. The disadvantage of heavily relying on credit score is banks would miss valuable customers who come from countries that are traditionally underbanked with no credit history or new immigrants who have repaying power but lack credit history.
- Due to the scope of the project and lack of computational resources, this analysis is not intended to be exhaustive, we only applied 3 classification machine learning models

Scope of the Project-

- The purpose of this project is to conduct quantitative analysis on credit card default risk by applying 3 classification machine learning models.
- Despite machine learning and big data have been adopted by the banking industry, the current applications are mainly focused on credit score predicting.
- Despite machine learning and big data have been adopted by the banking industry, the current applications are mainly focused on credit score predicting.
- Due to the scope of the project and lack of computational resources, this analysis is not intended to be exhaustive, we only applied 3 classification machine learning models

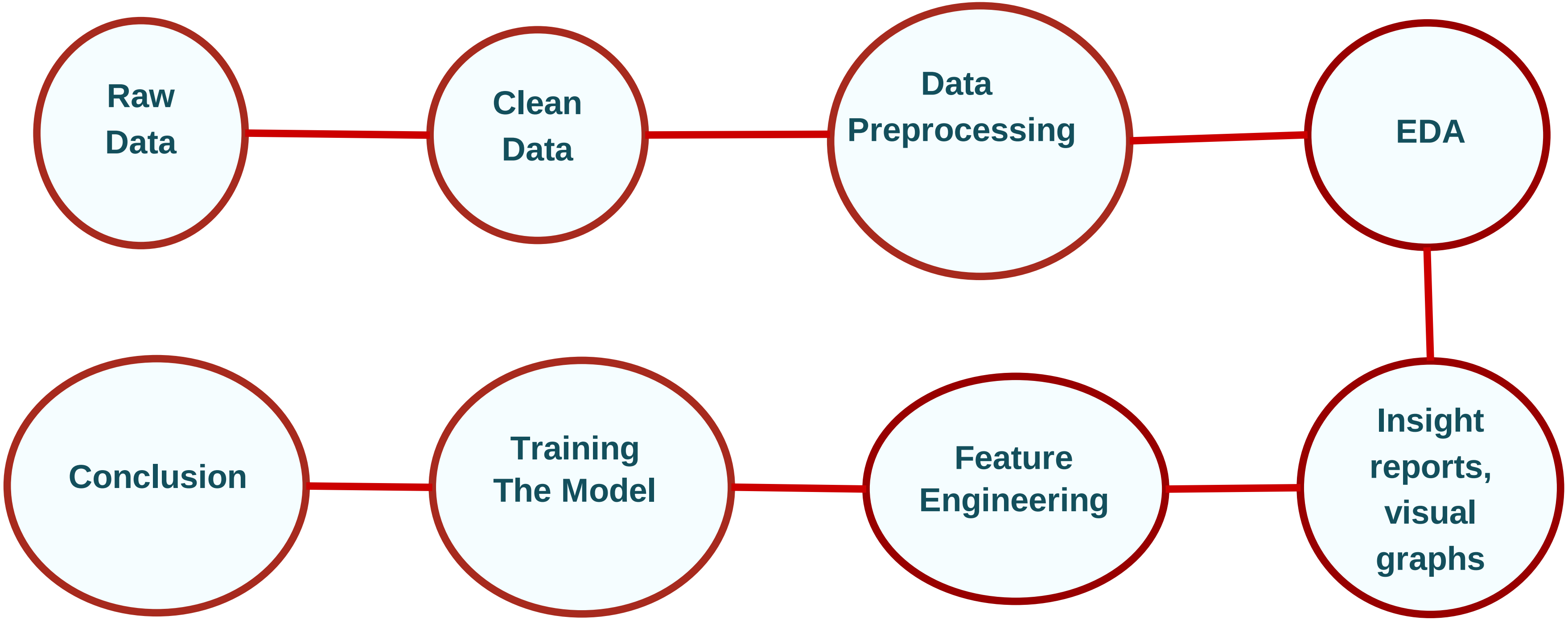
Problem Statement

- This project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. We can use the K-S chart to evaluate which customers will default on their credit card payments

Data Summary

- This research employed a binary variable, default payment (Yes = 1, No = 0), as the respon se variable. This study reviewed the literature and used the following 23 variables as expl anatory variables:
- X1: Amount of the given credit (NT dollar): it includes both the individual consumer credi t and his/her family (supplementary) credit.
- X2: Gender (1 = male; 2 = female).
- X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- X4: Marital status (1 = married; 2 = single; 3 = others).
- X5: Age (year).
- X6,X7,X8,X9,X10,X11: the repayment status in September, August,July.June,may,april2005 respectively. The measurement scale for the repayment status is: -
  1 = pay duly; 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.
- X12,X13,X14,X15,X16,X17: Amount of bill statement (NT dollar) in September, August, July, June, May, April 2005 respectively
- X18,X19,X20,X21,X22,X23: Amount of previous payment (NT dollar).amount paid in Septem ber, August, July, June, May, April ,205 respectively

Process Flow-



Libraries used-

1. numpy
2. Pandas
3. scipy
4. matplotlib.pyplot
5. seaborn
6. imblearn
7. Sklearn
8. statsmodel
9. Math
10. Xgboost
11. warnings

