DATA 240
Predict Default of Credit Card Payment

TEAM 5:
Lakshmi Naga Meghana Polisetty
Sakshi Jain
Sakshi Tongia
Snehal Dashrath Karad
Vaibhav Yalla

Motivation

- According to the GOBanking Rates survey, about 6%, that is 14 million Americans have credit card debt over $10,000
- 33% of Americans believe that it would take more than two years to clear their credit card debt
- Increase in credit-card debt acts negatively for the customer and also reduces the bottom line for the lenders
- Using Machine Learning to predict credit card default will save the lenders from issuing credit card to risky customers or to identify if a customer is going to default a payment in future and thereby can reduce their credit limit

Feature Information

There are 25 variables:

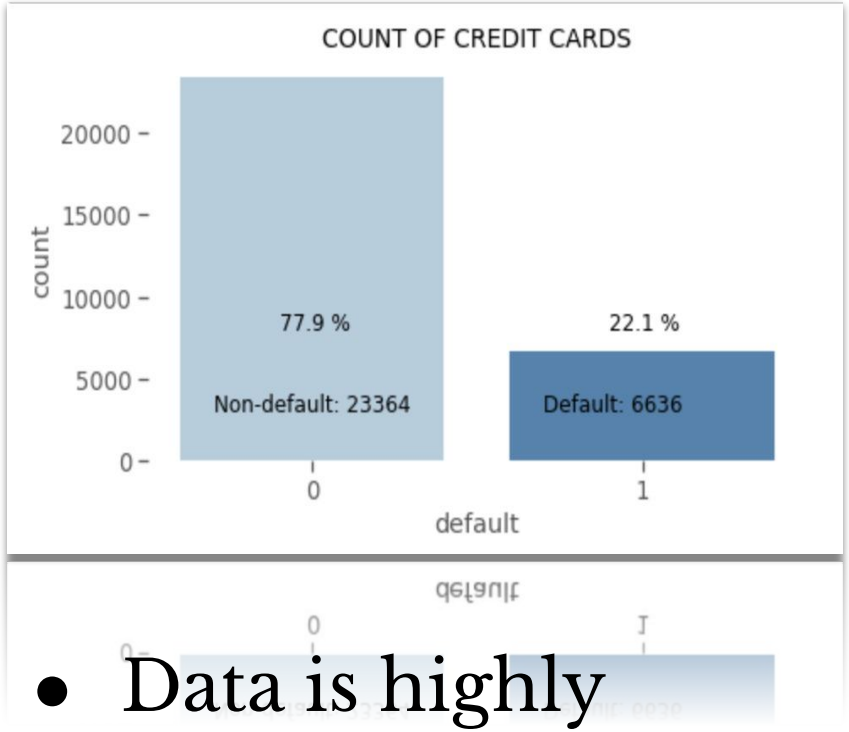
- ID: ID of each client
- LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years
- PAY_0: Repayment status in September, 2005 (-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)
- PAY_2: Repayment status in August, 2005 (scale same as above)
- PAY_3: Repayment status in July, 2005 (scale same as above)
- PAY_4: Repayment status in June, 2005 (scale same as above)
- PAY_5: Repayment status in May, 2005 (scale same as above)
- PAY_6: Repayment status in April, 2005 (scale same as above)
- BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
- BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
- BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
- BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
- BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
- BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
- PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
- PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
- PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
- PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
- PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
- PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
- default.payment.next.month: Default payment (1=yes, 0=no)

Raw Data

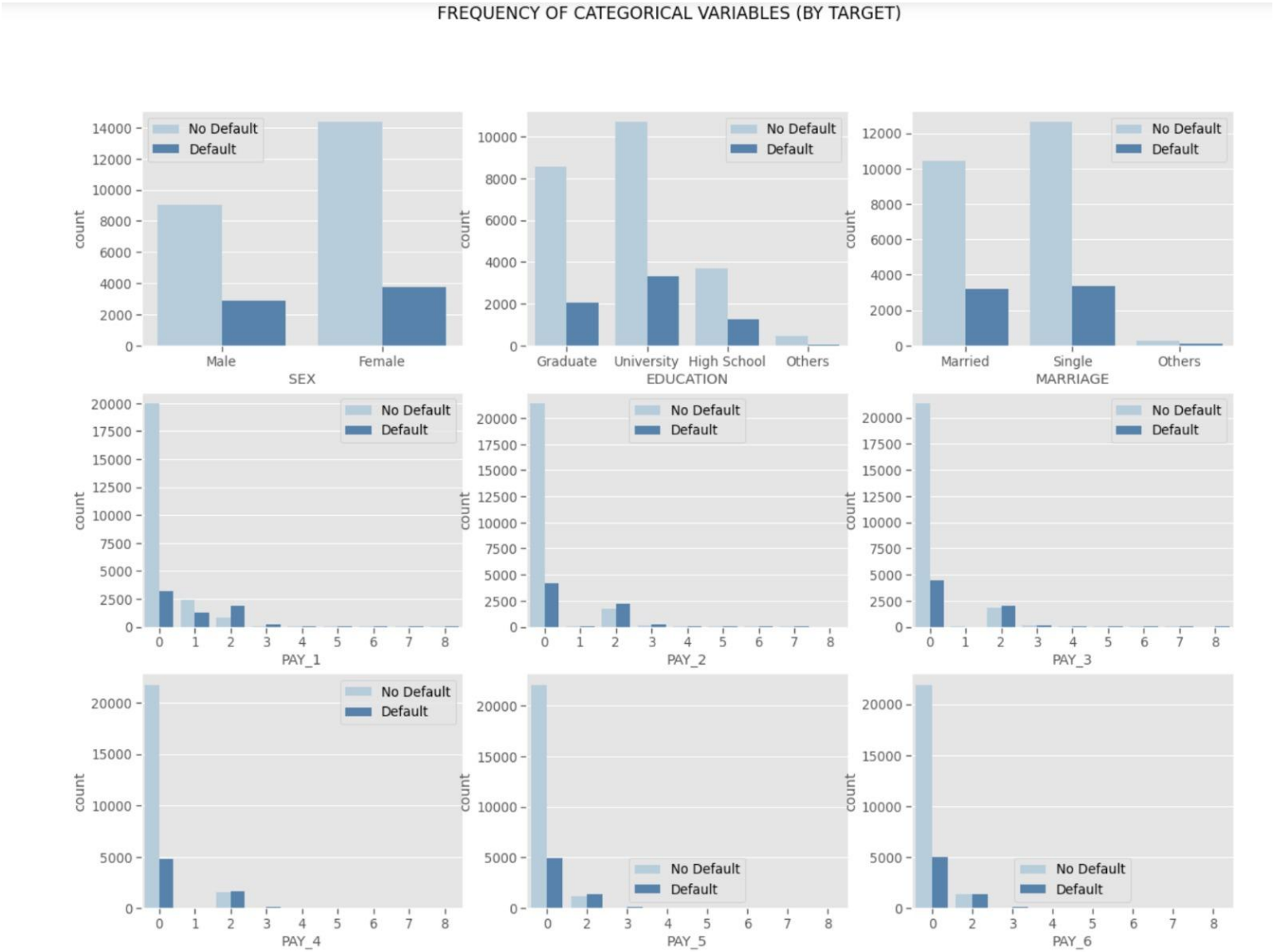
ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default payment next month
1	20000	2	2	1	24	2	2	-1	-1	...	0	0	0	0	689	0	0	0	0	1
2	120000	2	2	2	26	-1	2	0	0	...	3272	3455	3261	0	1000	1000	1000	0	2000	1
3	90000	2	2	2	34	0	0	0	0	...	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
4	50000	2	2	1	37	0	0	0	0	...	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
5	50000	1	2	1	57	-1	0	-1	0	...	20940	19146	19131	2000	36681	10000	9000	689	679	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
29996	220000	1	3	1	39	0	0	0	0	...	88004	31237	15980	8500	20000	5003	3047	5000	1000	0
29997	150000	1	3	2	43	-1	-1	-1	-1	...	8979	5190	0	1837	3526	8998	129	0	0	0
29998	30000	1	2	2	37	4	3	2	-1	...	20878	20582	19357	0	0	22000	4200	2000	3100	1
29999	80000	1	3	1	41	1	-1	0	0	...	52774	11855	48944	85900	3409	1178	1926	52964	1804	1
30000	50000	1	2	1	46	0	0	0	0	...	36535	32428	15313	2078	1800	1430	1000	1000	1000	1

