Capstone Project - 3

Credit Card Default Prediction

Project By :
Sanjay Yadav

Contents

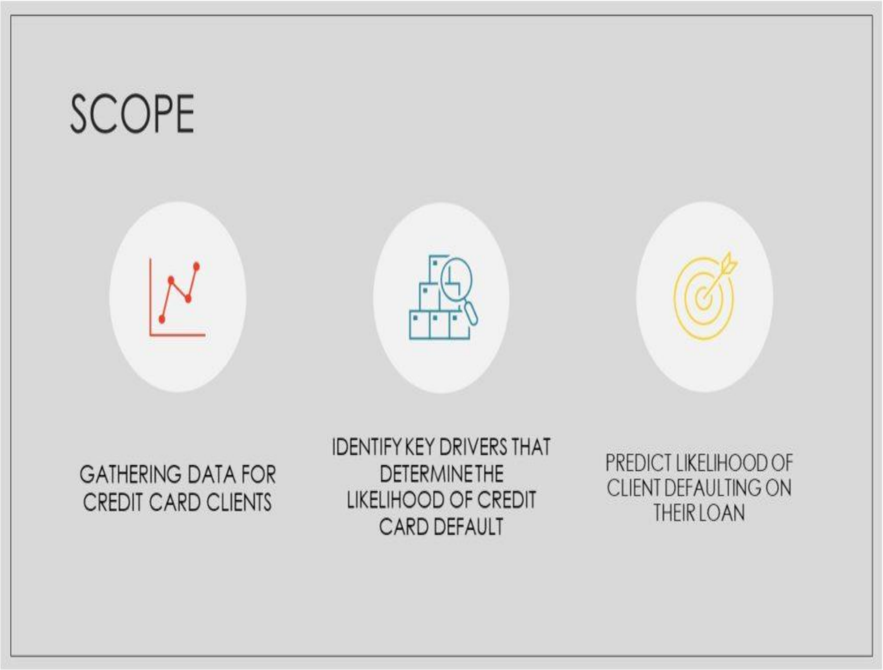
Introduction

Credit risk plays a major role in the banking industry business. Banks main activities involve granting loan, credit card, investment, mortgage, and others. Credit card has been one of the most booming financial services by banks over the past years.

However, with the growing number of credit card users, banks have been facing an escalating credit card default rate. Data analytics can provide solutions to tackle the current phenomenon and management credit risks.

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.



Data Summary

Features

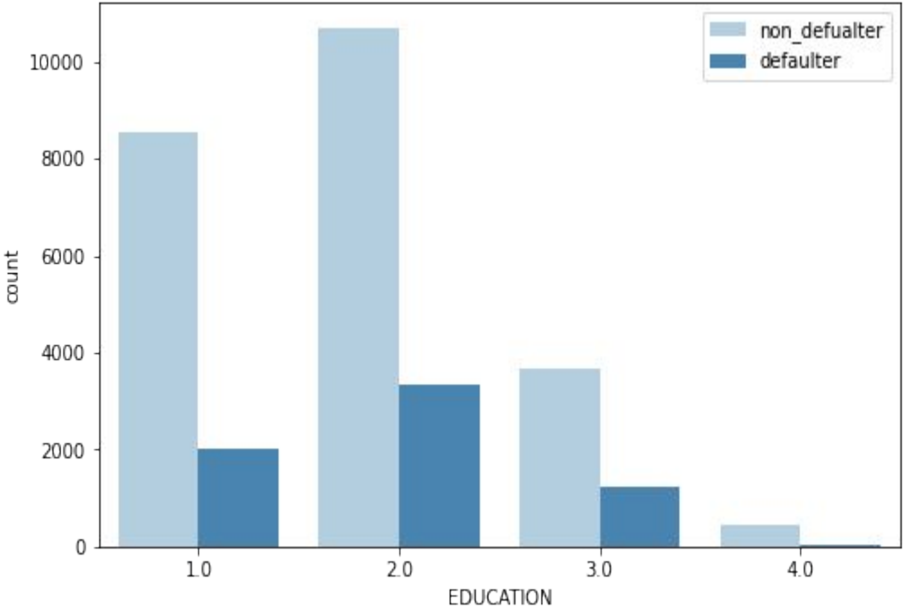
- Credit Info: Credit line
- Demographics: Gender, Highest education degree, Age, and Marital Status
- Payment History (Apr ~ Sep 2005): repayment status, payment amount, and bill amount by month

Target

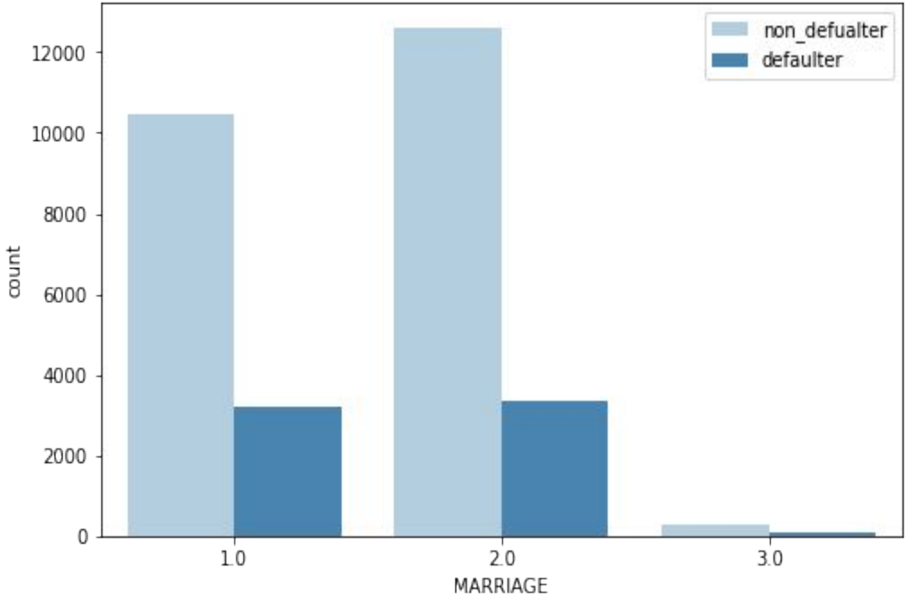
Whether the credit card client will default or not next month

Name	Description
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 (-2=no consumption, -1=pay duly, 0=the use of revolving credit, 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)

Exploratory Data Analysis

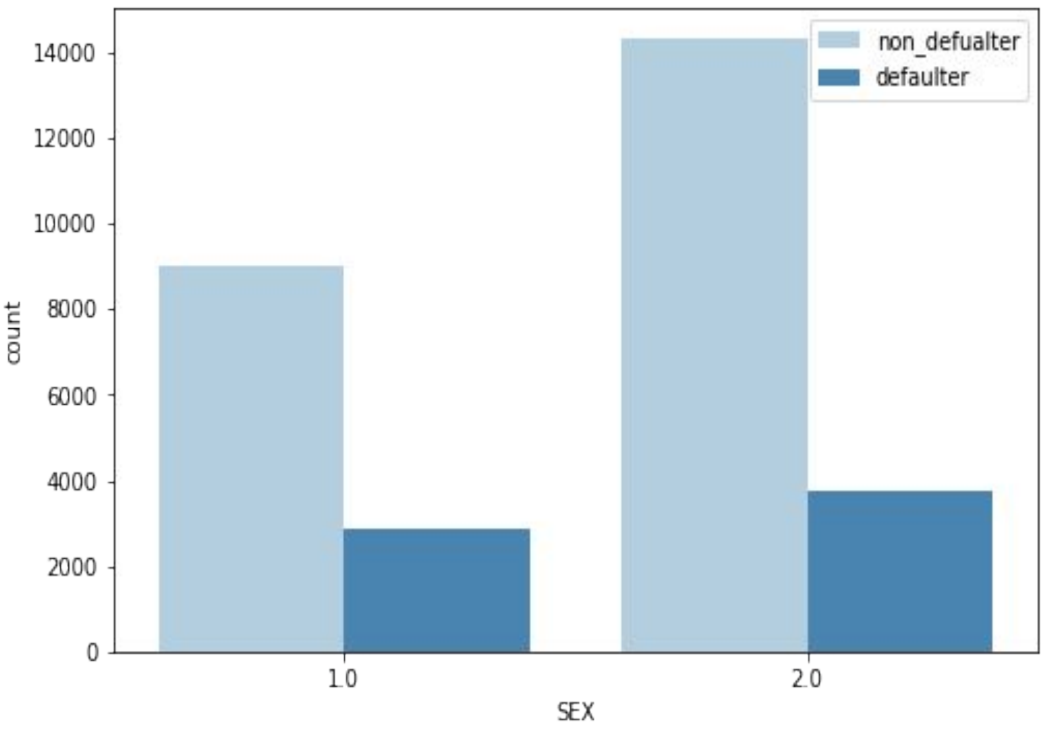



```
 1 = graduate school
2 = university
3 = high school
4 = others
```

