CAPSTONE PROJECT
CREDIT CARD DEFAULT PREDICTION

Shubham Kumar

POINTS TO DISCUSS

- Problem Statement
- Abstract
- Data Summary
- Data Pipeline
- Exploratory Data Analysis
- Observations on EDA
- Model building
- Comparison of models used
- Conclusion

PROBLEM STATEMENT

The fundamental objective of the project is to analyze credit card data collected from Taiwan-based credit card issuer and predict whether or not a consumer will default on their credit cards, as well as identify the key drivers behind this. This would inform the issuer's on who to give a credit card.

ABSTRACT

Credit risk plays a major role in the banking industry business. Bank's main activities involve granting loans, credit cards, investments, mortgages, 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. Also, many customers used their credit card beyond their repayment capacity leading to high debt accumulation. This situation creates a problem for the bank. With the help of historical data, we need to predict the defaulters. As such, machine learning offers solutions for addressing the current issue and handling credit risk, Thus this project used ML algorithms of classification to get predictions with accuracy to detect the default users based on previous six months data.

DATA SUMMARY

The provided data set has following different columns of over 25 variables and they are given below

CATEGORICAL VARIABLES

- **GENDER** - [1 – Male ,2 – female]
- **EDUCATION-**
  [1 – Graduate School,

  2 – University ,

  3 – High School,

  4 – Others]
- **MARRIAGE** -
  [1 – Married, 2 – Single ,3 – Others]

- **REPAYMENT HISTORY**

  PAY_1 – September

  PAY_2 – August

  PAY_3 – July

  PAY_4 – June

  PAY_5 – May

  PAY_6 – April

- **DEFAULT - [1- YES , 0 - NO]**

NUMERICAL VARIABLES

- **LIMIT_BALANCE** - Amount of Credit given in New Taiwan Dollars
- **AGE** - Age in years
- **BILL AMOUNT**

  BILL_AMT1 – September

  BILL_AMT2 – August

  BILL_AMT3 – July

  BILL_AMT4 – June

  BILL_AMT5 – May

  BILL_AMT6 – April

- **PAY_AMOUNT**

  PAY_AMT1 – September

  PAY_AMT2 – August

  PAY_AMT3 – July

  PAY_AMT4 – June

  PAY_AMT5 – May

  PAY_AMT6 – April

DATA PIPELINE

We checked for missing values, column type, column name, duplicate records, identified outliers, numerical categorical data and cleaned the dataset for EDA

Exploratory data analysis using tables and graphs to derive the observations from the data

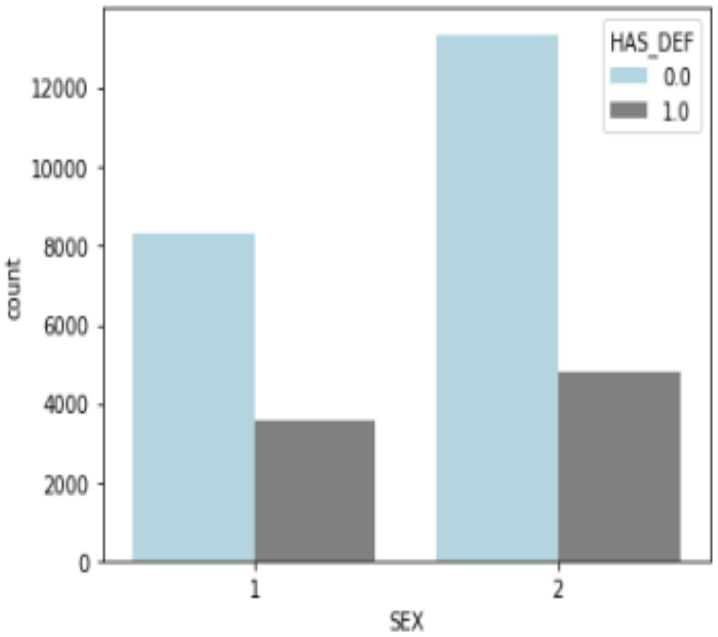
- Logistic
- SVM
- Bagging classifier
- Random Forest
- GBC
- XGBoost

EXPLORATORY DATA ANALYSIS

After Cleaning the data set, we did some exploratory data analysis using tables and graphs to derive the observations from the data and get the solution to the problem statement.