30000 rows × 25 columns

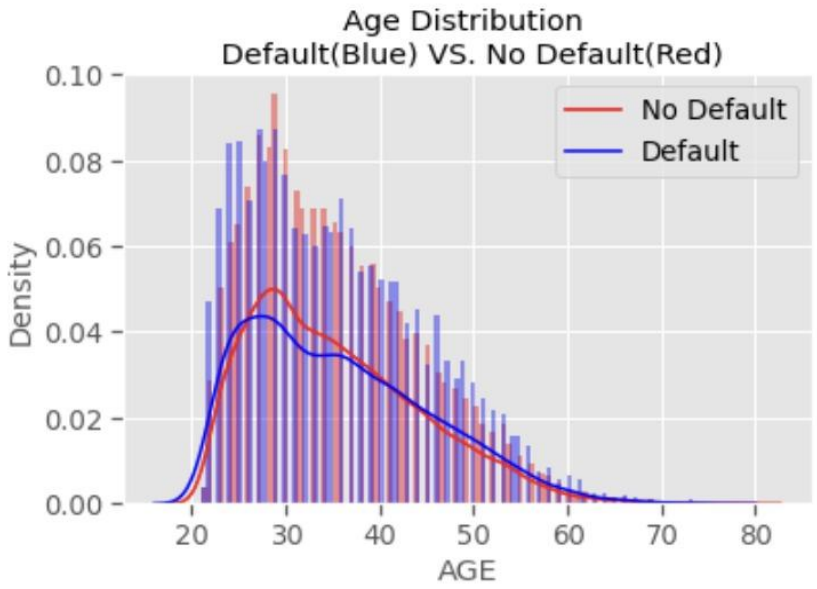
Exploratory Data Analysis



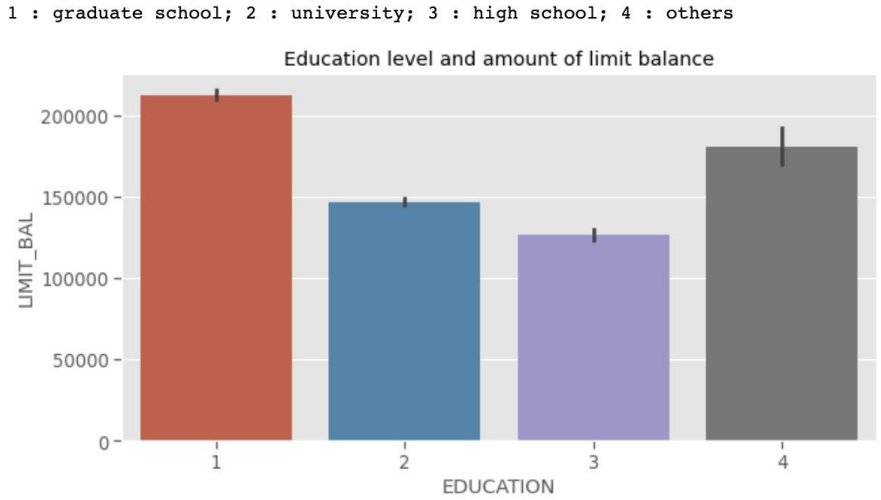
- Data is highly imbalanced with a ratio of about 78 : 22 percent