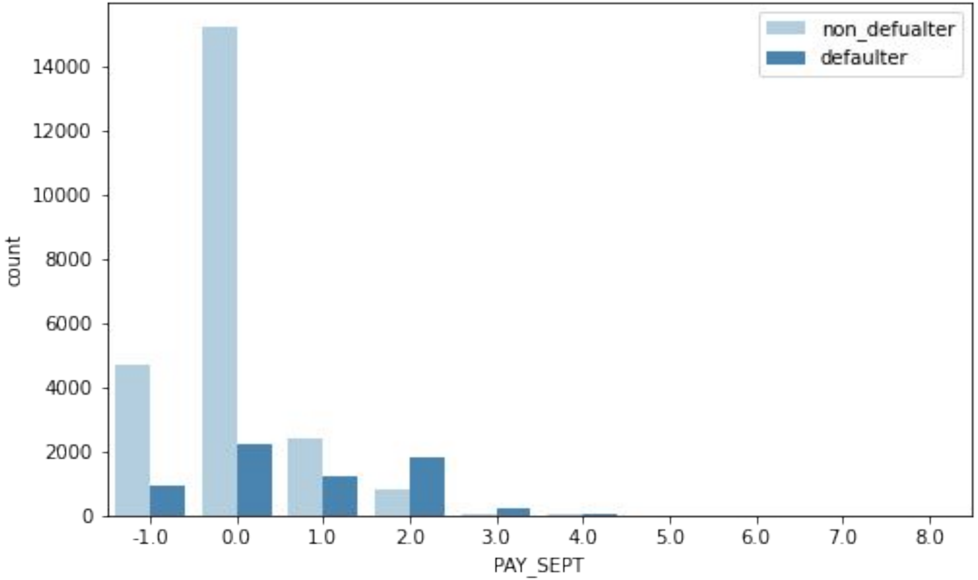


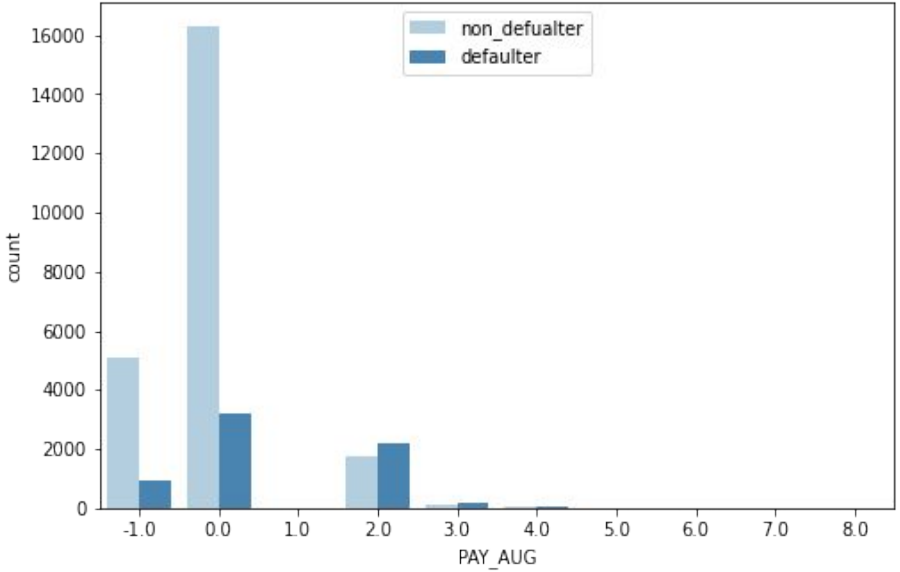

```
 1 = married
2 = single
3 = others
```

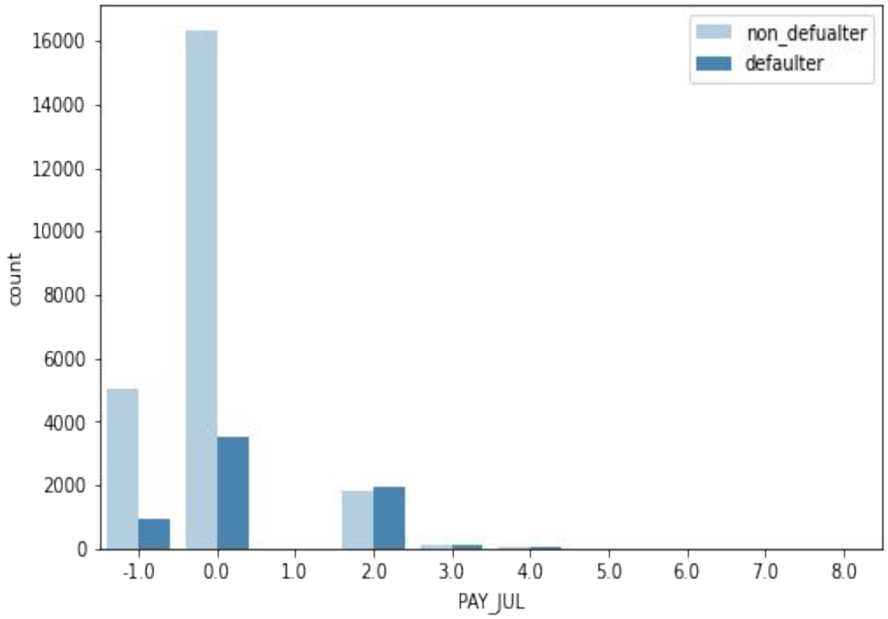
1 = male
2 = female

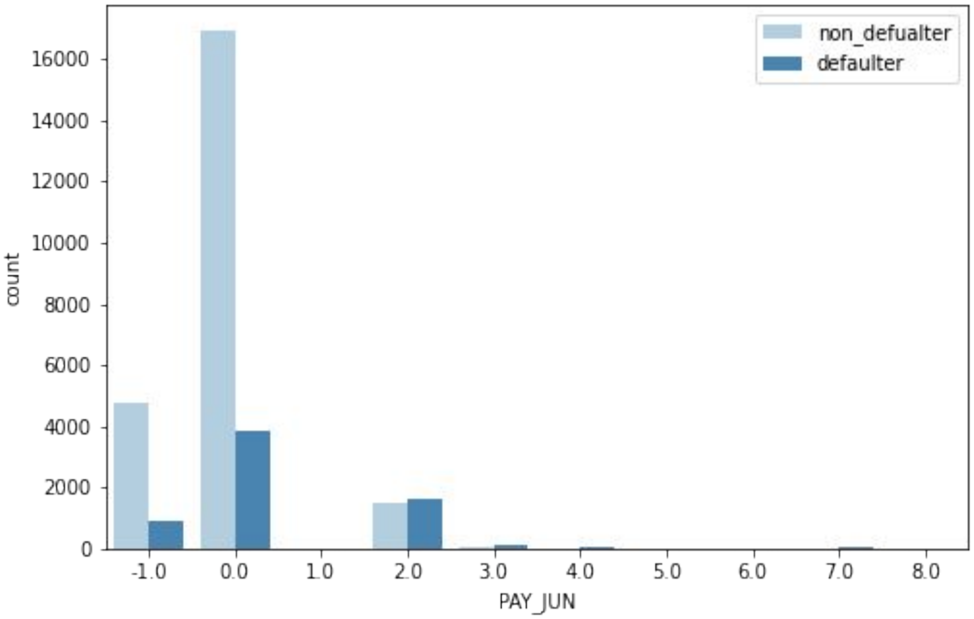


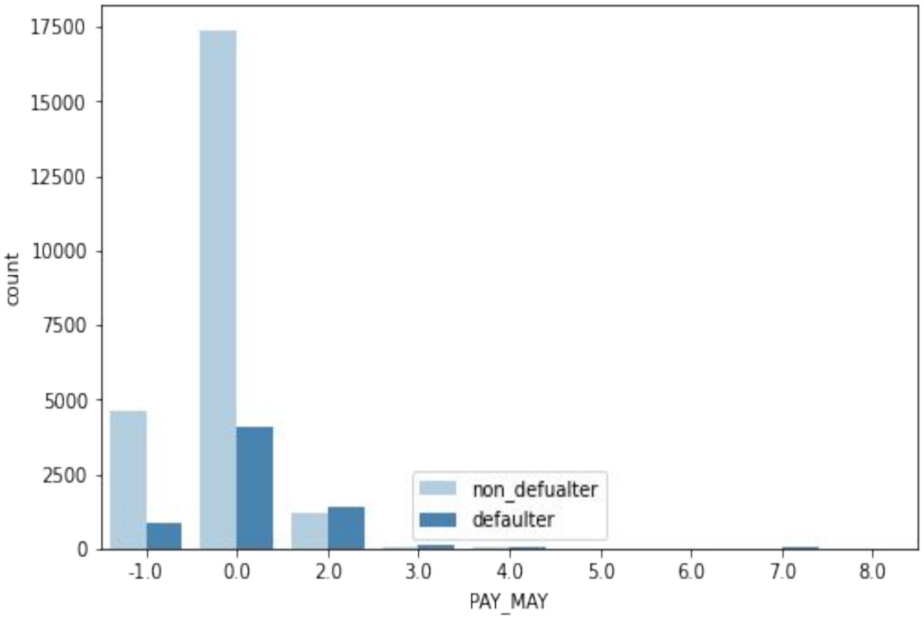
```
 -2= no consumption
-1= pay duly
1 = payment delay for one month
2 = payment delay for two months
...
8 = payment delay for eight months
9 = payment delay for nine months and above
```