While doing EDA we tried to answer few questions below using univariate, bivariate and Multivariate Analysis.

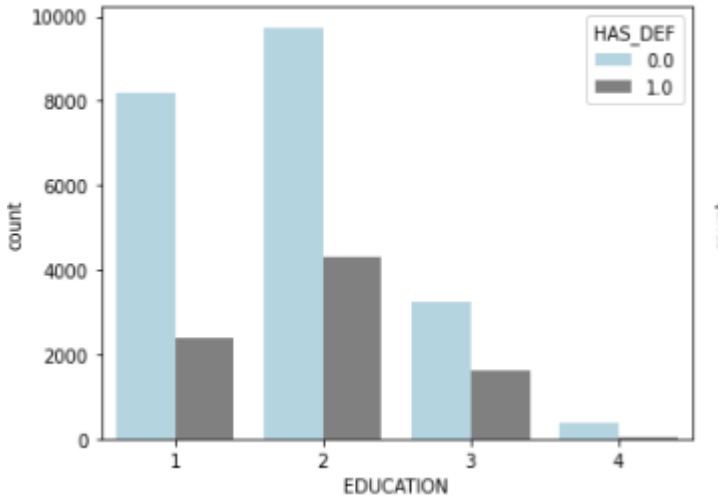
1. Is the proportion of defaults the same for men and women?



The data set contains 11888 males and 18112 females. 30% male have default payment while 26% female have default payment The proportion of defaults for men is slightly higher than the proportion of defaults for women.

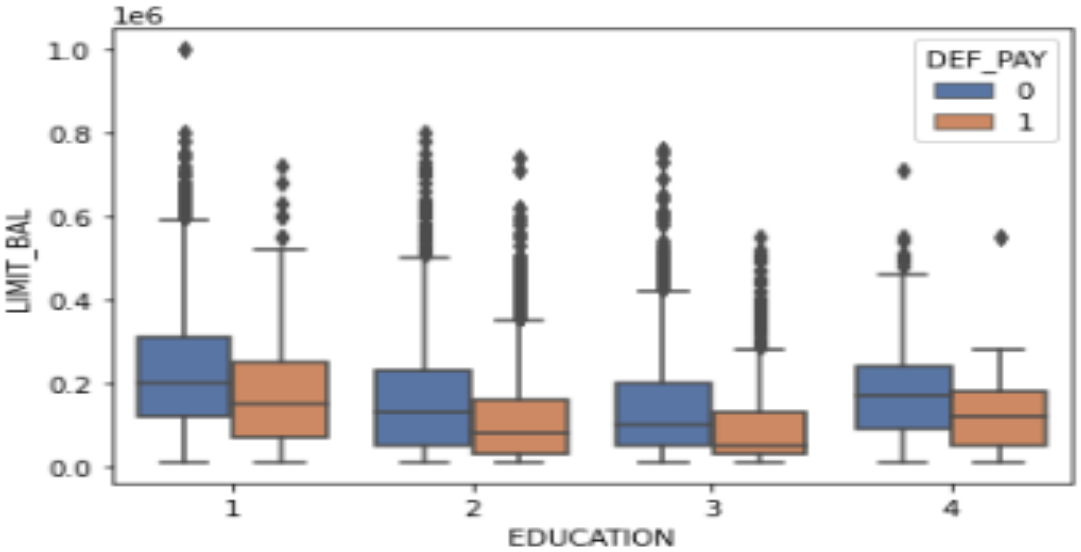
SEX	Male	Female	All
HAS_DEF			
Non-default	8291	13329	21620
Default	3597	4783	8380
All	11888	18112	30000