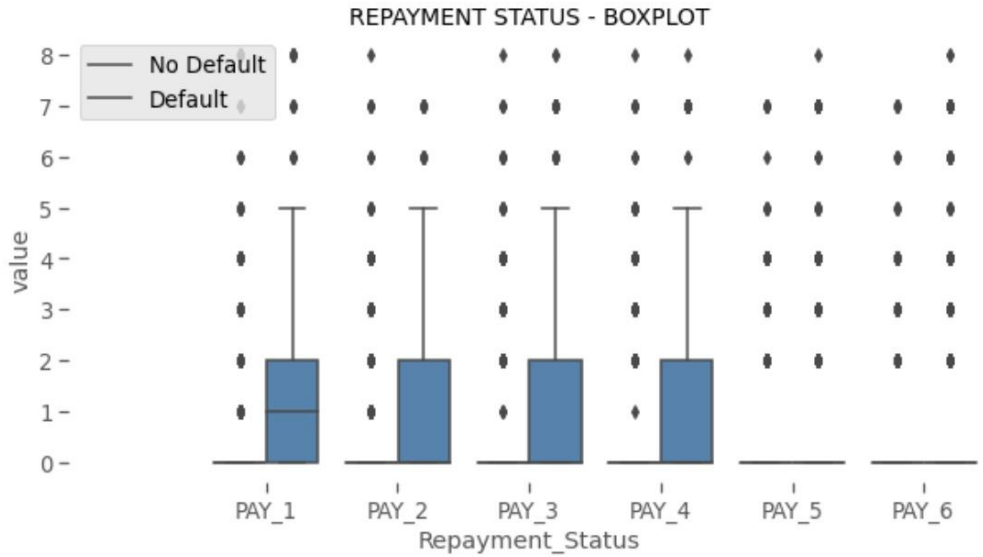


Exploratory Data Analysis

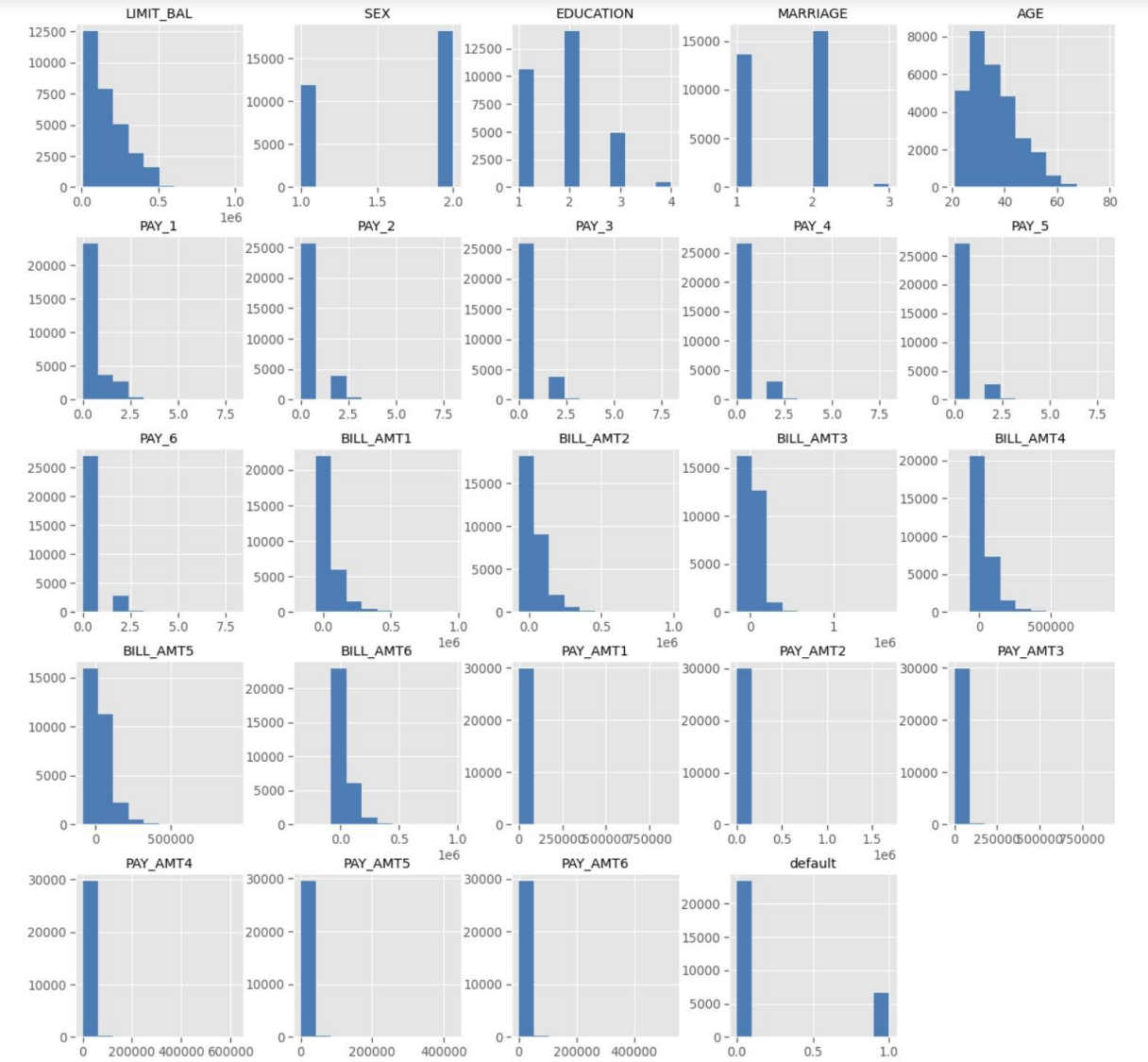


- As age increases to 30, the probability of default increases.
- Meanwhile, when clients are over 30, the probability decreases when aging.





Data Summary



Distributions for all Features :

1. The average value for the amount of credit card limit is 167K NT dollars. The standard deviation is 129K NT dollars, ranging from 10K to 1M NT dollars.

2. Education level is mostly Graduate school (1) and University (2).

3. Most of the clients are either married or single (less frequent in the other status).

4. Average age is 35.5 years, with a standard deviation of 9.2 years.

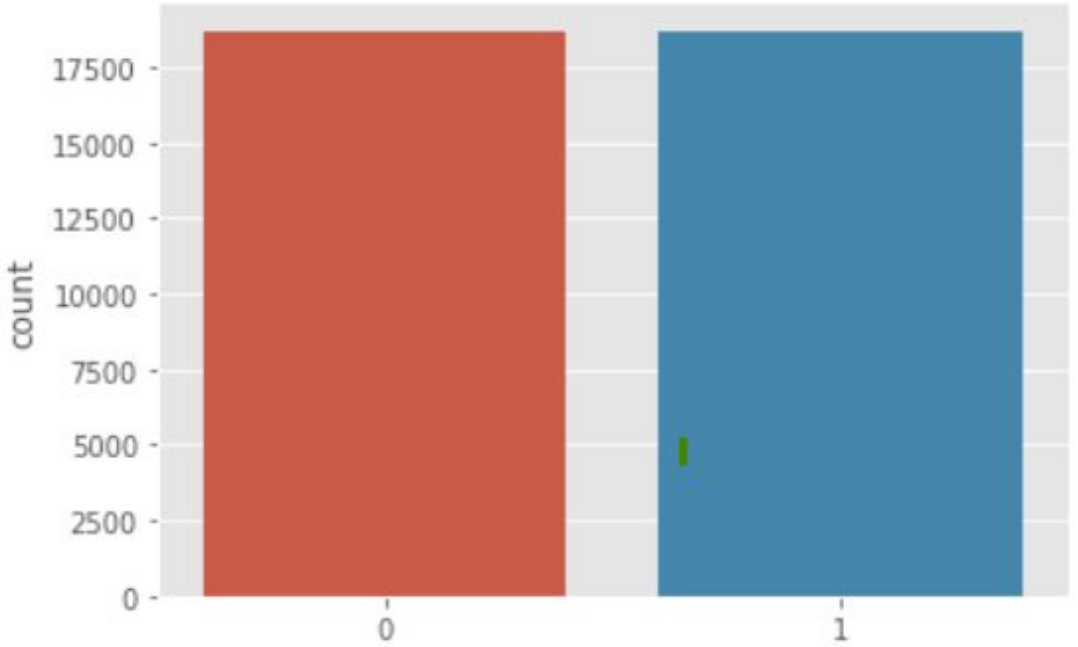
Data Transformation

Applying the MinMax Scaler:

- Transformation to normalize values

(If the distribution is not Gaussian or the standard deviation is very small, the min-max scaler works better.)

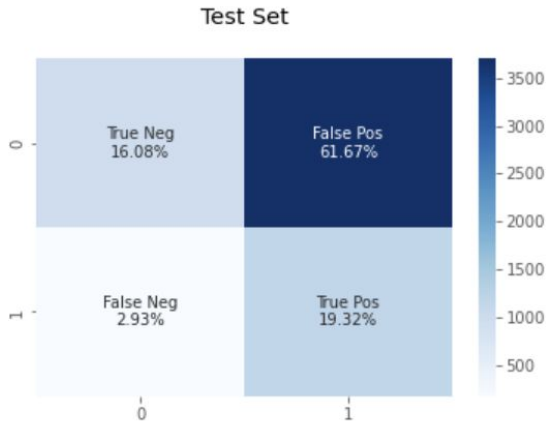
Applying SMOTE to Balance the Dataset:



Data Transformation

ROC_AUC_Score of train set is 0.9572243636763337.
ROC_AUC_Score of test set is 0.5428839352738338.
F1 of train set is 0.8831.
F1 of test set is 0.3742.

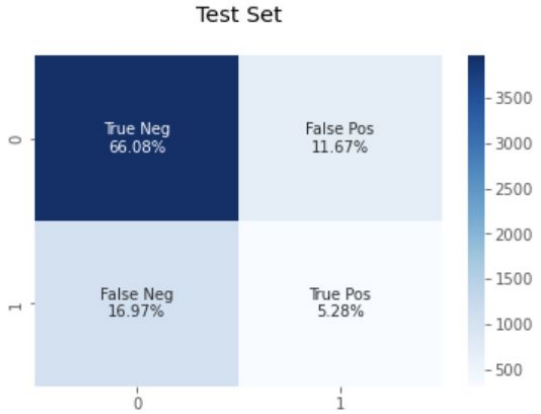
	precision	recall	f1-score	support
0	0.85	0.21	0.33	4665
1	0.24	0.87	0.37	1335
accuracy			0.35	6000
macro avg	0.54	0.54	0.35	6000
weighted avg	0.71	0.35	0.34	6000



XGBoost

ROC_AUC_Score of train set is 0.9997951309391232.
ROC_AUC_Score of test set is 0.5915151880085585.
F1 of train set is 0.9948.
F1 of test set is 0.2696.