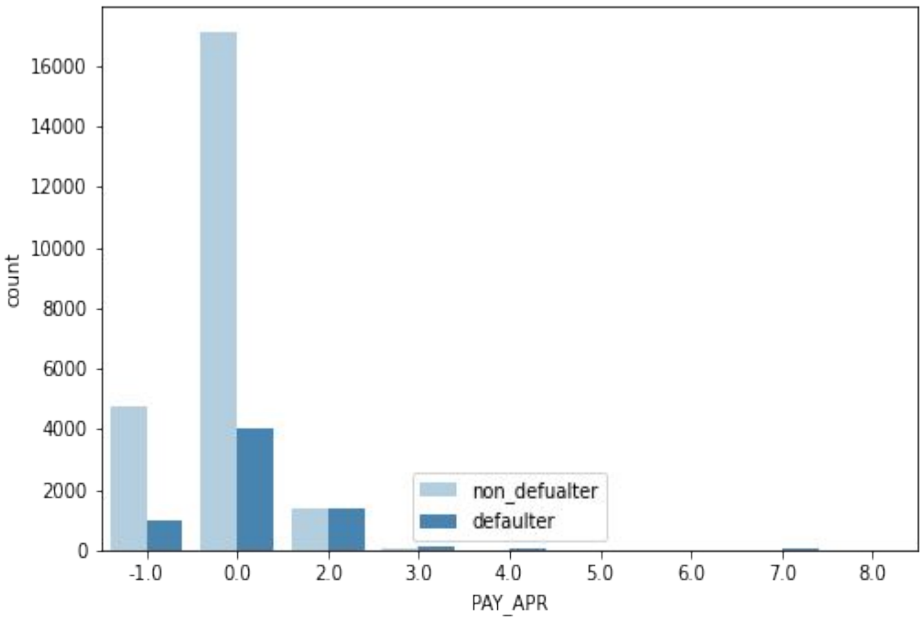


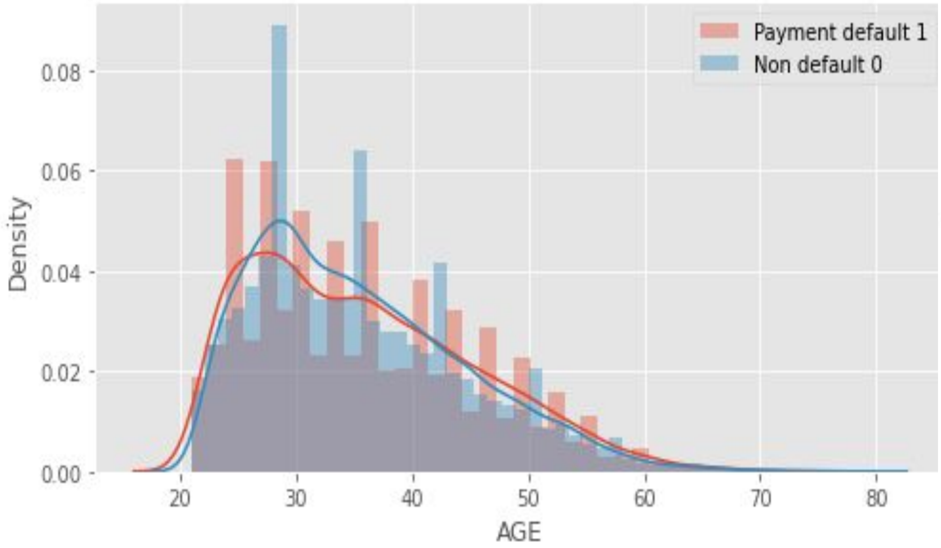


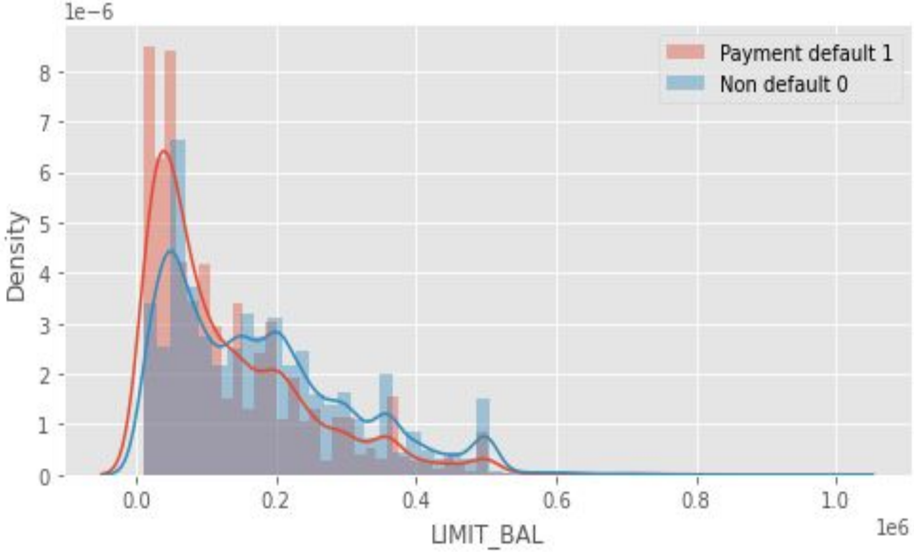

non_defualter
defaulter
count
16000
14000
12000
10000
8000
6000
4000
2000
0
-1.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
PAY_JUL


non_defualter
defaulter
count
16000
14000
12000
10000
8000
6000
4000
2000
0
-1.0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
PAY_JUN


17500
15000
12500
10000
7500
5000
2500
0
count
non_defualter
defaulter
-1.0
0.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
PAY_MAY

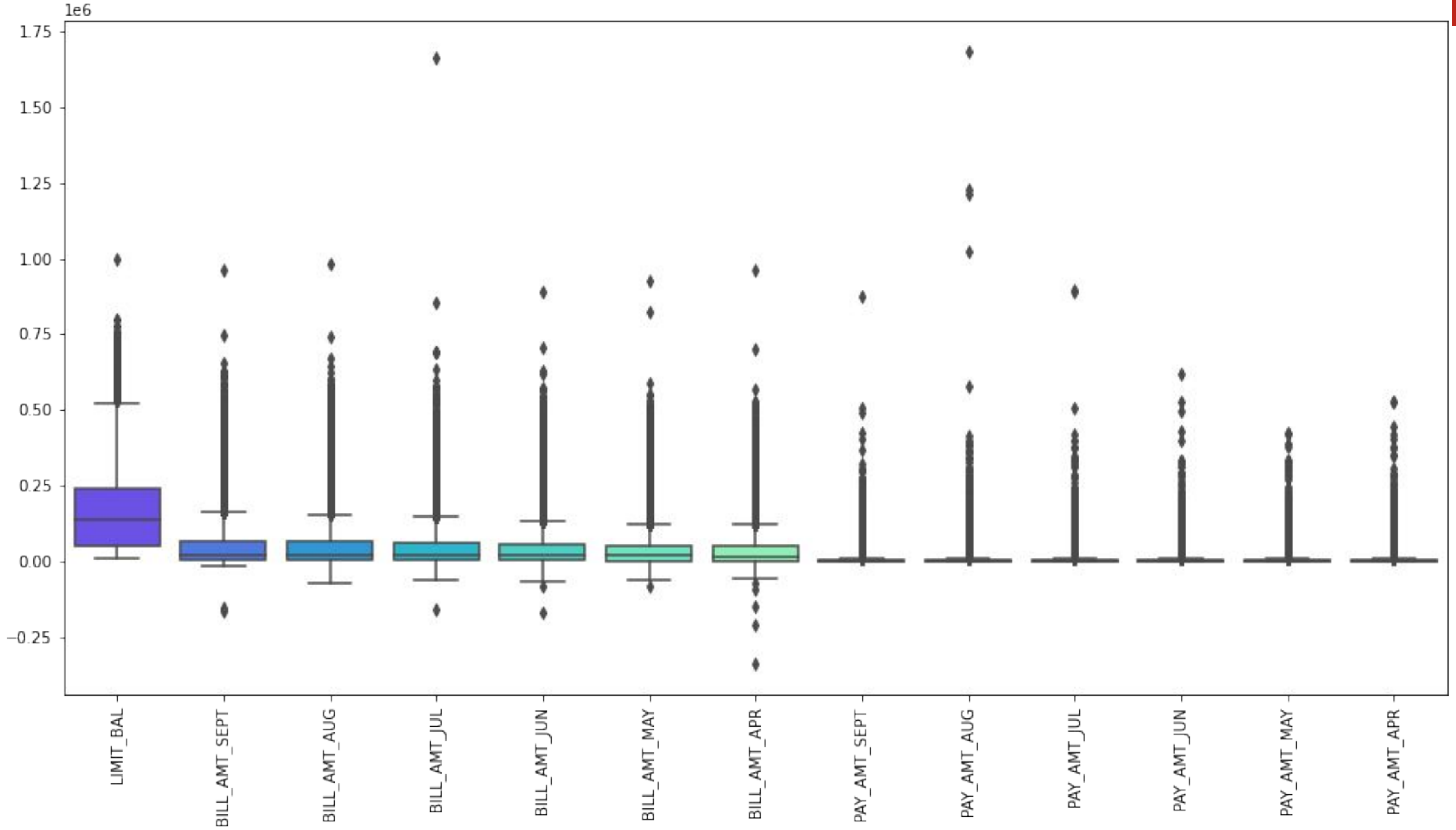

16000
14000
12000
10000
8000
6000
4000
2000
0
count
non_defualter
defaulter
-1.0
0.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
PAY_APR


Payment default 1
Non default 0
Density
AGE
0.00
0.02
0.04
0.06
0.08
20
30
40
50
60
70
80

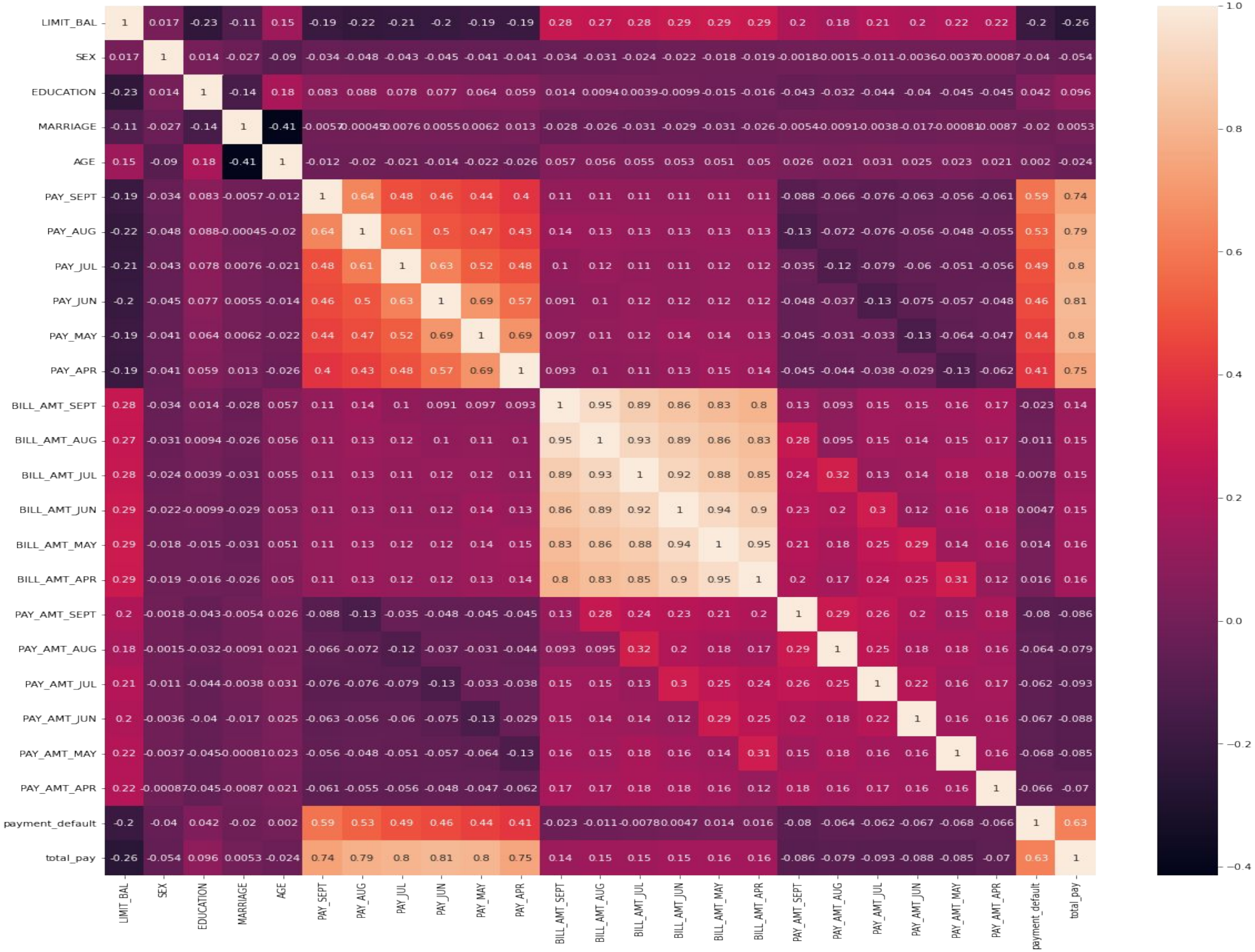

1e−6
Payment default 1
Non default 0
Density
LIMIT_BAL
1e6
0
1
2
3
4
5
6
7
8
0.0
0.2
0.4
0.6
0.8
1.0