Data Preprocessing

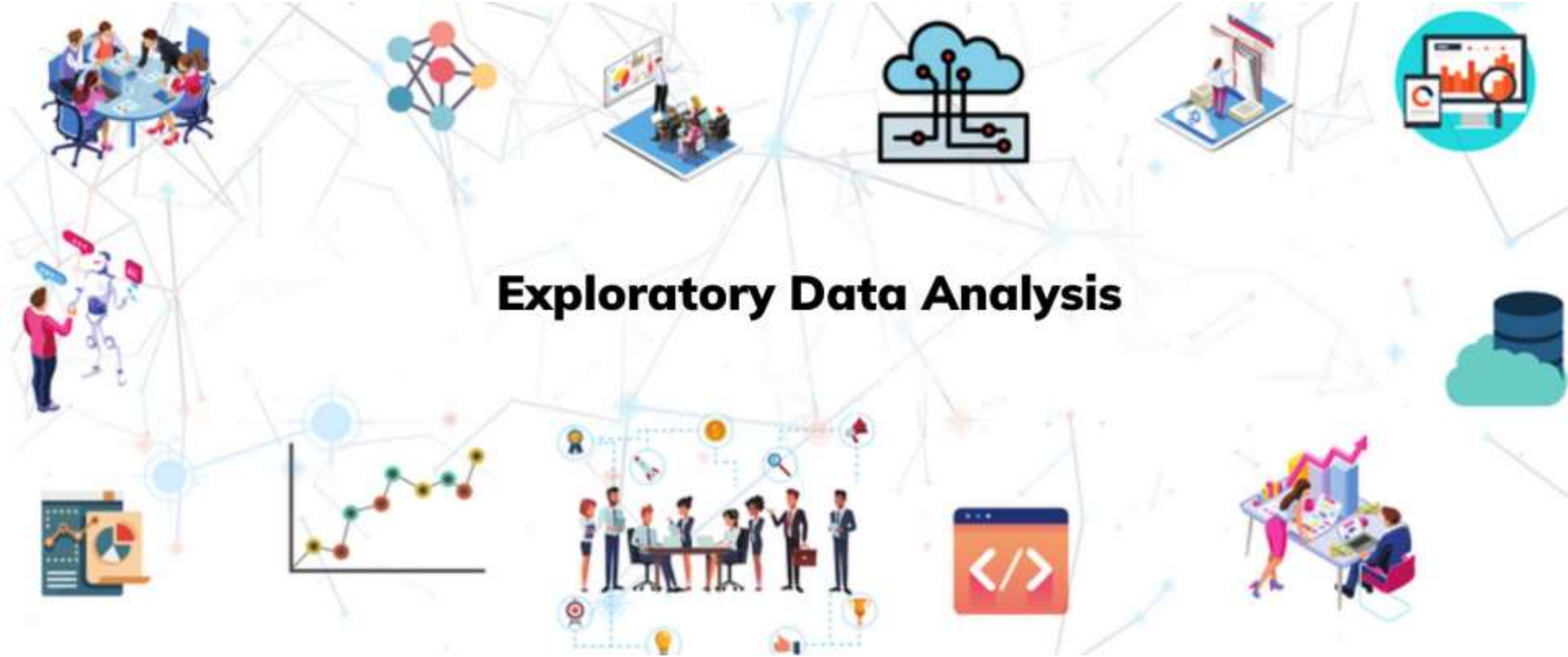
Checking the null values for cleaning the Dataset for further analysis.

id	0	bill_amt2	0
limit_bal	0	bill_amt3	0
sex	0	bill_amt4	0
education	0	bill_amt5	0
marriage	0	bill_amt6	0
age	0	pay_amt1	0
pay_1	0	pay_amt2	0
pay_2	0	pay_amt3	0
pay_3	0	pay_amt4	0
pay_4	0	pay_amt5	0
pay_5	0	pay_amt6	0
pay_6	0	default	0
bill_amt1	0		

We do not see any null values in the dataset.

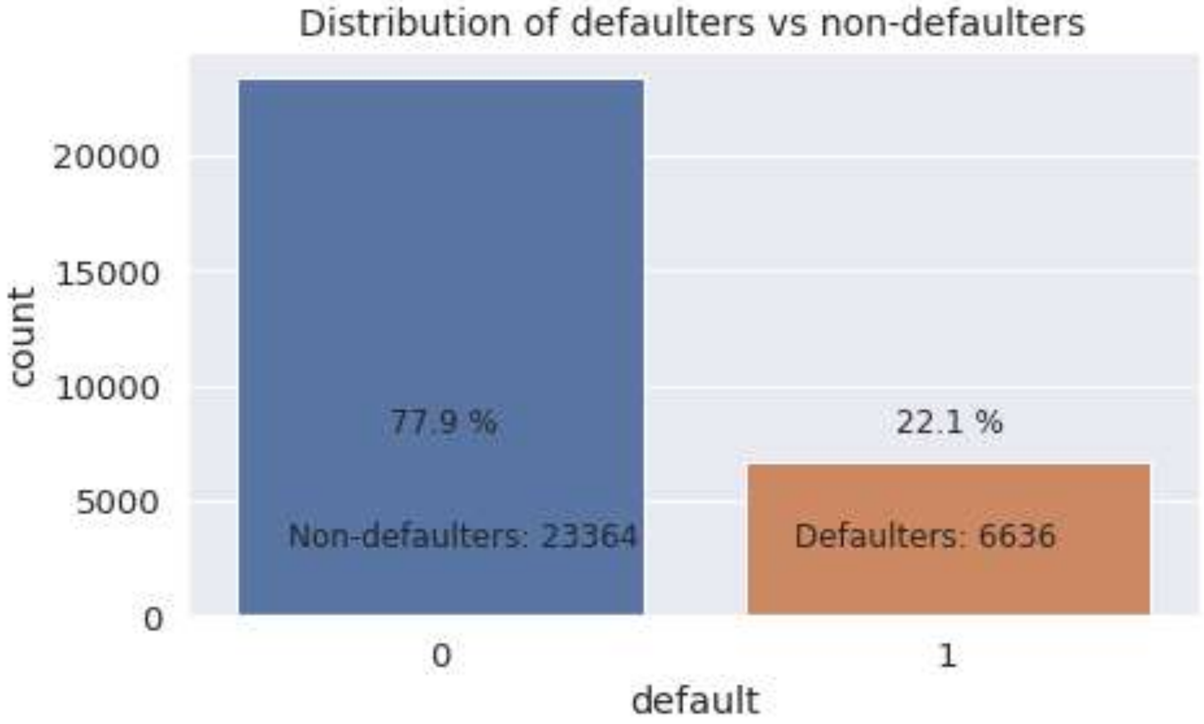
Checking the unique values for Analyzing the dataset for Further analysis.