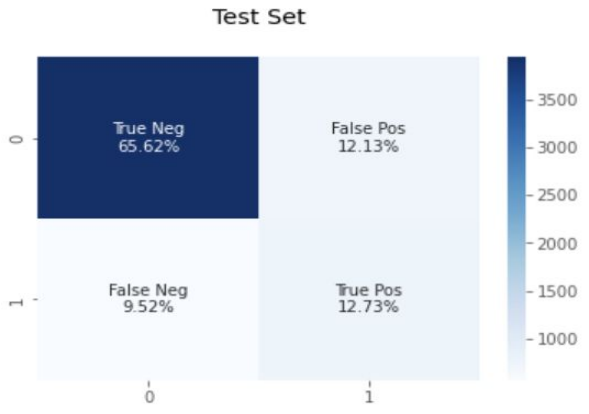
	precision	recall	f1-score	support
0	0.80	0.85	0.82	4665
1	0.31	0.24	0.27	1335
accuracy			0.71	6000
macro avg	0.55	0.54	0.55	6000
weighted avg	0.69	0.71	0.70	6000



Random Forest

ROC_AUC_Score of train set is 0.7614349850067325.
ROC_AUC_Score of test set is 0.7713328596489115.
F1 of train set is 0.648.
F1 of test set is 0.5405.

	precision	recall	f1-score	support
0	0.87	0.84	0.86	4665
1	0.51	0.57	0.54	1335
accuracy			0.78	6000
macro avg	0.69	0.71	0.70	6000
weighted avg	0.79	0.78	0.79	6000



Logistic Regression

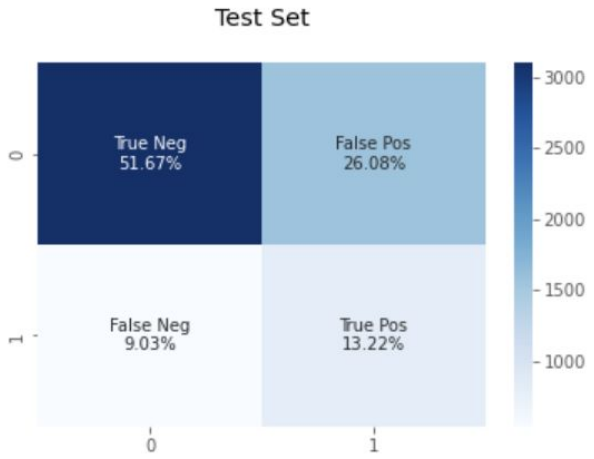
Models on Transformed Dataset

Data Transformation

```
ROC_AUC_Score of train set is 0.9547819608526235.
ROC_AUC_Score of test set is 0.6653638739357154.
F1 of train set is 0.8725.
F1 of test set is 0.4295.
              precision    recall  f1-score   support

           0       0.85      0.66      0.75      4665
           1       0.34      0.59      0.43      1335

    accuracy                           0.65      6000
   macro avg       0.59      0.63      0.59      6000
weighted avg       0.74      0.65      0.68      6000
```

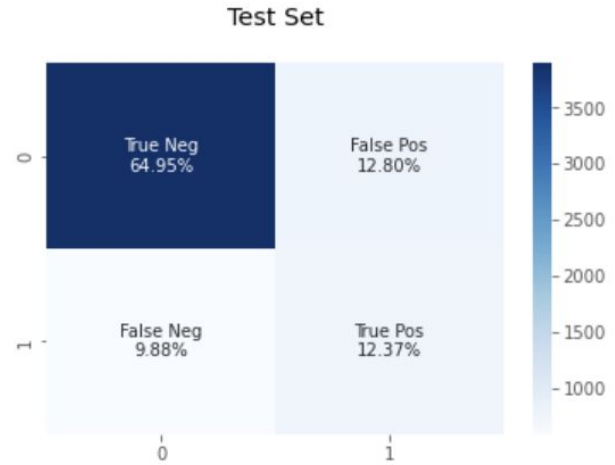


K-Nearest Neighbour

```
ROC_AUC_Score of train set is 0.7667199464076059.
ROC_AUC_Score of test set is 0.7228431502422614.
F1 of train set is 0.6461.
F1 of test set is 0.5216.
              precision    recall  f1-score   support

           0       0.87      0.84      0.85      4665
           1       0.49      0.56      0.52      1335

    accuracy                           0.77      6000
   macro avg       0.68      0.70      0.69      6000
weighted avg       0.78      0.77      0.78      6000
```



Decision Tree Classifier

Models on Transformed Dataset

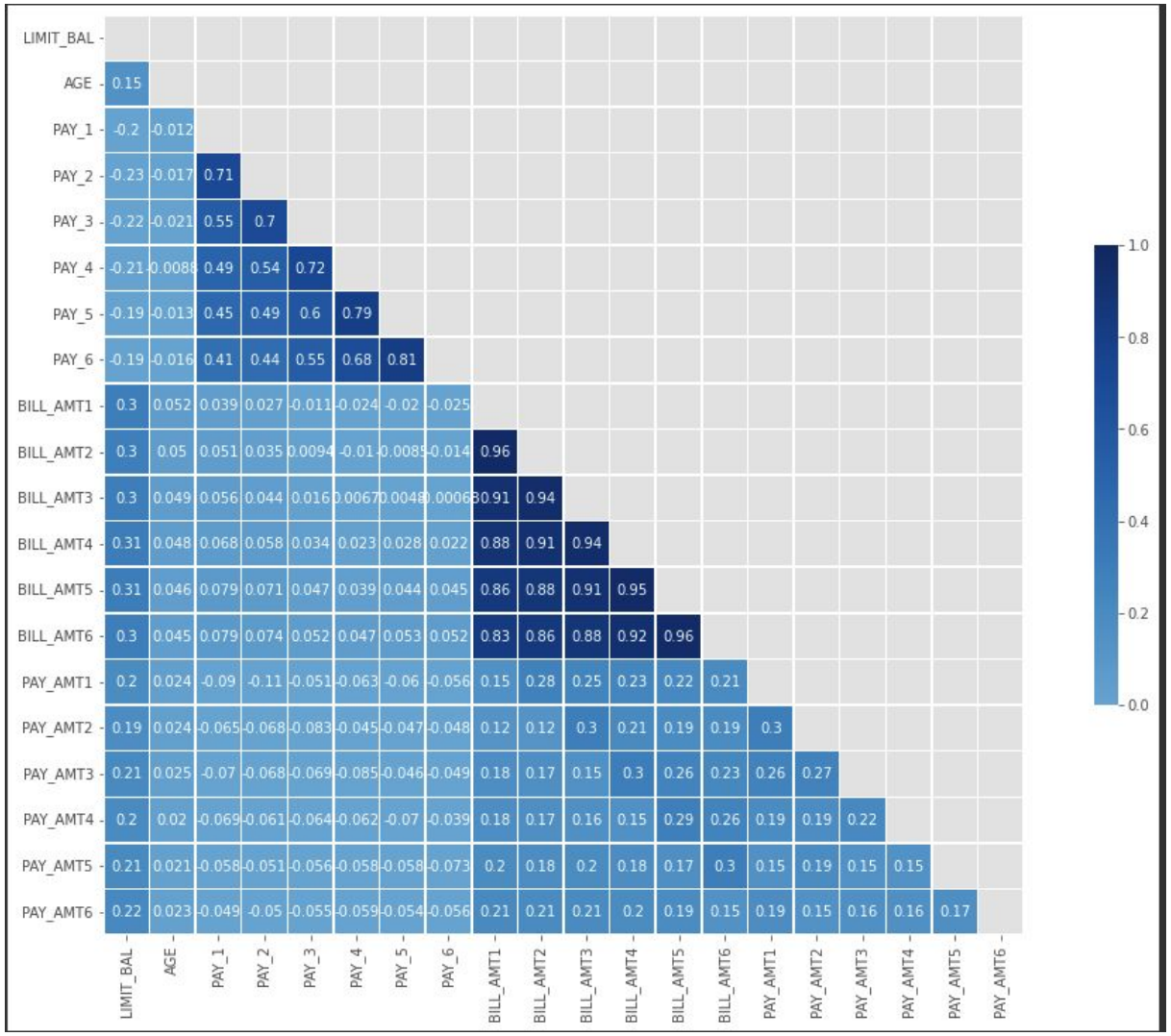
Reference

1. Length: 10-minute presentation.
2. Contents:
i. Motivation and Data summary (2-3 min)
ii. Results with different metrics (Tables) (1-2 min)
(example) (Please color the best and the worst)
iii. Explanation: why some features are important (based on your domain knowledge)? Why is a certain method better than other methods based on your data (not a general one)? What is the extracted knowledge from this data?