2. Did customers with higher education have less number of delayed payment?



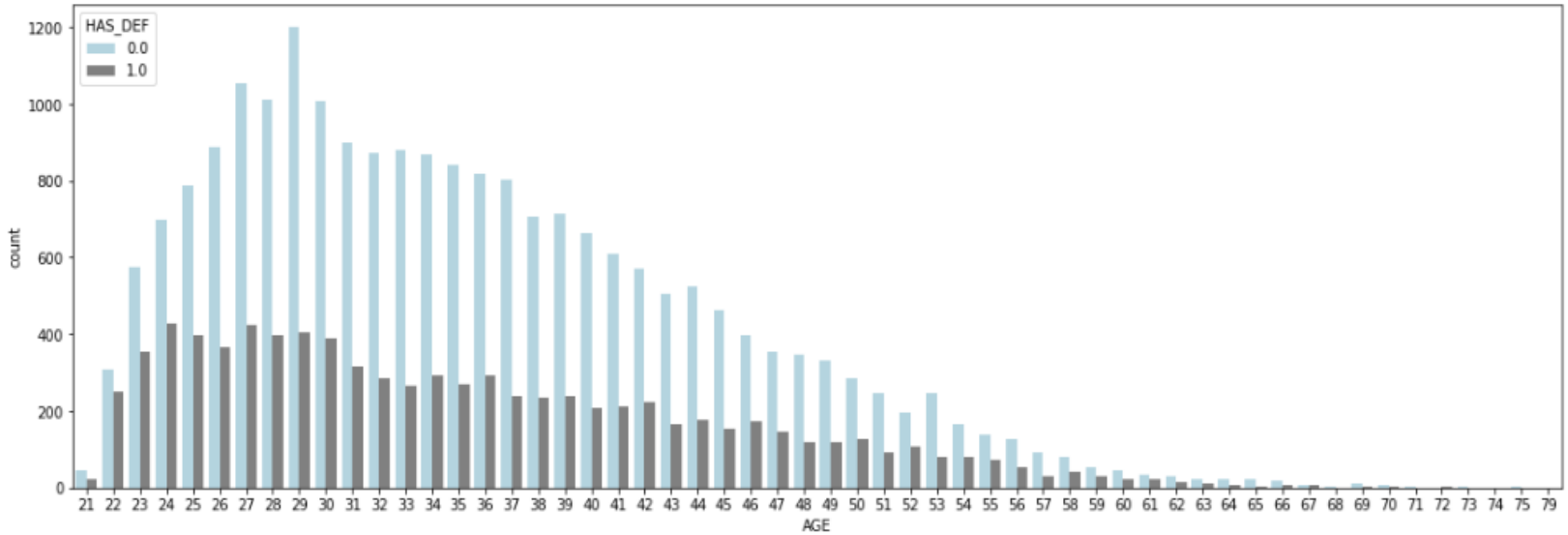
The proportion of defaults seems to decrease as the level of education increases. Customers with high school and university educational level had higher default proportions than customers with grad school education. More number of credit holders are university students followed by Graduates and then High school students

3. Did customers with a higher education level get higher credit limits?



clients who have lower than high school level education tend to have default payment more.