id	30000	bill_amt2	22346
limit_bal	81	bill_amt3	22026
sex	2	bill_amt4	21548
education	7	bill_amt5	21010
marriage	4	bill_amt6	20604
age	56	pay_amt1	7943
pay_1	11	pay_amt2	7899
pay_2	11	pay_amt3	7518
pay_3	11	pay_amt4	6937
pay_4	11	pay_amt5	6897
pay_5	10	pay_amt6	6939
pay_6	10	Default	2
bill_amt1	22723		



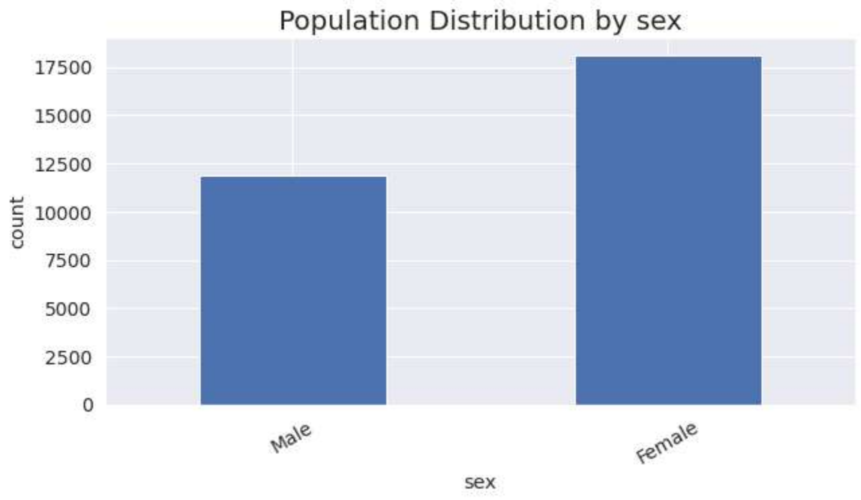
Exploratory Data Analysis

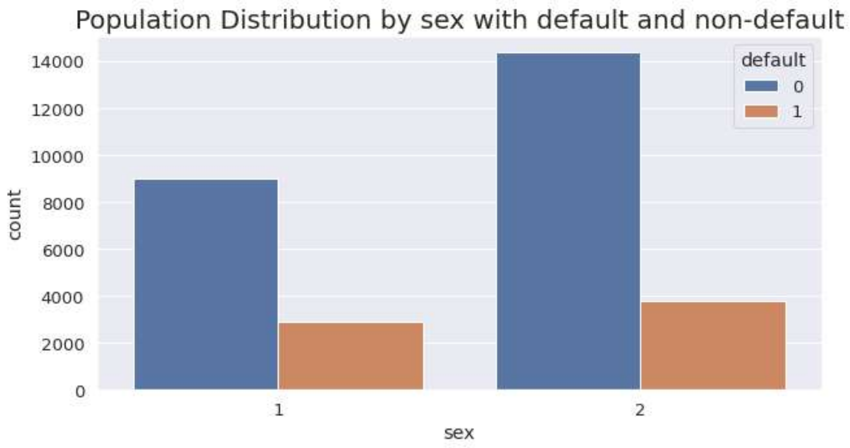
Distribution of defaulters vs non-defaulters



Relationship between the variables and default

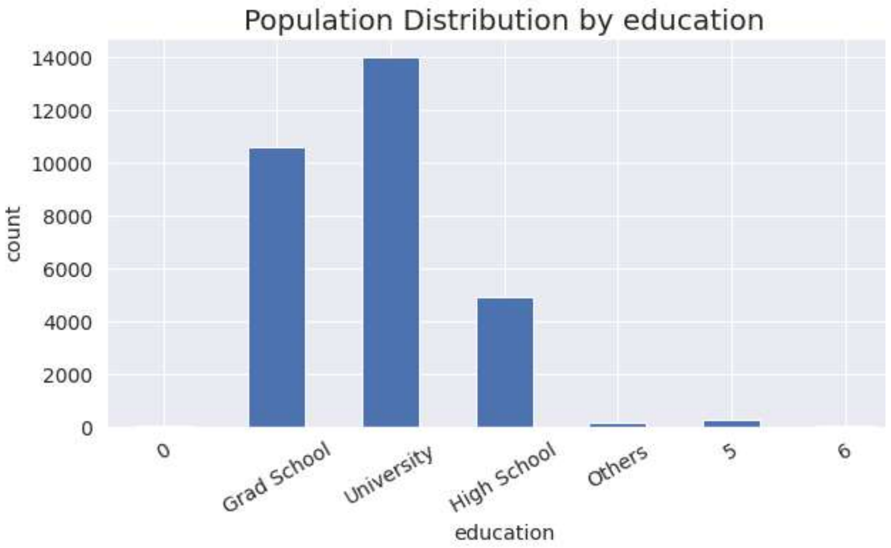
Is default proportion affected by gender?