Feature Selection

- Technique used to reduce the number of dependent variables
- Exhaustive dataset can be reduced in size by pruning away the redundant features that reduce the model’s accuracy
- Helps reduce the computational expense by discarding the redundant features that reduce model’s accuracy
-

Feature Selection:Correlation



- Linear relation between variables

- Some features show high correlations with each other
  - BILL_AMT1 and BILL_AMT2 have 0.95
  - BILL_AMT2 and BILL_AMT3 have 0.93
  - BILL_AMT4 and BILL_AMT5 have 0.94

- Not meaningful to keep all highly correlated features, so we kept BILL_AMT1 and removed rest highly correlated features

Feature Selection: Hypothesis Test

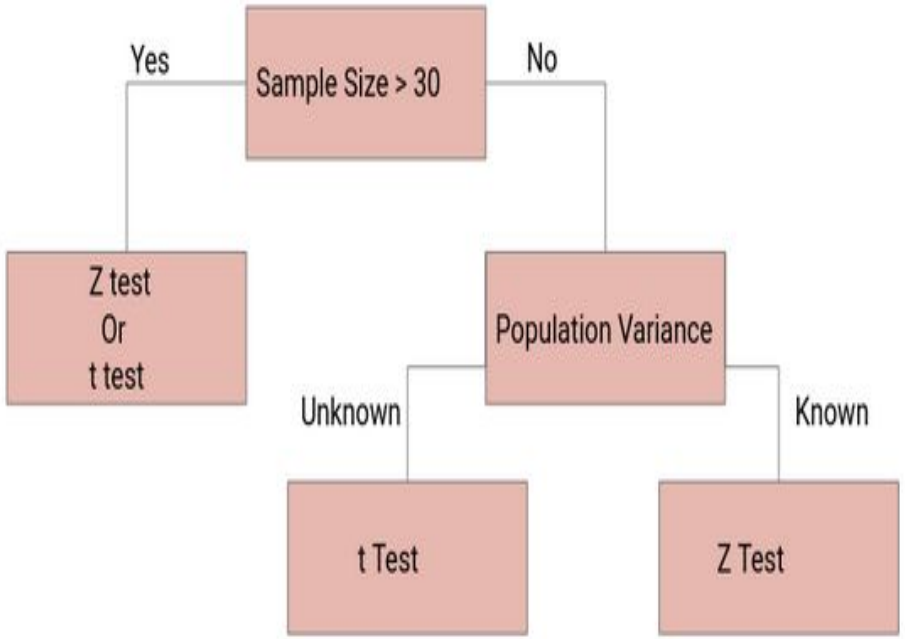
"There are two possible outcomes: if the result confirms the hypothesis, then you've made a measurement. If the result is contrary to the hypothesis, then you've made a discovery — Enrico Fermi"

Our Hypothesis : Our Null Hypothesis states that;
H0 -> No difference between features and target variable, so we should remove them
H1 -> Difference between features and target variable, so we should keep them
if p>0.01 , can't reject null hypothesis | if p<0.01, can reject null hypothesis

How?
We apply T-Test between two classes for each feature.
The features rejecting our null hypothesis
will be kept, as they are most different .

Why?

1) T-test works best when samples are greater than 30
   (Z-Test and T-Test have negligible difference).
2) It works with features that have numerical data.
3) It is a good measure of evaluation of relation in a classification problem to define yes or no.
4) Using paired t-test because comparison is done between saame iems of the sample on a unique condition.



Feature Selection:Chi-square

```
Relationship Between  default    EDUCATION
chi-square statistic:- 130.36289614432826
critical_value: 3.841458820694124
p-value: 0.0
Significance level:  0.05
Degree of Freedom:  1
Reject H0,There is a relationship between the 2 categorical variables
Reject H0,There is a relationship between the 2 categorical variables


Relationship Between  default    SEX
chi-square statistic:- 117.26473121793711
critical_value: 3.841458820694124
p-value: 0.0
Significance level:  0.05
Degree of Freedom:  1
Reject H0,There is a relationship between the 2 categorical variables
Reject H0,There is a relationship between the 2 categorical variables


Relationship Between  default    MARRIAGE
chi-square statistic:- 24.449204617239992
critical_value: 3.841458820694124
p-value: 7.629500311523429e-07
Significance level:  0.05
Degree of Freedom:  1
Reject H0,There is a relationship between the 2 categorical variables
Reject H0,There is a relationship between the 2 categorical variables
```

- Chi-square test(X^2) is used to test the independence of two events.
- It can be used to interpret correlation between two categorical variables.
- The null hypothesis assumes there is no difference between observed and expected frequencies. So if we accept the null hypothesis (e.g. p-value > 0.05) then we say the variables are not dependent on one other.

Why?

- There are categorical features in the dataset, having 2 or more categories each.
- It works best when the data is being sampled randomly.
- We have independent features in dataset
- Dataset contains categorical features that are nominal in nature.

Feature Selection:Logistic Regression

Iterations 7

Results: Logit

Model:	Logit	Pseudo R-squared:	0.166
Dependent Variable:	y	AIC:	43281.7178
Date:	2022-12-01 07:02	BIC:	43477.8934
No. Observations:	37398	Log-Likelihood:	-21618.
Df Model:	22	LL-Null:	-25922.
Df Residuals:	37375	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	7.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
LIMIT_BAL	-1.2120	0.0891	-13.5960	0.0000	-1.3867	-1.0373
SEX	-0.1623	0.0239	-6.7980	0.0000	-0.2090	-0.1155
EDUCATION	-0.1468	0.0507	-2.8941	0.0038	-0.2461	-0.0474
MARRIAGE	-0.2670	0.0490	-5.4495	0.0000	-0.3630	-0.1709
AGE	0.2912	0.0769	3.7873	0.0002	0.1405	0.4419
PAY_1	7.3809	0.1720	42.9165	0.0000	7.0438	7.7180
PAY_2	0.3836	0.1722	2.2271	0.0259	0.0460	0.7212
PAY_3	1.0749	0.1801	5.9683	0.0000	0.7219	1.4279
PAY_4	0.8766	0.2020	4.3394	0.0000	0.4807	1.2725
PAY_5	0.4853	0.2312	2.0993	0.0358	0.0322	0.9384
PAY_6	1.1587	0.2000	5.7925	0.0000	0.7666	1.5507
BILL_AMT1	-3.9374	0.6751	-5.8319	0.0000	-5.2607	-2.6141
BILL_AMT2	4.1835	0.7994	5.2336	0.0000	2.6168	5.7502
BILL_AMT3	0.8612	1.6182	0.5322	0.5946	-2.3105	4.0328
BILL_AMT4	0.6282	0.8612	0.7295	0.4657	-1.0598	2.3162
BILL_AMT5	-1.6181	0.6548	-2.4712	0.0135	-2.9015	-0.3347
BILL_AMT6	0.8187	0.4944	1.6559	0.0977	-0.1503	1.7878
PAY_AMT1	-18.0359	1.6310	-11.0581	0.0000	-21.2326	-14.8392
PAY_AMT2	-9.9243	2.1786	-4.5554	0.0000	-14.1943	-5.6544
PAY_AMT3	-3.0292	1.1296	-2.6817	0.0073	-5.2431	-0.8153
PAY_AMT4	-3.0234	0.8583	-3.5224	0.0004	-4.7057	-1.3411
PAY_AMT5	-2.7005	0.5469	-4.9378	0.0000	-3.7724	-1.6286
PAY_AMT6	-1.9629	0.4772	-4.1134	0.0000	-2.8982	-1.0276

How?