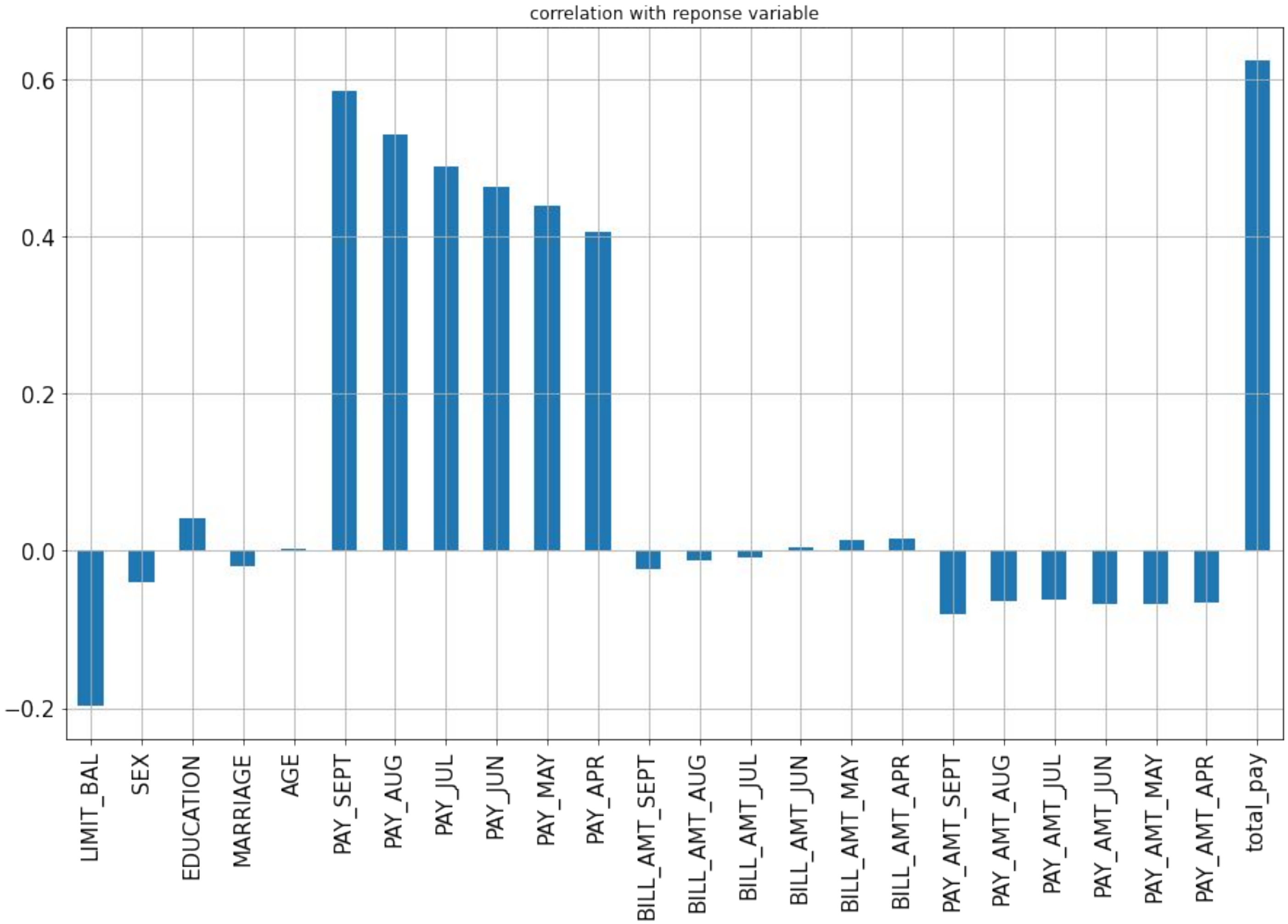
Inferences from EDA

- As the value 0 for default payment means 'not default' and value 1 means 'default', the mean of 0.221 means that there are 22.1% of credit card contracts that will default next month.There is huge difference between non-defaulter(0) and defaulter(1).
- Number of Male credit holder(represented as 1) is less than Female(represented as 2).
- More number of credit holders are university students(represented as 2) followed by Graduates(represented as 1) and then High school students(represented as 3).
- More number of credit cards holder are Single.
- Mostly , payments are not due(0) from april to september.
- The average value for the amount of credit card limit is 167,484 NT dollars. The standard deviation is 129,747 NT dollars, ranging from 10,000 to 1M NT dollars.
- Average age is 35.5 years, with a standard deviation of 9.2 years. Age above 60 years old rarely uses the credit card.

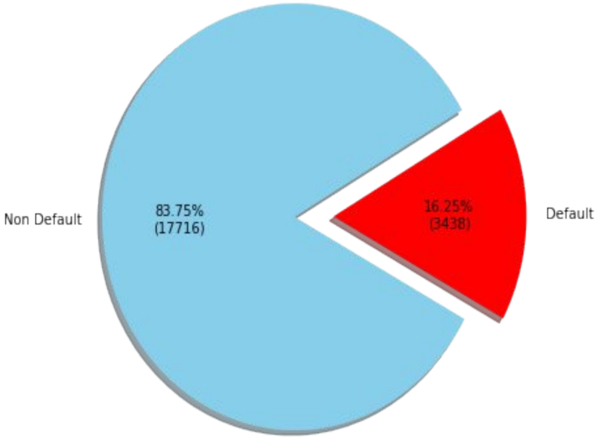


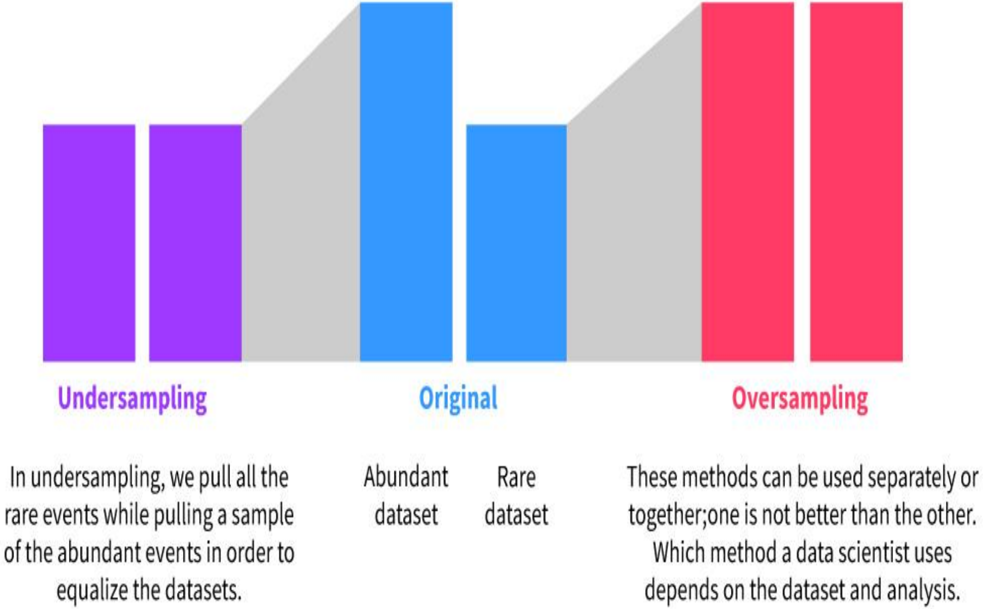
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
LIMIT_BAL
BILL_AMT_SEPT
BILL_AMT_AUG
BILL_AMT_JUL
BILL_AMT_JUN
BILL_AMT_MAY
BILL_AMT_APR
PAY_AMT_SEPT
PAY_AMT_AUG
PAY_AMT_JUL
PAY_AMT_JUN
PAY_AMT_MAY
PAY_AMT_APR


LIMIT_BAL
SEX
EDUCATION
MARRIAGE
AGE
PAY_SEPT
PAY_AUG
PAY_JUL
PAY_JUN
PAY_MAY
PAY_APR
BILL_AMT_SEPT
BILL_AMT_AUG
BILL_AMT_JUL
BILL_AMT_JUN
BILL_AMT_MAY
BILL_AMT_APR
PAY_AMT_SEPT
PAY_AMT_AUG
PAY_AMT_JUL
PAY_AMT_JUN
PAY_AMT_MAY
PAY_AMT_APR
payment_default
total_pay
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


correlation with reponse variable
0.6
0.4
0.2
0.0
−0.2
LIMIT_BAL
SEX
EDUCATION
MARRIAGE
AGE
PAY_SEPT
PAY_AUG
PAY_JUL
PAY_JUN
PAY_MAY
PAY_APR
BILL_AMT_SEPT
BILL_AMT_AUG
BILL_AMT_JUL
BILL_AMT_JUN
BILL_AMT_MAY
BILL_AMT_APR
PAY_AMT_SEPT
PAY_AMT_AUG
PAY_AMT_JUL
PAY_AMT_JUN
PAY_AMT_MAY
PAY_AMT_APR
total_pay

Handling Class Imbalance





Model Evaluation Parameters

A 'Confusion Matrix' is a consolidation of the number of times a model gives a correct or an incorrect inference or simply, the number of times a model rightly identifies the truth (actual classes) and the number of times it gets confused in identifying one class from another.