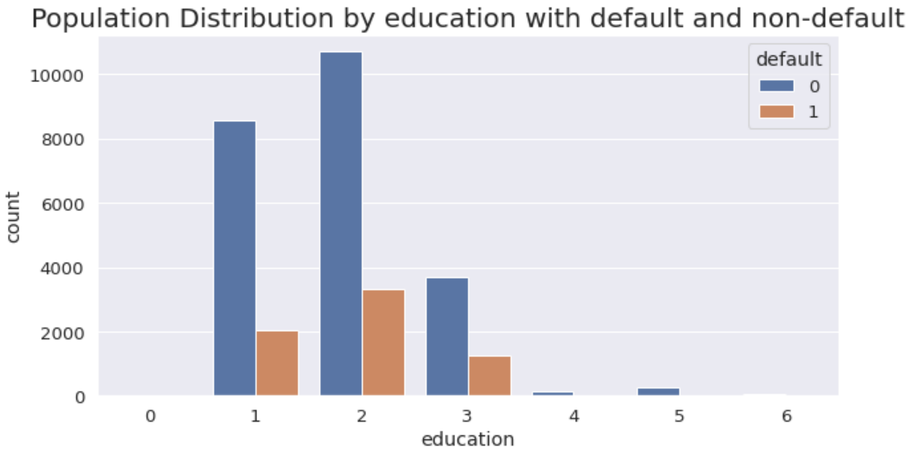




- Although there are more female credit card holders, the default proportion among men is higher. I will do a hypothesis test to see if the difference is statistically significant.

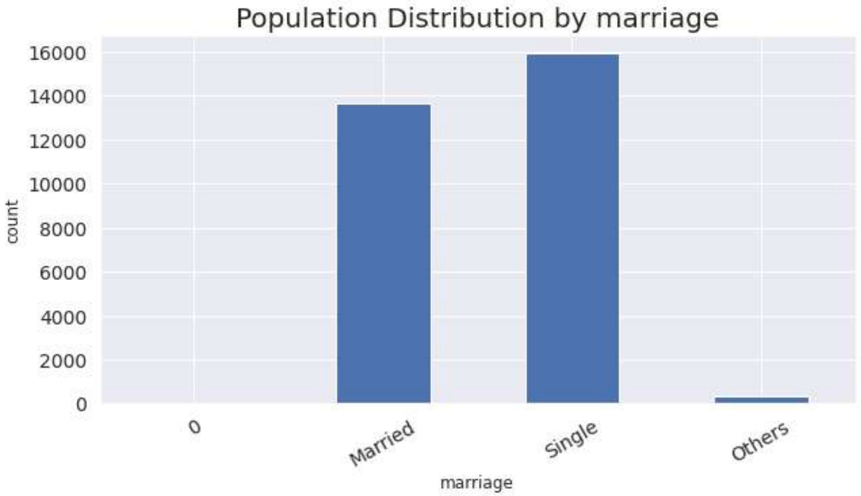
Is default proportion affected by education?

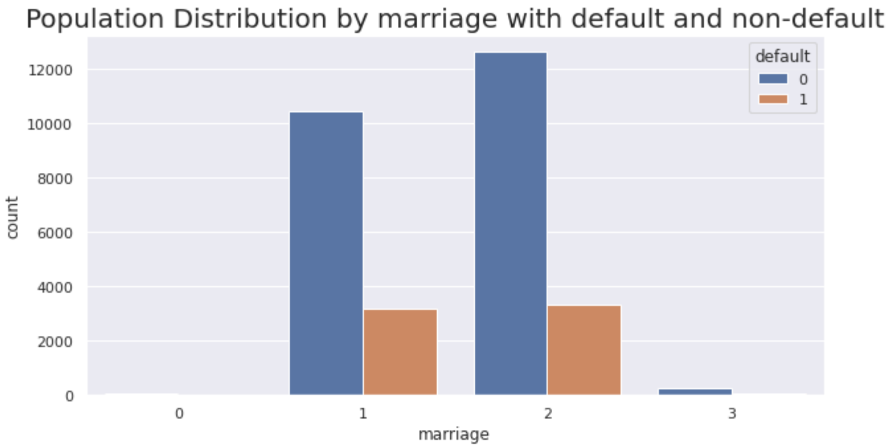




- The default proportion decreases with higher education level.

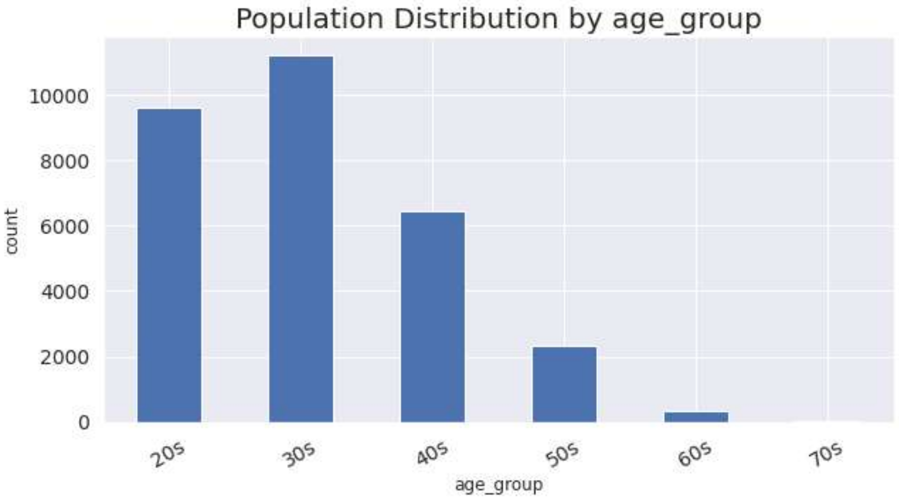
Is default proportion affected by marital status?