- This is the Logistic regression-based model where the amount of variation between features is interpreted by computing p-value.
- It is an iterative process where if the initial p-value for a feature exceeds the threshold for exclusion, it is removed from dataset, else we add and move on the next feature, till we meet convergence.
- The features with p-value less than 0.05 are considered to be the more relevant feature.

Why?

- Although this method is computationally expensive on a large dataset , it works well for the current dataset as its a moderately sized dataset.
- It is one of the fastest automatic feature selection methods and is capable of managing high number of variable features.

***But, keeping this in mine, it has its disadvantages too. It is based on a greedy algorithm so this might not be an optimum choice. Results will be compared in accuracy of models.

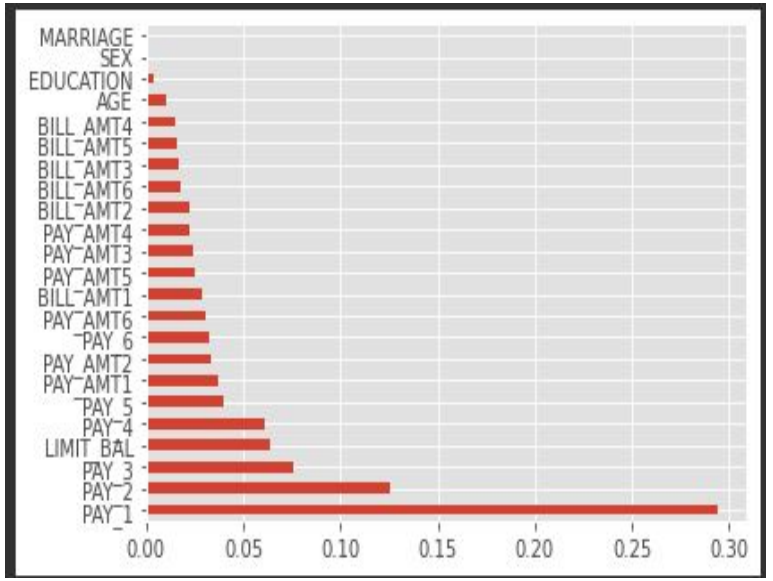
Feature Selection:Random Forest

HOW?
A measure of impurity (Gini impurity) is computed and a decrease is noted on this value after each tree. Lower impurity means that the feature is important.

In random forest, an average is calculated for each accuracy decrease for tree is calculated classification, we used impurity gini impurity to determine the the importance of a feature variable.

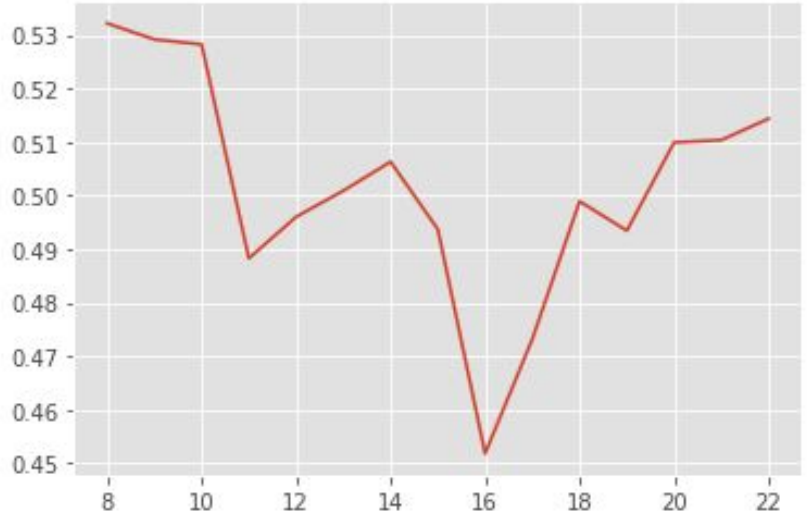
Features on the top of the list are chosen as important features.

A recall chart can be referred to choose the top n features to be selected.



WHY?

1. This method provides relative values of the computed importances. The advantage in this method is a speed of computation, all the needed values are computed during the Random Forest training.



Feature Selection Methods/Results

Feature Selection Method	Selected Features	Removed Features
T-Test	Limit_BAL, AGE, PAY_1, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, BILL_AMT1, PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6	BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6
Correlation	PAY_1, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, BILL_AMT1, PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6	BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6
CHI-Square	SEX, EDUCATION, MARRIAGE	None
Logistic Regression	Limit_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY_1, PAY_3, PAY_4, PAY_5, PAY_6, BILL_AMT1, BILL_AMT6, , PAY_AMT1, , PAY_AMT2	BILL_AMT3, BILL_AMT4, BILL_AMT6
Random Forest Feature Importance	PAY_1, PAY_2, PAY_3, PAY_4, PAY_5, PAY_AMT1, PAY_6, Limit_BAL, BILL_AMT1, PAY_AMT3	PAY_6, PAY_AMT6, BILL_AMT1, PAY_AMT3, PAY_AMT5, PAY_AMT4, BILL_AMT2, BILL_AMT6, BILL_AMT3, BILL_AMT5, BILL_AMT4, AGE, EDUCATION, SEX, MARRIAGE

Model Comparison (Accuracy)

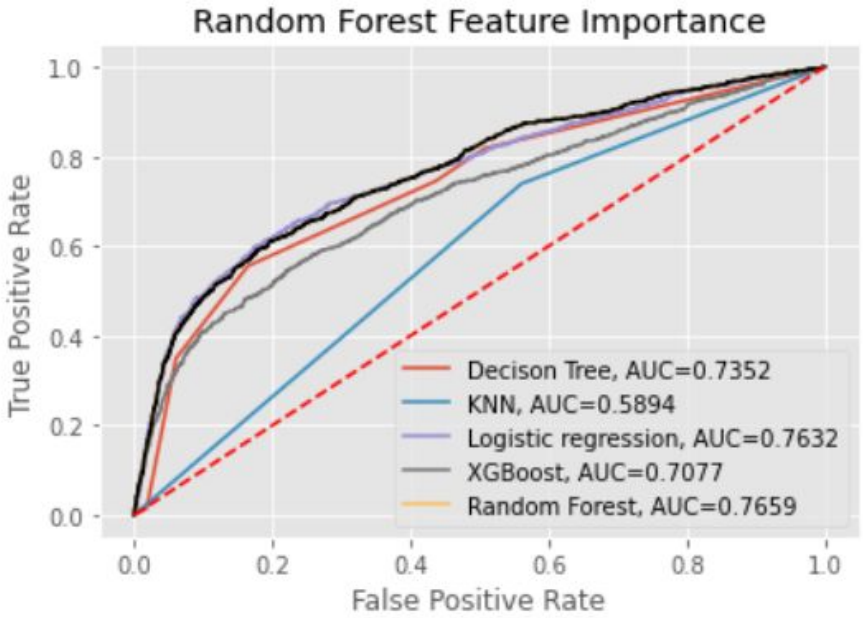
Method	Logistic Regression	KNN	Decision Tree	Random Forest	XGBoost
Without Feature Selection	78%	69%	77%	71%	65%
T-Test, Chi-square	78%	67%	77%	67%	65%
Correlation, Chi-square	77%	70%	78%		78%
Feature Selection by Feature Importance (Logistic Regression)	78%, 55 %	70%, 65%	77%,	79%, 77%	66%, 76 %
Feature Selection by Feature Importance (Random Forest)	78%	51%	77%	76%	63%