CONFUSUSION MATRIX	ACTUAL	
PREDICTED	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

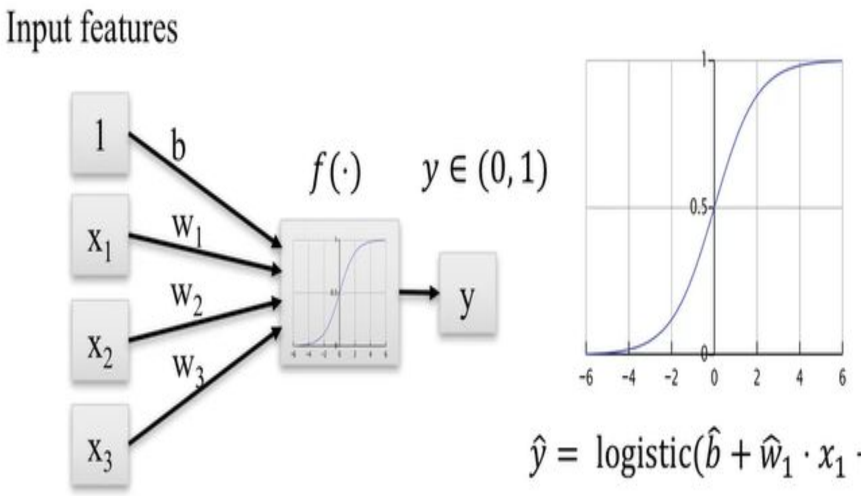
$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{F1-Score} = \frac{2*Precision*Recall}{Precision+Recall}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

Logistic Regression

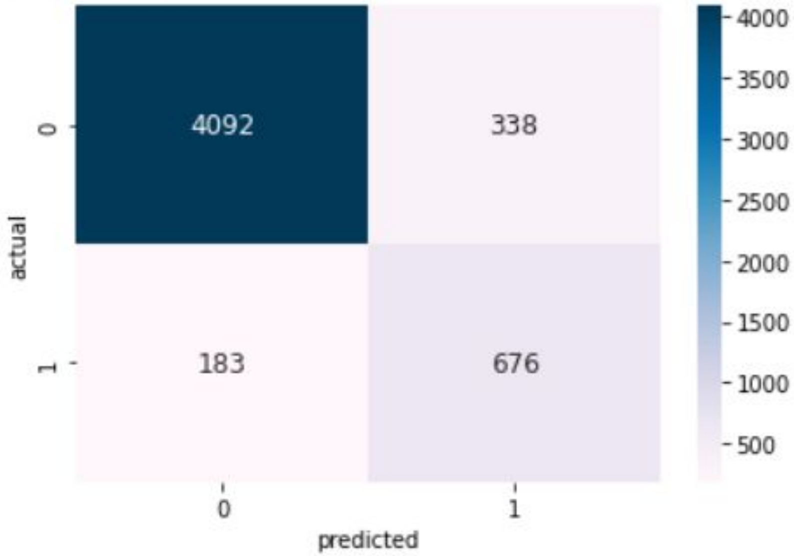


$$\hat{y} = \text{logistic}(\hat{b} + \hat{w}_1 \cdot x_1 + \cdots \hat{w}_n \cdot x_n)$$

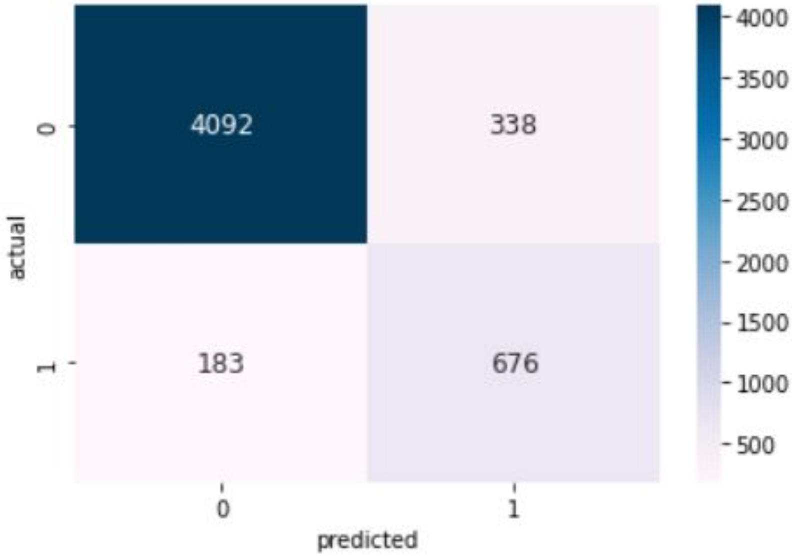
$$= \frac{1}{1 + \exp\left[-(\hat{b} + \hat{w}_1 \cdot x_1 + \cdots \hat{w}_n \cdot x_n)\right]}$$

- Poor performance on non-linear data
- Not very powerful
- Requires moderate or no multicollinearity between independent variables

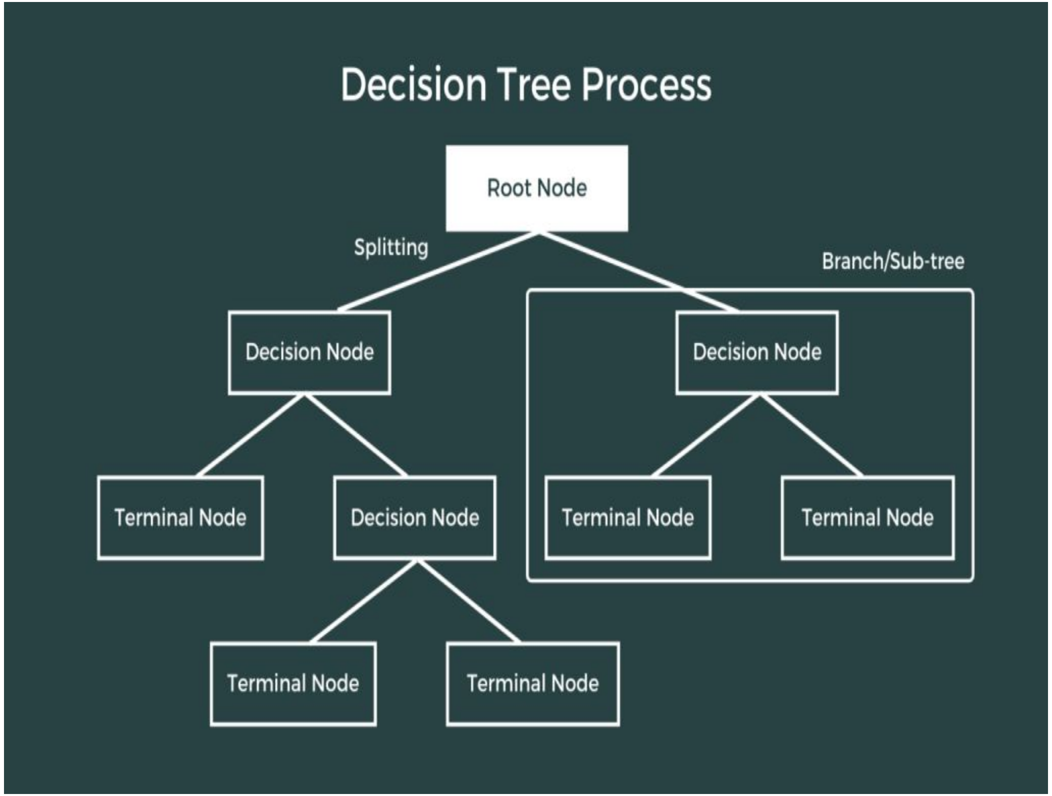
Model : Logistic Regression
Recall : 0.787
Accuracy : 0.901
Precision : 0.667
AUC_ROC : 0.855
F1: 0.722



Model : Logistic Regression
Recall : 0.787
Accuracy : 0.901
Precision : 0.667
AUC_ROC : 0.855
F1: 0.722



Decision Tree



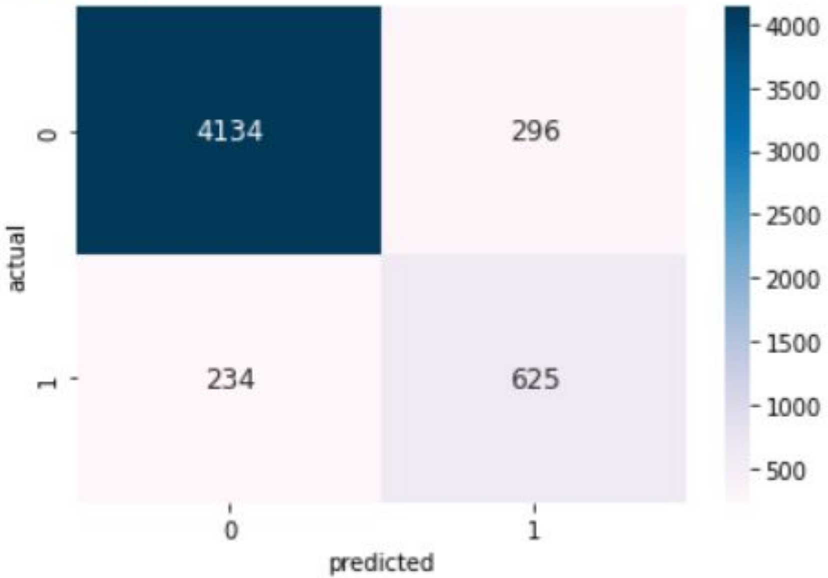
Pros:

- Normalization or scaling of data not needed.
- Handling missing values
- Easy to explain
- Automatic Feature selection

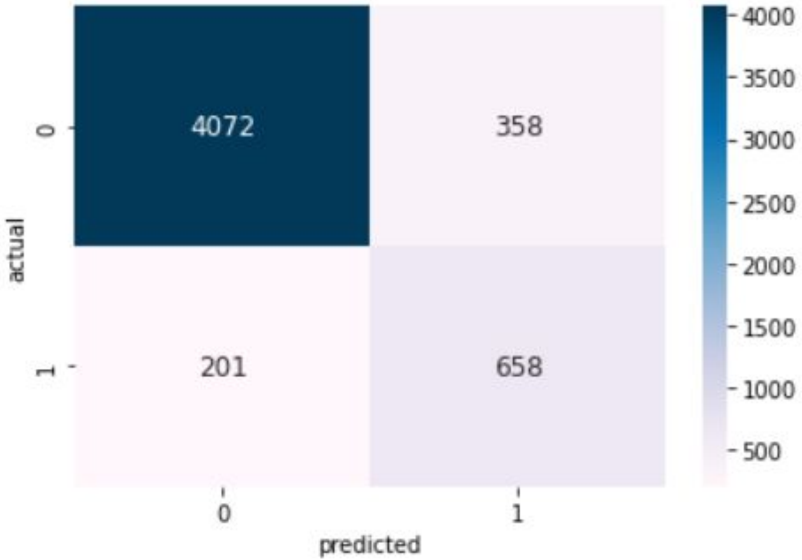
Cons:

- Prone to overfitting.
- Sensitive to data

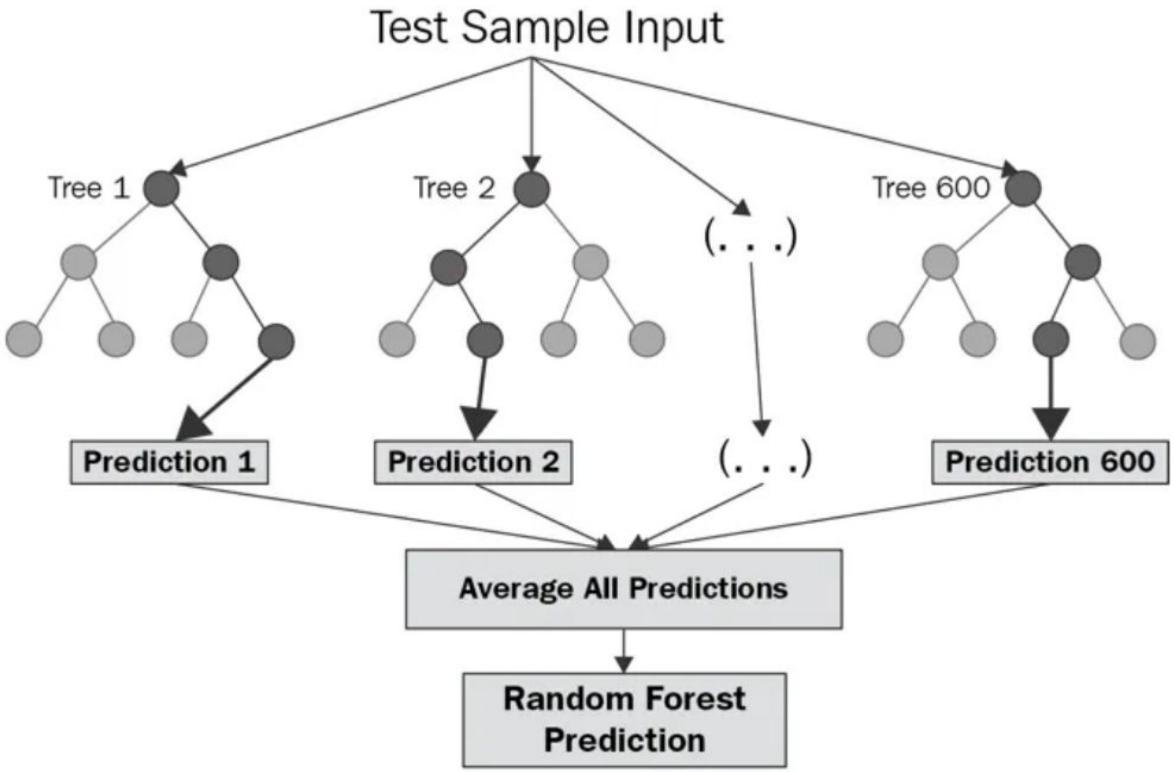
```
Model : Decision Trees
Recall :  0.728
Accuracy :  0.9
Precision :  0.679
AUC_ROC :  0.83
F1:  0.702
```




```
Model : Decision Trees
Recall :  0.766
Accuracy :  0.894
Precision :  0.648
AUC_ROC :  0.843
F1:  0.702
```



Random Forest



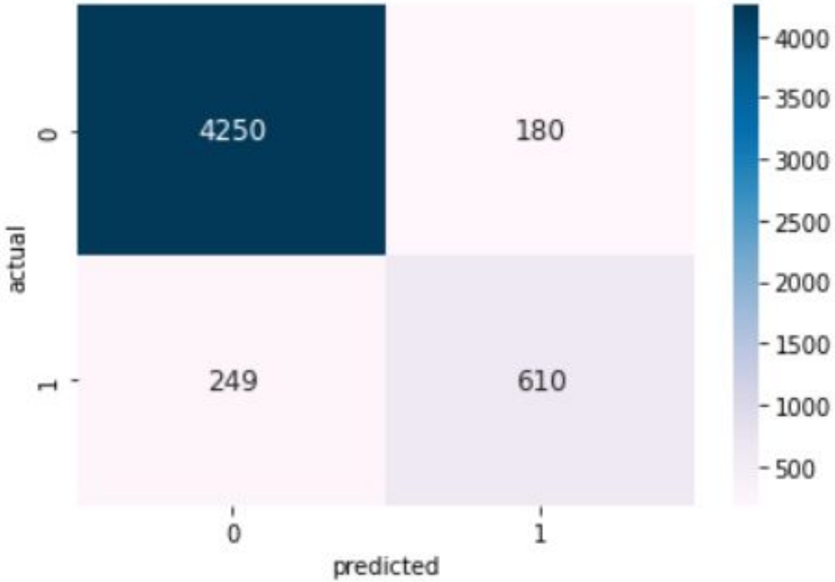
Pros:

- Good Performance on Imbalanced datasets.
- Handling of huge amount of data
- No problem of overfitting

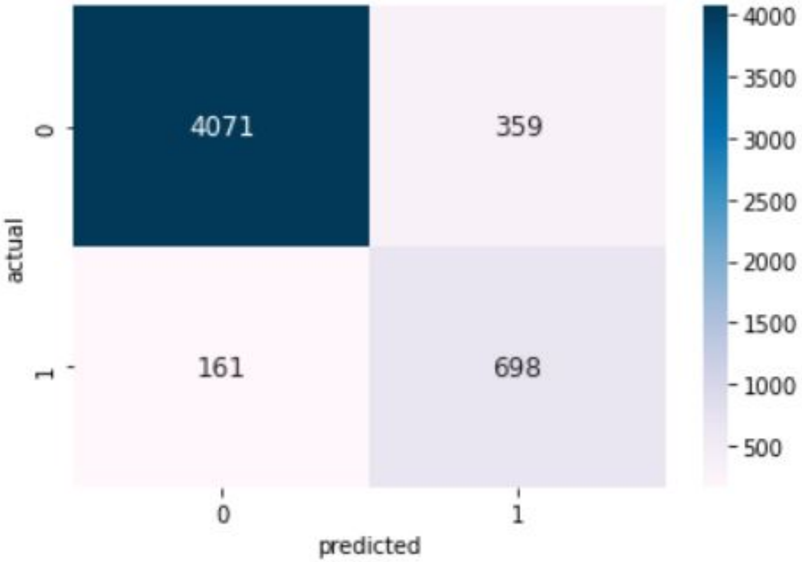
Cons:

- Features need to have some predictive power
- Appears as Black Box

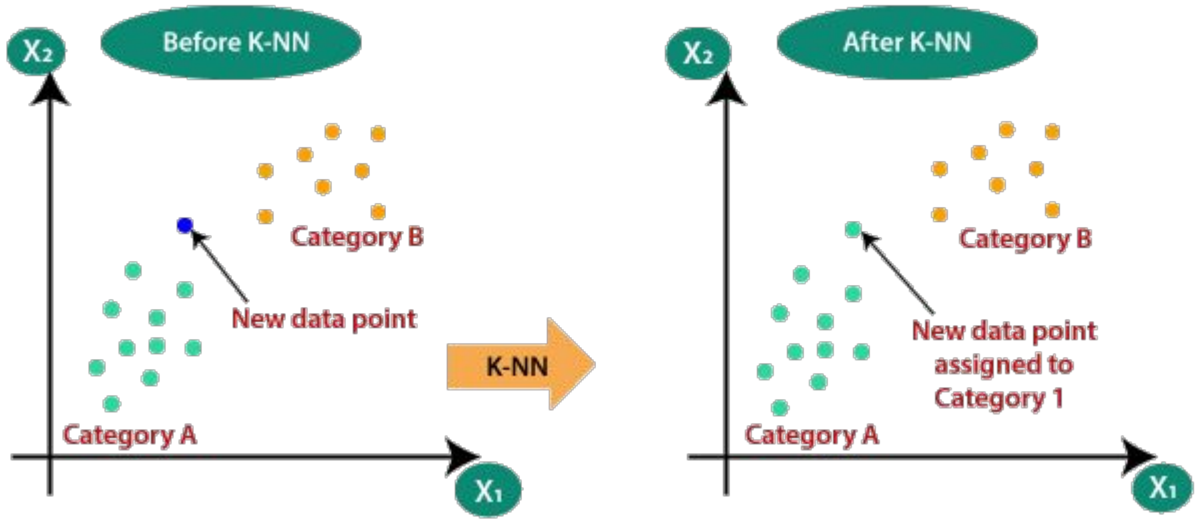
```
Model : Random Forest
Recall :  0.71
Accuracy :  0.919
Precision :  0.772
AUC_ROC :  0.835
F1:  0.74
```




```
Model : Random Forest
Recall :  0.813
Accuracy :  0.902
Precision :  0.66
AUC_ROC :  0.866
F1:  0.729
```



K-Nearest Neighbor



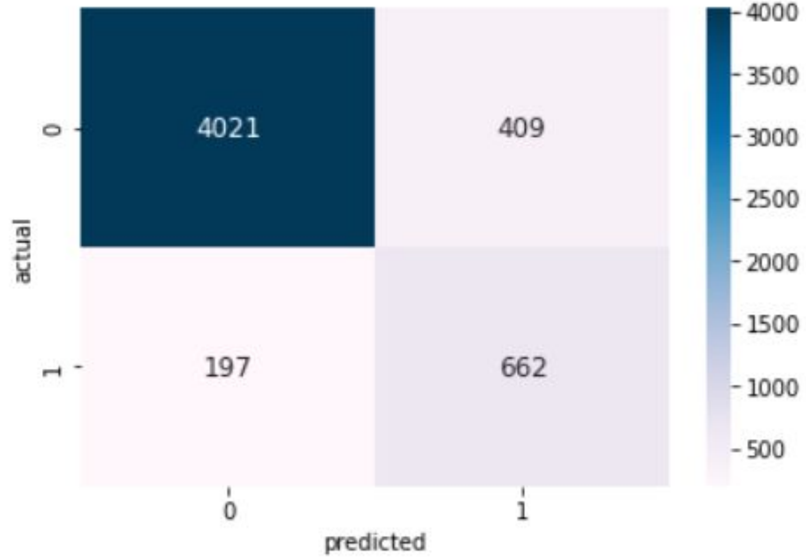
Pros:

- Simple to understand and implement.
- No assumption about data
- Constantly evolving

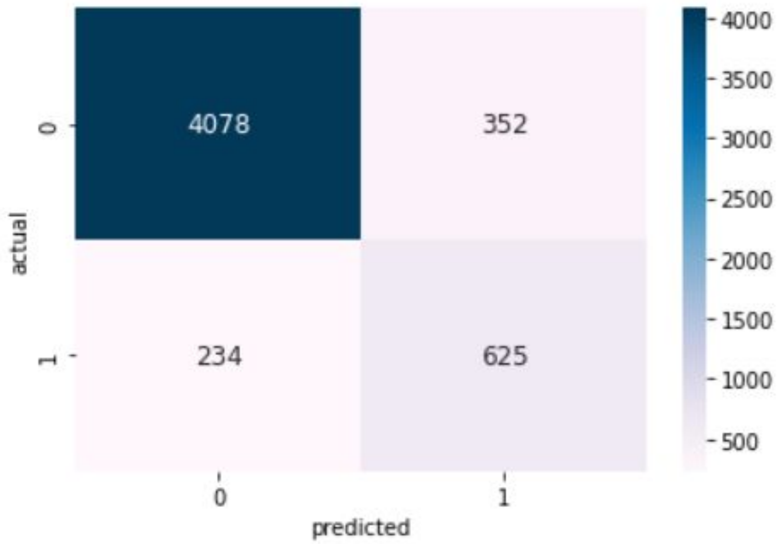
Cons:

- Slow for large datasets.
- Scaling of data absolute must.
- Curse of dimensionality
- Does not work well on Imbalanced data.