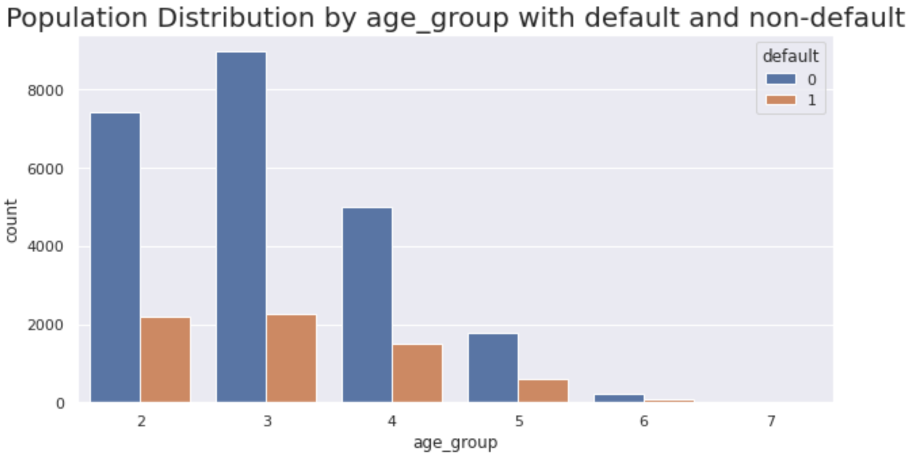




- Married people have higher default proportions than single folks. While there are intuitiv e arguments for and against it, closer inspection is needed. For example, is there a differ ence between married men and married women?

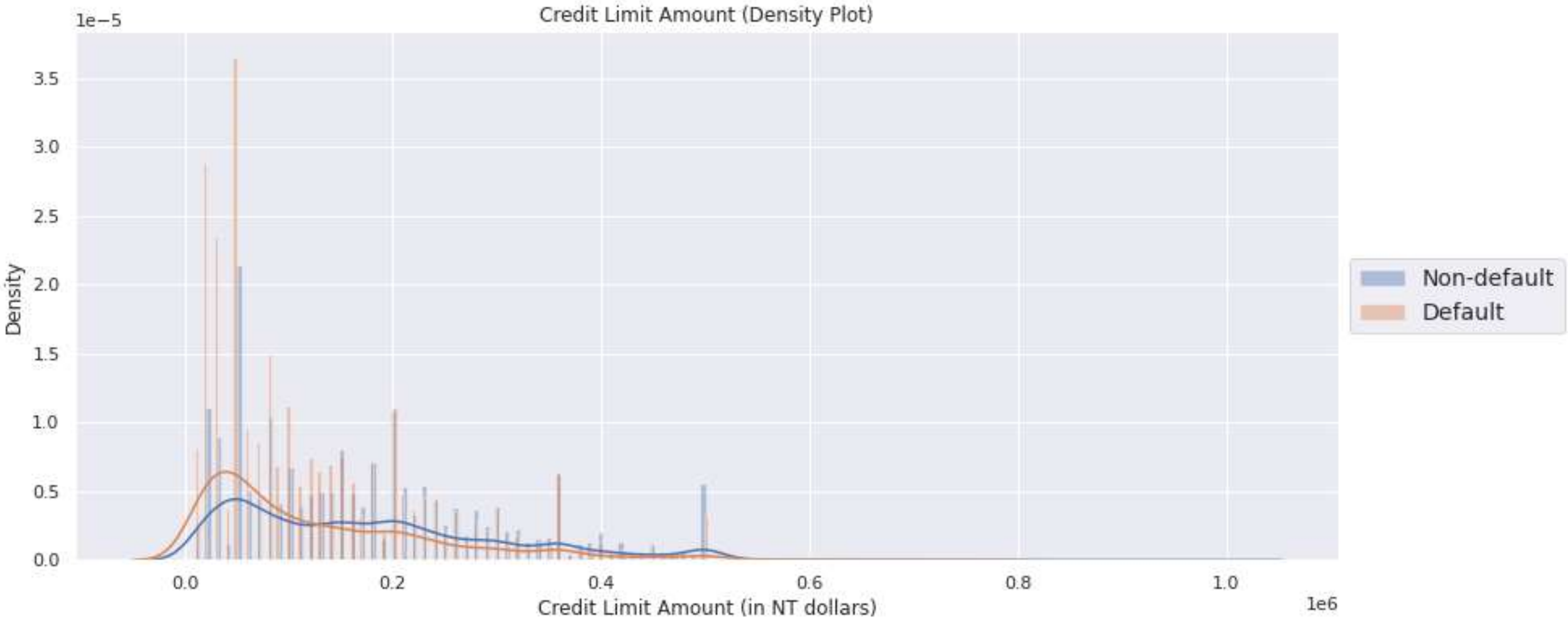
Is the proportion of defaults correlated with age?