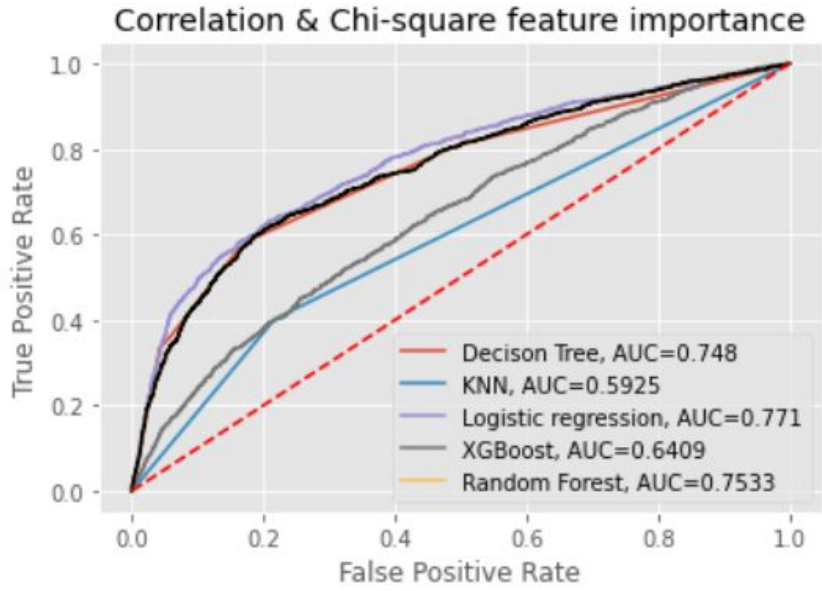
Model Comparison (F1)

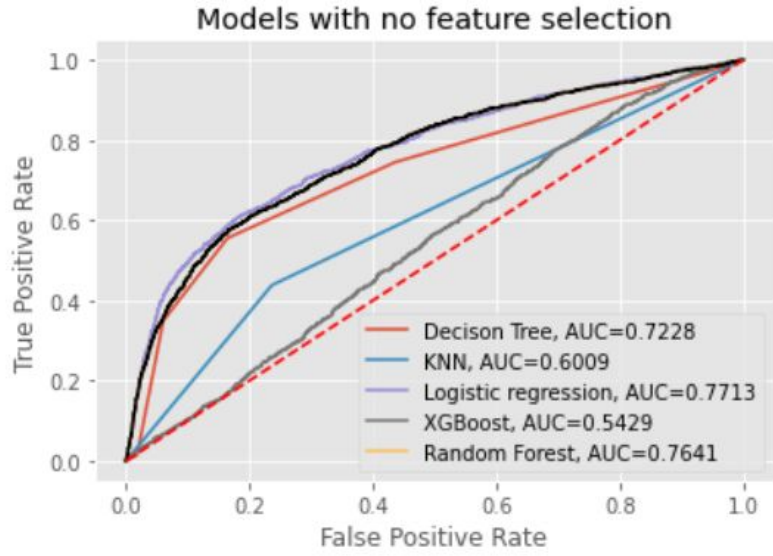
Method	Logistic Regression	KNN	Decision Tree	Random Forest	XGBoost
Without Feature Selection					0.45
T-Test, Chi-Square	0.54	0.37	0.52	0.28	0.45
Correlation, Chi-Square	0.54	0.37	0.52		0.44
Feature Selection by Feature Importance (Logistic Regression)	0.54	0.39	0.52	0.49	0.45, 0.
Feature Selection by Feature Importance (Random Forest)	0.53	0.39	0.52	0.52	0.46

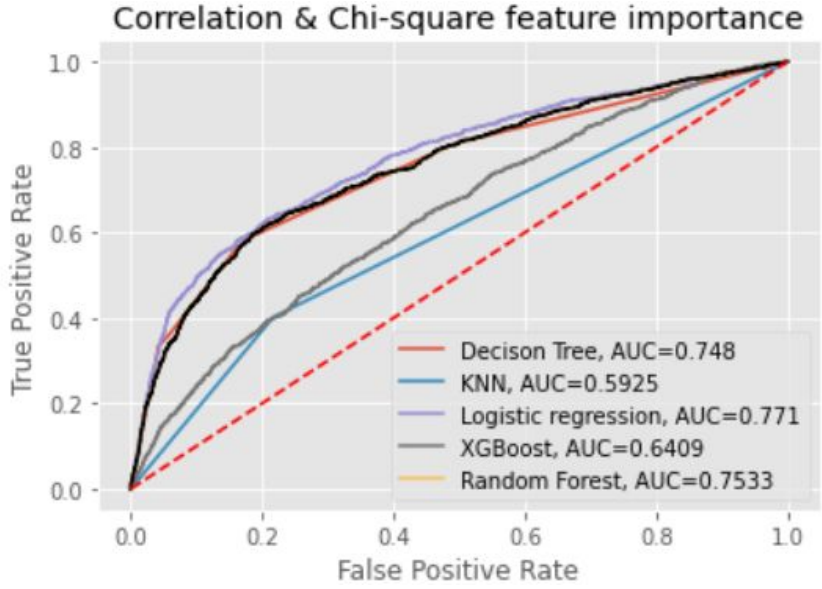
Model Comparison (AUC score)

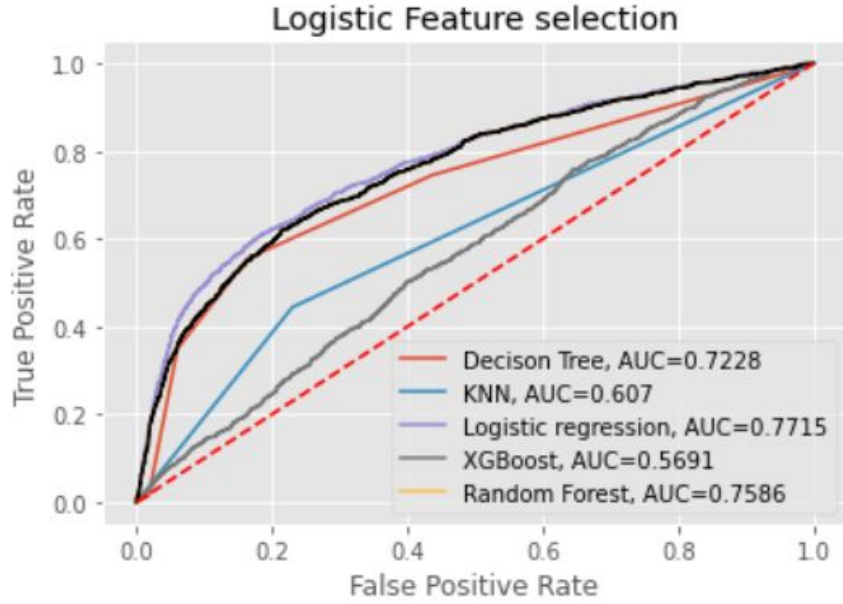
Method	Logistic Regression	KNN	Decision Tree	Random Forest	XGBoost
Without Feature Selection	0.77	0.60	0.72	0.59	0.69
T-Test, Chi-Square	0.77	0.58	0.72	0.55	0.69
Correlation, Chi-Square	0.77	0.59	0.74	0.75	0.73
Feature Selection by Feature Importance (Logistic Regression)	0.77	0.60	0.72	0.75	0.69
Feature Selection by Feature Importance (Random Forest)	0.76	0.58	0.73	0.76	0.73