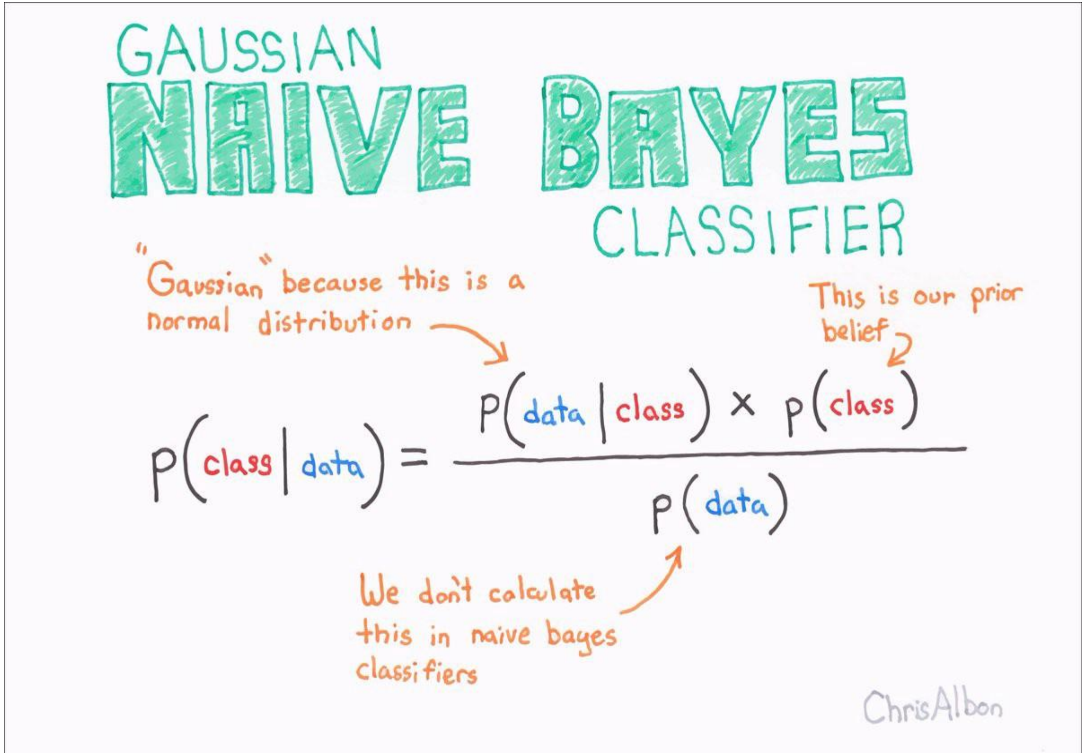
```
Model : K-Nearest Neighbors
Recall :  0.771
Accuracy :  0.885
Precision :  0.618
AUC_ROC :  0.839
F1:  0.686
```




```
Model : K-Nearest Neighbors
Recall :  0.728
Accuracy :  0.889
Precision :  0.64
AUC_ROC :  0.824
F1:  0.681
```



Naive Bayes



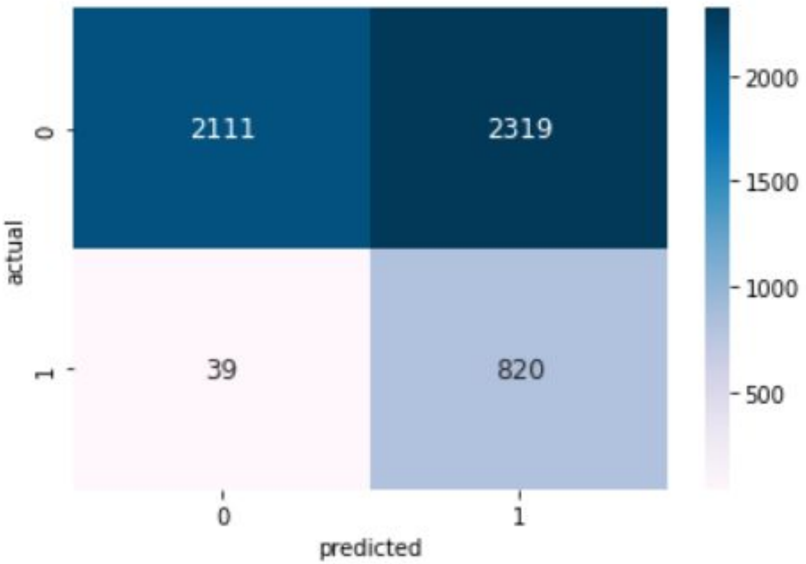
Pros:

- It is very fast and can be used in real time.
- Insensitive to irrelevant features.
- Good performance with high dimensional data.

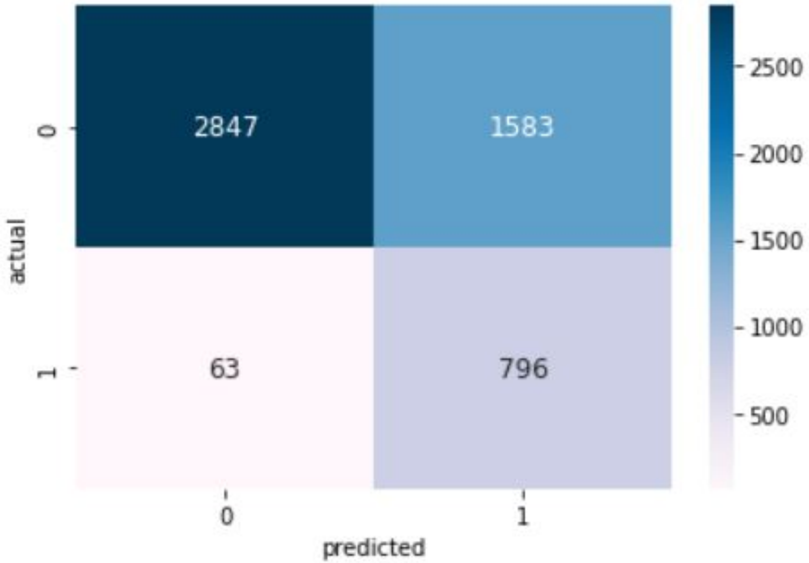
Cons:

- Independence of features does not hold.
- Bad estimator.

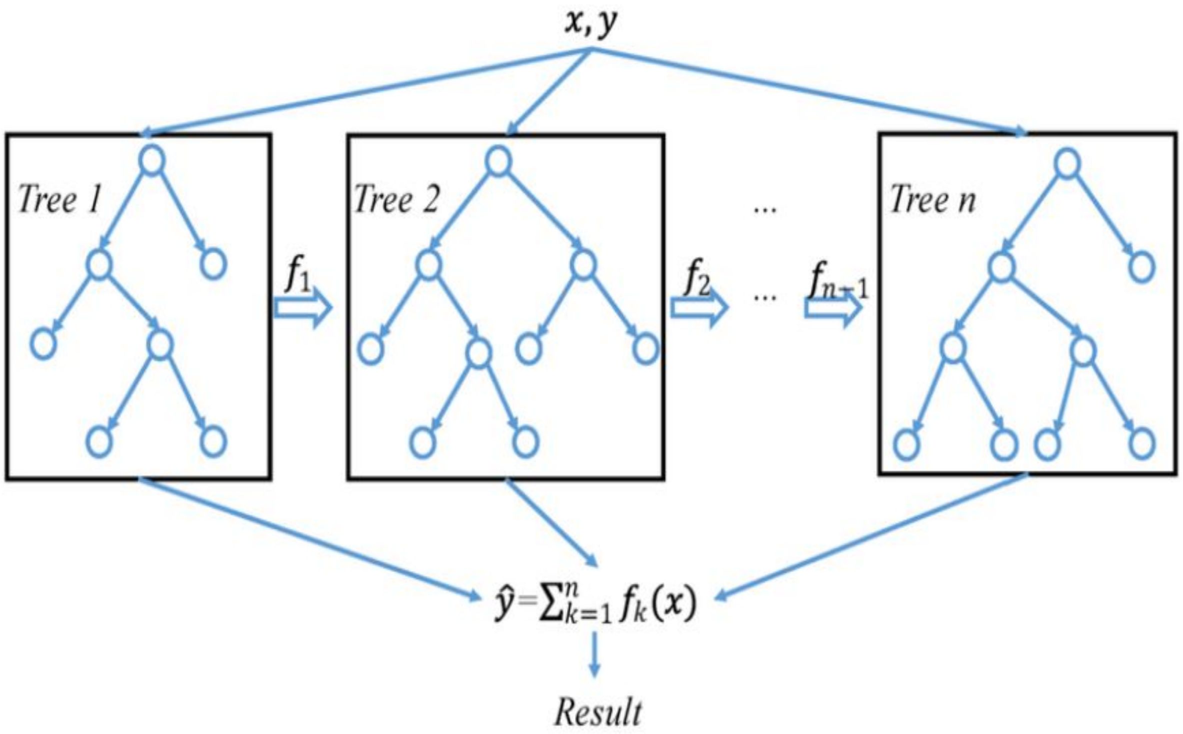
```
Model : Naive Bayes
Recall :  0.955
Accuracy :  0.554
Precision :  0.261
AUC_ROC :  0.716
F1:  0.41
```




```
Model : Naive Bayes
Recall :  0.927
Accuracy :  0.689
Precision :  0.335
AUC_ROC :  0.785
F1:  0.492
```