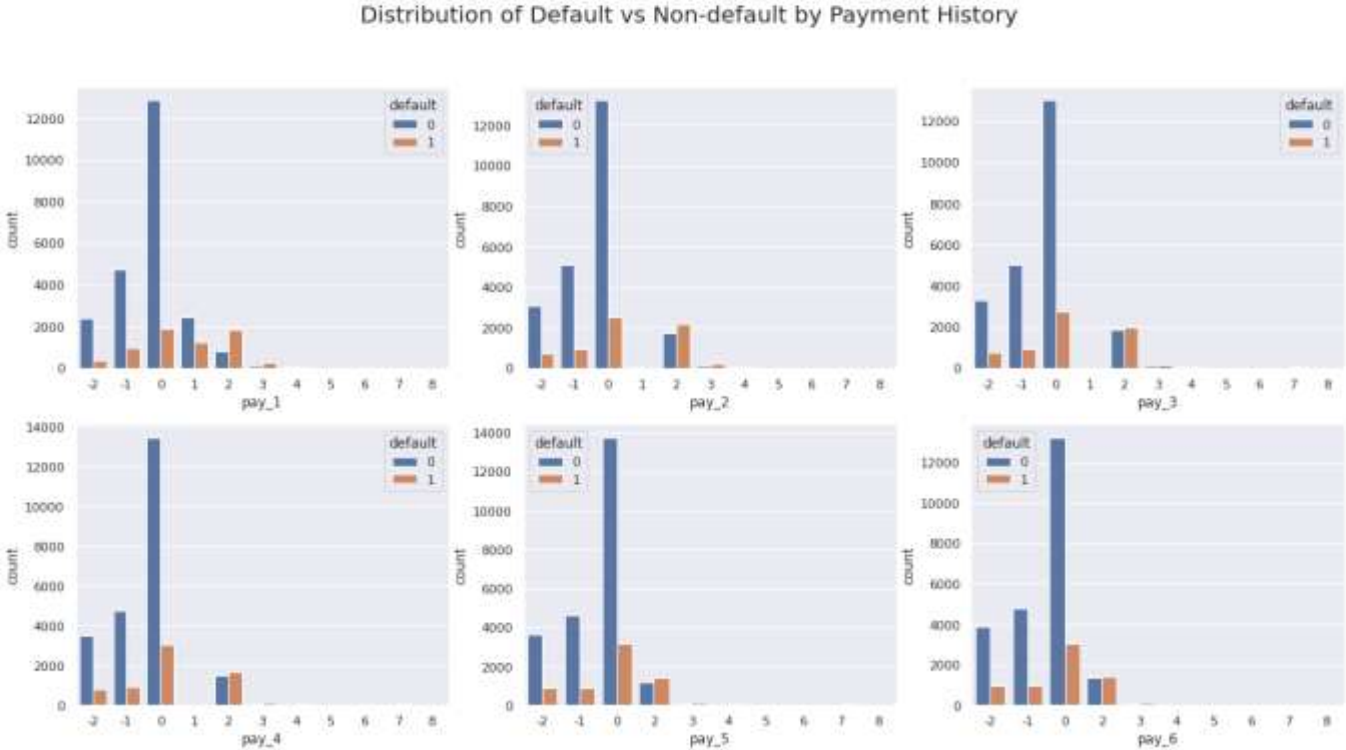
The age group of 30s has the high number of counts of defaults.

Is the default proportion affected by credit limit?



- It seems that people with higher credit limit have significantly lower default proportion.

Is the default proportion affected by history of past repayment status?



- I notice that if the person has defaulted for 2 months or more in the past two months, there is a very high chances of them defaulting.

Correlation between the variables

- High correlation among the payment history features and the bill amount features.

Plotting the bill amount density and their scatter plot

The distribution of the bill amounts are skewed.

Statistical Inference.

Does the gender affect the default rate? I have tried to answer this question with hypothesis test.

I have use the significance level of $\alpha = 0.05$.

The bounds of the confidence interval are given by $\frac{\alpha}{2}, 1 - \frac{\alpha}{2} = [0.025, 0.975]$

I state the null and alternate hypotheses:

$H_0 : p_m = p_w$

$H_a : p_m \neq p_w$

As the p value is 0.05 the null hypothesis is rejected.

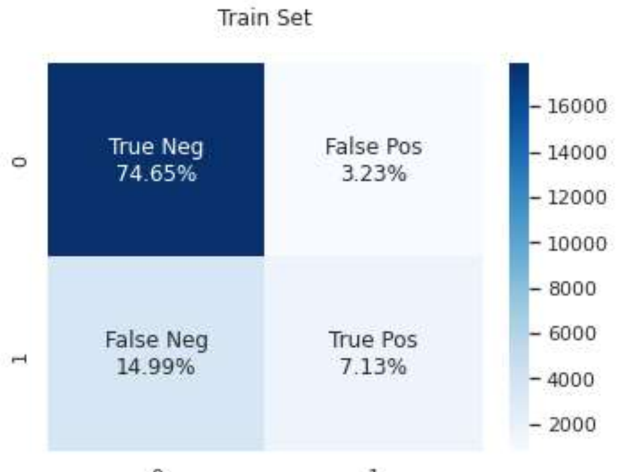
Feature Engineering

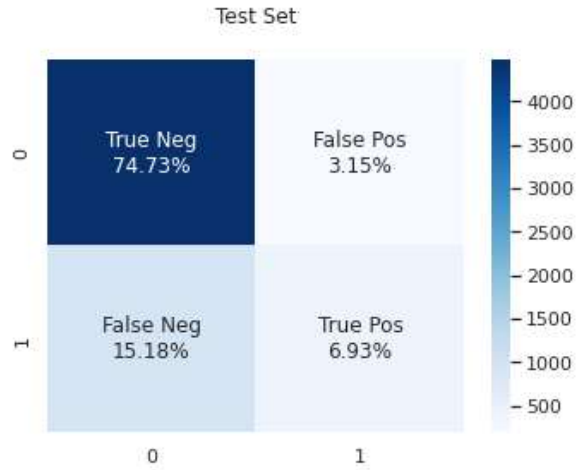
I have added some more features like as follows:

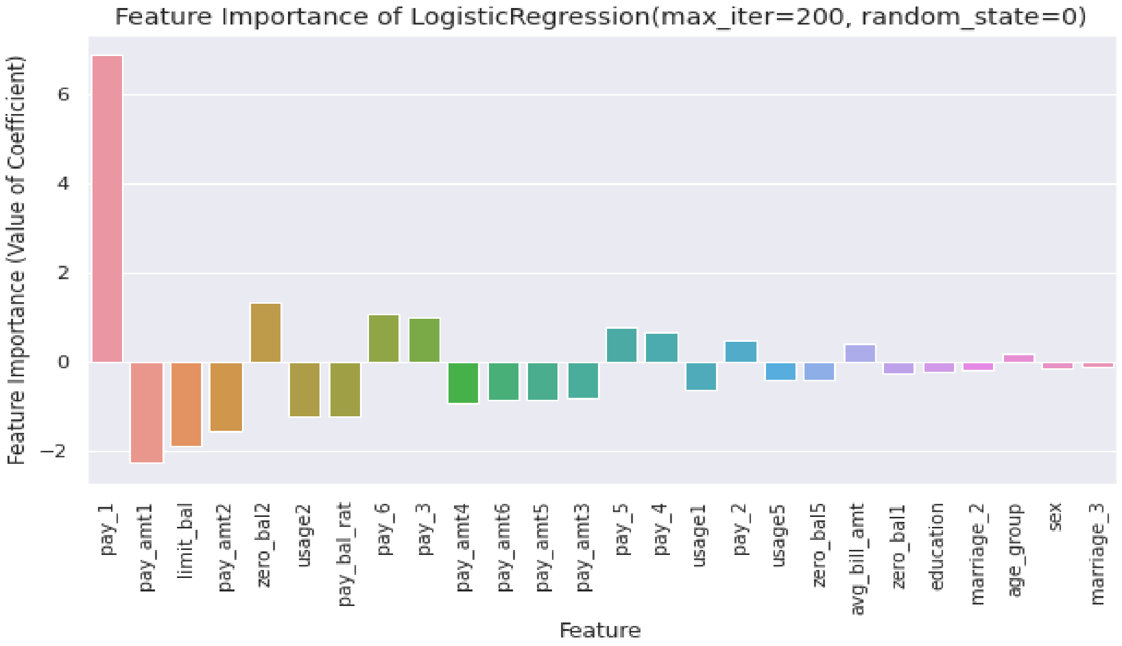
- avg_default
- avg_bill_amt
- avg_pay_amt
- pay_bill_rat
- bill_bal_rat
- pay_bal_rat
- overdraft