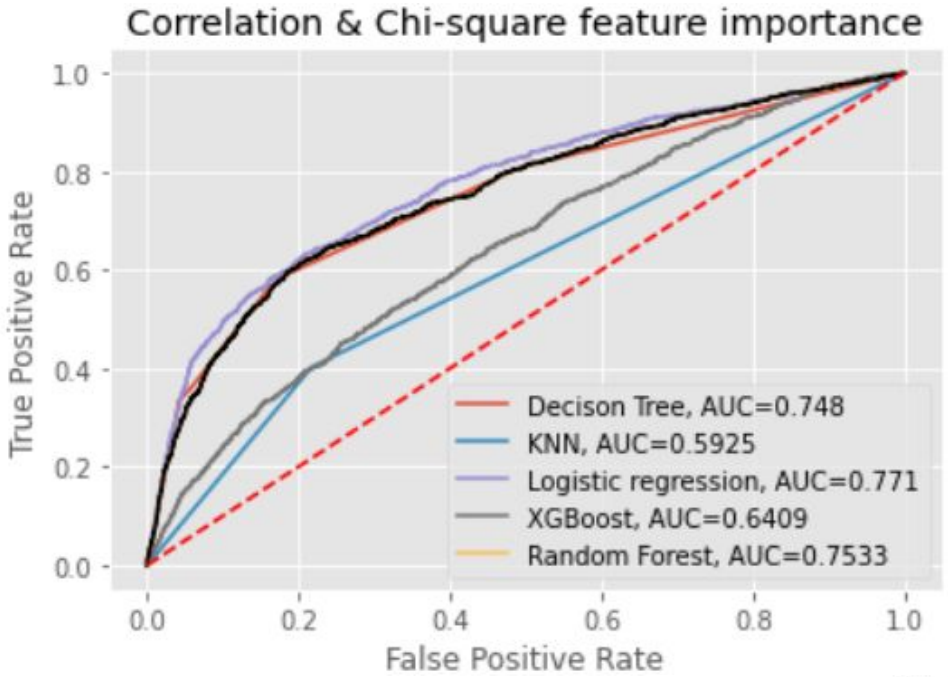

Random Forest Feature Importance
True Positive Rate
False Positive Rate
Decison Tree, AUC=0.7352
KNN, AUC=0.5894
Logistic regression, AUC=0.7632
XGBoost, AUC=0.7077
Random Forest, AUC=0.7659

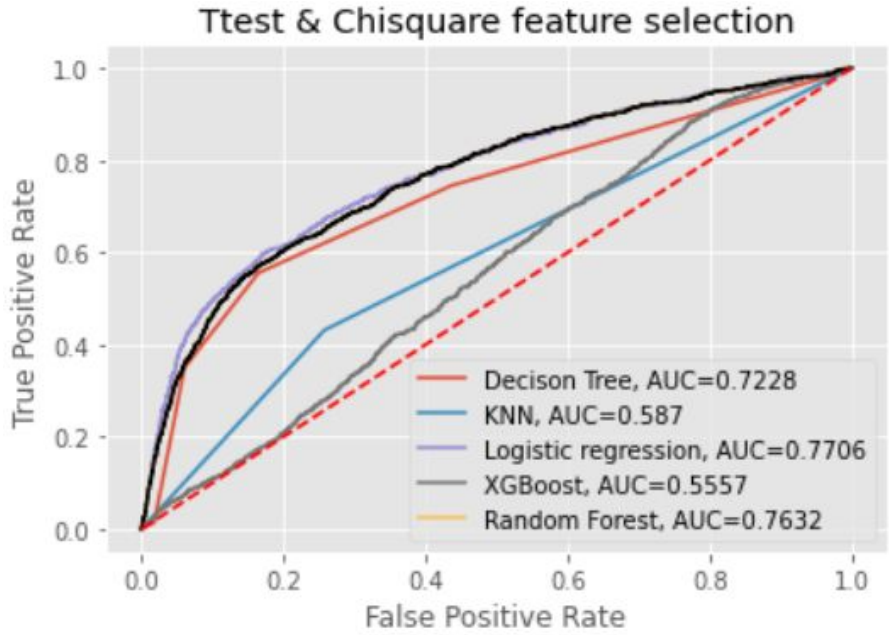

Correlation & Chi-square feature importance
True Positive Rate
False Positive Rate
Decison Tree, AUC=0.748
KNN, AUC=0.5925
Logistic regression, AUC=0.771
XGBoost, AUC=0.6409
Random Forest, AUC=0.7533

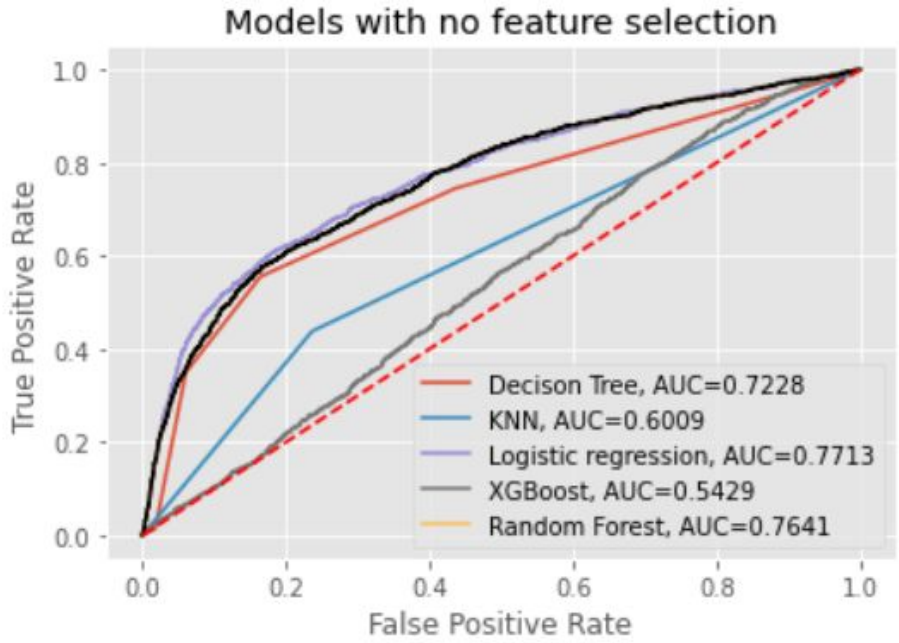

Models with no feature selection
True Positive Rate
False Positive Rate
Decison Tree, AUC=0.7228
KNN, AUC=0.6009
Logistic regression, AUC=0.7713
XGBoost, AUC=0.5429
Random Forest, AUC=0.7641


Correlation & Chi-square feature importance
True Positive Rate
False Positive Rate
Decison Tree, AUC=0.748
KNN, AUC=0.5925
Logistic regression, AUC=0.771
XGBoost, AUC=0.6409
Random Forest, AUC=0.7533


Logistic Feature selection
True Positive Rate
False Positive Rate
Decison Tree, AUC=0.7228
KNN, AUC=0.607
Logistic regression, AUC=0.7715
XGBoost, AUC=0.5691
Random Forest, AUC=0.7586


Correlation & Chi-square feature importance
True Positive Rate
False Positive Rate
Decison Tree, AUC=0.748
KNN, AUC=0.5925
Logistic regression, AUC=0.771
XGBoost, AUC=0.6409
Random Forest, AUC=0.7533


Ttest & Chisquare feature selection
True Positive Rate
False Positive Rate
Decison Tree, AUC=0.7228
KNN, AUC=0.587
Logistic regression, AUC=0.7706
XGBoost, AUC=0.5557
Random Forest, AUC=0.7632


Models with no feature selection
True Positive Rate
False Positive Rate
Decison Tree, AUC=0.7228
KNN, AUC=0.6009
Logistic regression, AUC=0.7713
XGBoost, AUC=0.5429
Random Forest, AUC=0.7641

Hypothesis test ->

Correlation Method ->

Chi-square -

Random Forest ->

Logistic Regression ->

Pros and Cons Feature Selection Methods

Method	Pros	Cons
T-Test, Chi-Square		
Correlation, Chi-Square		
Feature Selection by Feature Importance (Logistic Regression)		
Feature Selection by Feature Importance (Random Forest)	Speed of computation, all the needed values are computed during the Random Forest training.	**They are completely useless if your model is weak** **Strongly Influenced by correlation features** **They are biased towards numerical and high cardinality features**

Conclusion

Final Feature selected method-

Final selected features -

Final best model