XGBoost



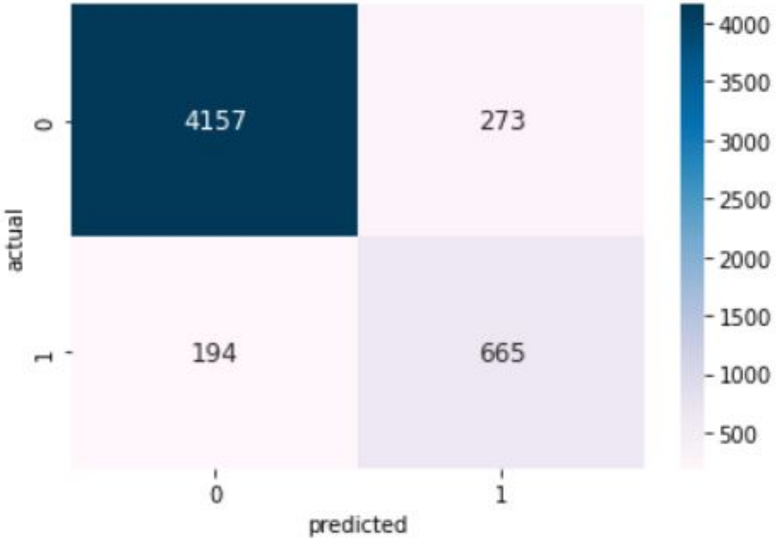
Pros:

- Less feature engineering required
- Handles large sized datasets
- Good model performance
- Less prone to overfitting

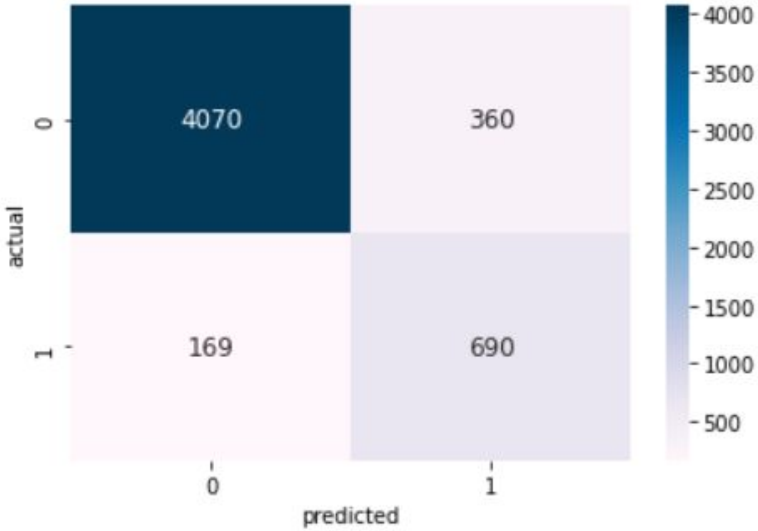
Cons:

- Difficult interpretation
- Harder to tune

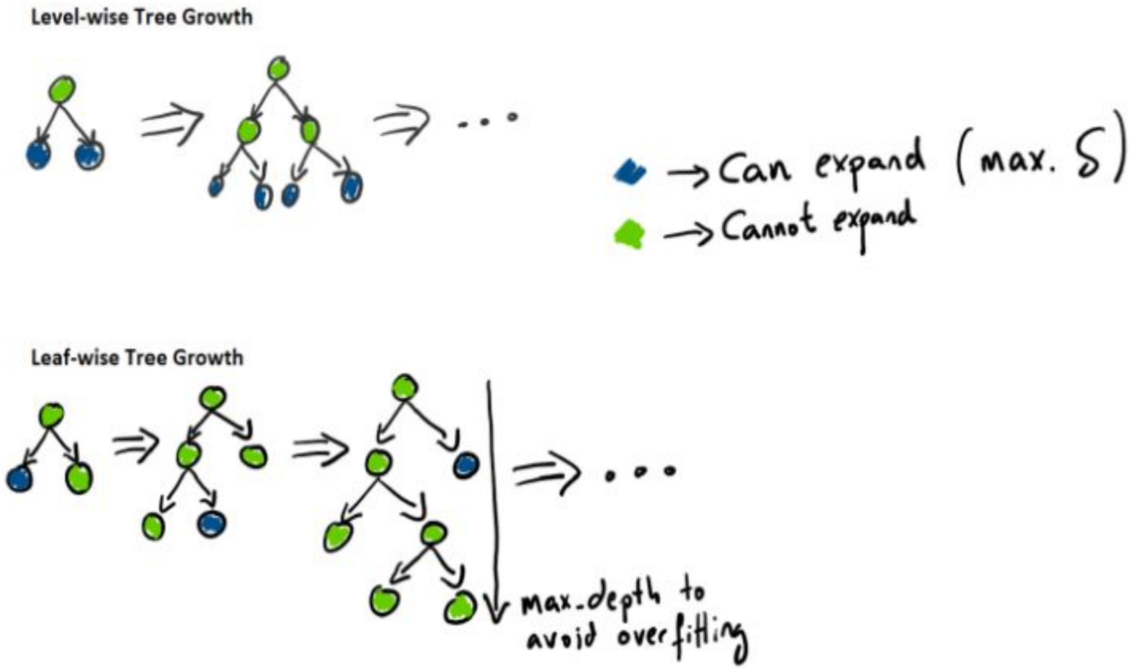
Model : XGB
Recall : 0.774
Accuracy : 0.912
Precision : 0.709
AUC_ROC : 0.856
F1: 0.74



Model : XGB
Recall : 0.803
Accuracy : 0.9
Precision : 0.657
AUC_ROC : 0.861
F1: 0.723



LightGBM



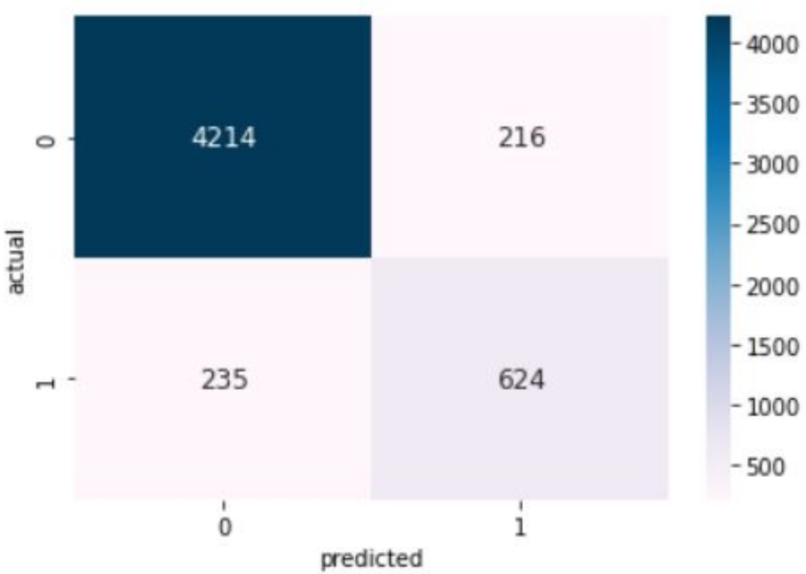
Pros:

- Faster training speed and higher efficiency
- Lower memory usage
- Compatibility with Large Datasets

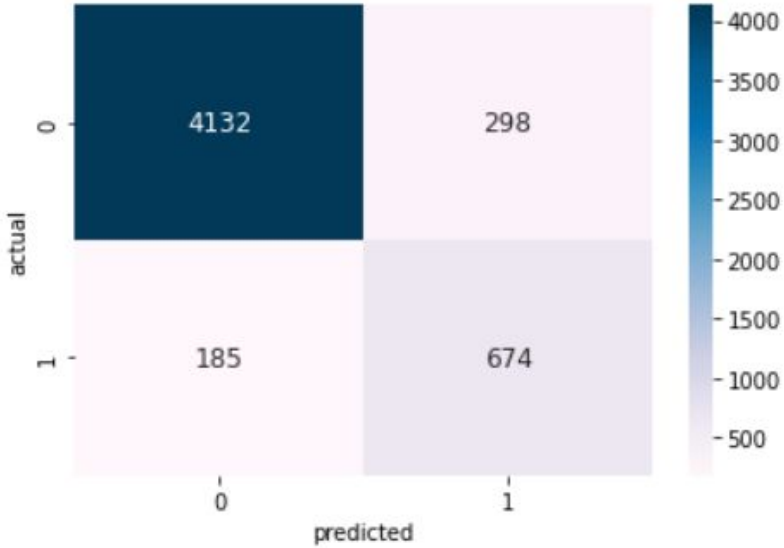
Cons:

- It is not suitable for small data set as it will overfit.

```
Model : LGBM
Recall :  0.726
Accuracy :  0.915
Precision :  0.743
AUC_ROC :  0.839
F1:  0.735
```




```
Model : LGBM
Recall :  0.785
Accuracy :  0.909
Precision :  0.693
AUC_ROC :  0.859
F1:  0.736
```



Model Selection

Model	Recall	Accuracy	Precision	AUC	F1
Logistic Regression	0.79	0.9	0.67	0.86	0.72
Decision Tree	0.77	0.89	0.65	0.84	0.7
Random Forest	0.81	0.9	0.66	0.87	0.73
K-Nearest Neighbors	0.72	0.89	0.64	0.82	0.68
Naive Bayes	0.92	0.69	0.33	0.79	0.5
XGBoost	0.8	0.9	0.66	0.86	0.72
LGBM	0.79	0.91	0.69	0.86	0.74

For our task, we can consider that achieving a high recall is more important since we would like to detect as maximum default transactions as possible to prevent losses.

Random Forest Classifier and XGBoost both have high recall values more than 80% and high AUC scores so we can choose either one of these models for our task.

Conclusion

● In general, all models have comparable accuracy. Nevertheless, because the classes are imbalanced (the proportion of non-default credit cards is higher than default) this metric is misleading.

● Furthermore, accuracy does not consider the rate of false positives (non-default credits cards that were predicted as default) and false negatives (default credit cards that were incorrectly predicted as non-default).

● Both cases have negative impact on the bank, since false positives leads to unsatisfied customers and false negatives leads to financial loss.

● XGBoost Classifier and Random Forest Classifier are giving us the best Recall, F1-score, and AUC Score among other algorithms. We can conclude that these two algorithms are the best to predict whether the credit card is default or not default according to our analysis.

Challenges

- Imbalanced dataset for minority class.
- Hyperparameter tuning
- Problem of overfitting.
- Choosing the right features for modelling.
- Choosing the right model to get the best scores.

THANK YOU !!