Machine learning : Classification models

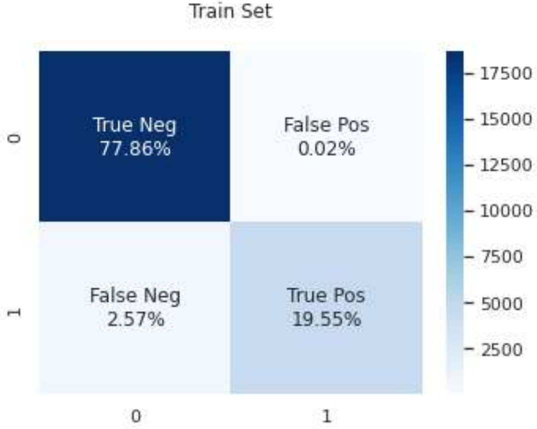
Logistic Regression model

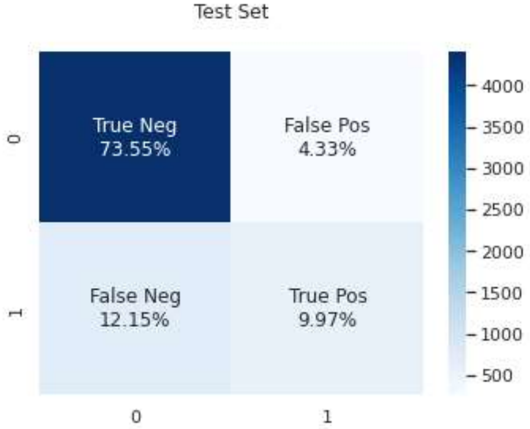


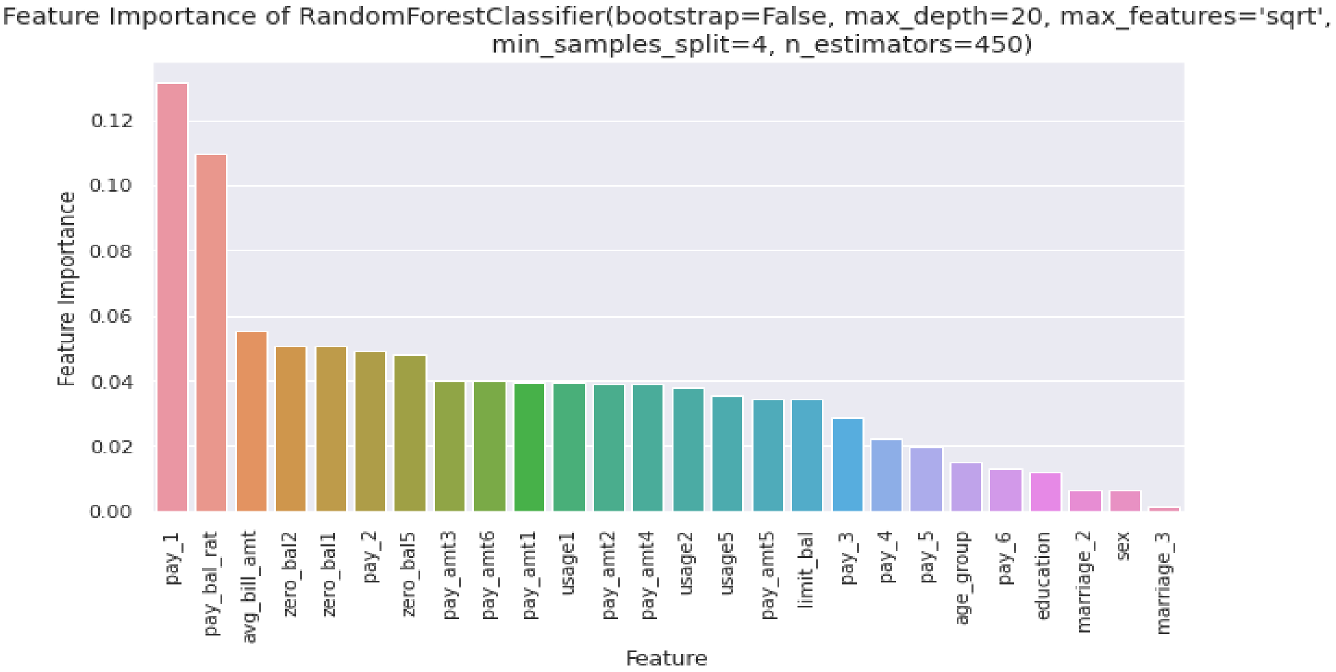




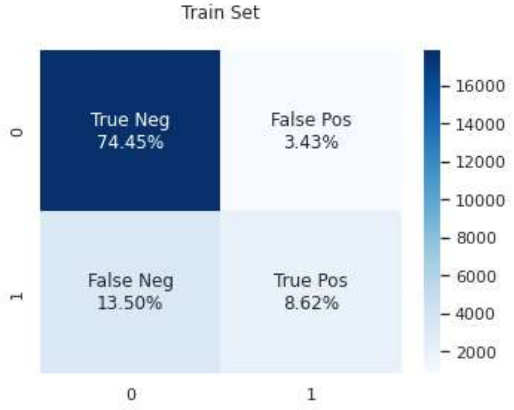
Random-Forest Classifier

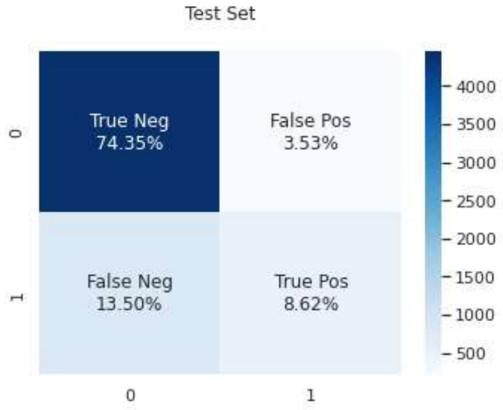


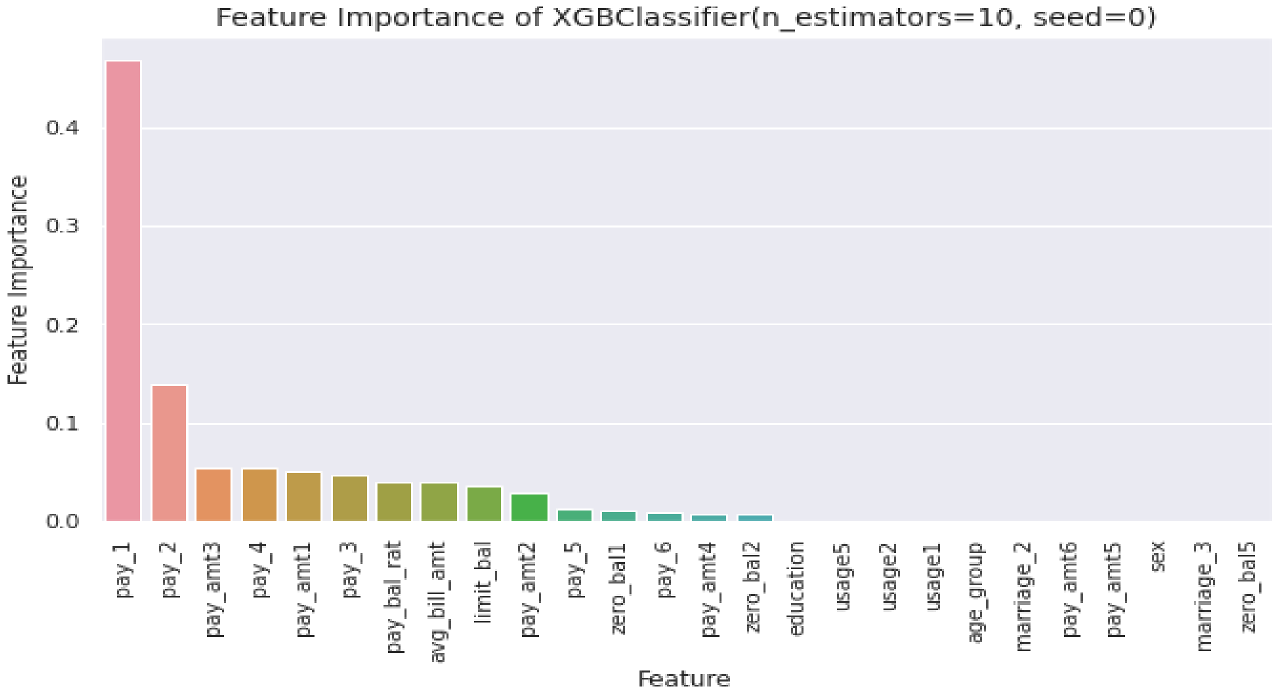




Gradient Boosting Classifier







ML Models and Metrics

	Model	Accuracy	Precision	recall	F1_score	AUC_ROC
0	Logistic Regression (original data)	0.778500	0.000000	0.000000	0.000000	0.641673
1	Logistic Regression (engineered data)	0.816667	0.687603	0.313489	0.430642	0.755872
2	Untuned Random Forest Model	0.833500	0.693396	0.443105	0.540690	0.795104
3	Tuned Random Forest Model	0.835167	0.696970	0.450641	0.547368	0.799170
4	Gradient Boosting Classifier	0.829667	0.709191	0.389601	0.502918	0.776989

Challenges

- For finding what should be the dependent variable
- From the given dataset Filtering discrete values.
- Features to be selected to get required output for ease the further analysis.

Conclusion

We used different type of classification algorithms to train our model like, logistic regression, Random Forest Classifier, Gradient Boosting Classifier also found the important features for training the model. Out of them random forest with tuned hyperparameters gave the best result.

Q & A