4. Does age has any relation with Credit Limit and Default Payments

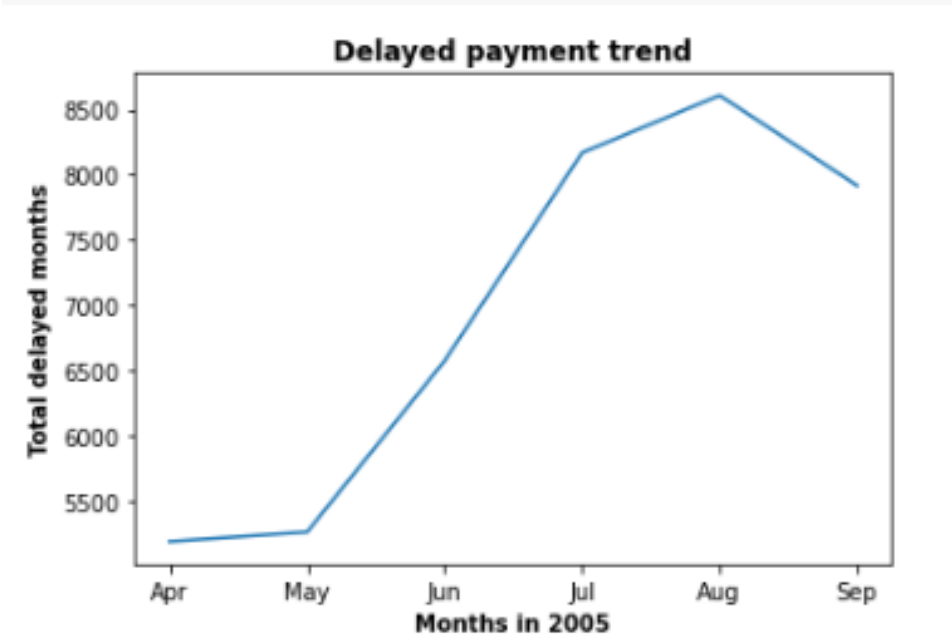


There are more adults as compared to old people above 40 and below 60

There are very low senior citizens

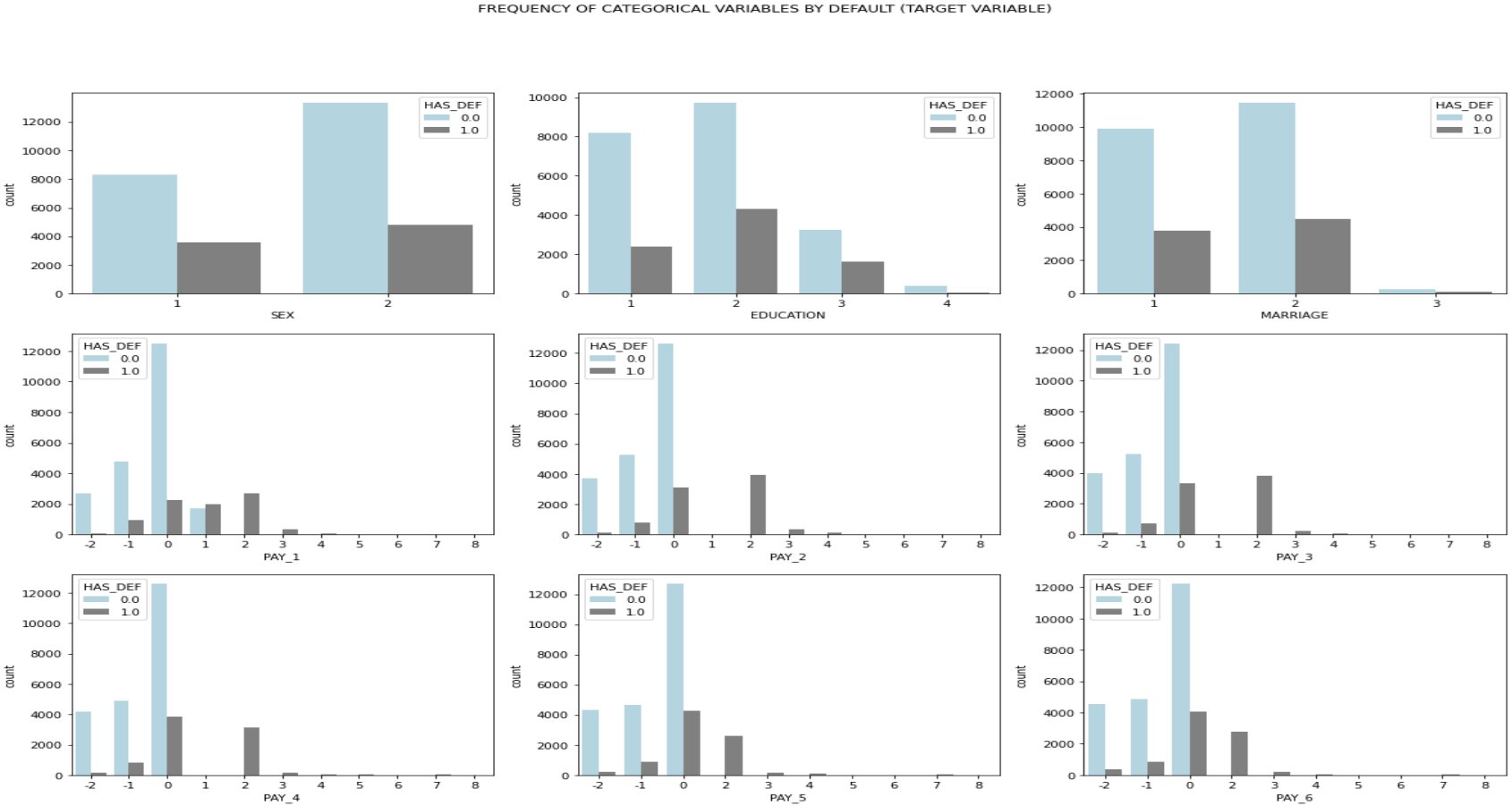
Customers aged between 30-50 had the lowest delayed payment rate, while younger groups (20-30) and older groups (50-70) all had higher delayed payment rates. However, the delayed rate dropped slightly again in customers older than 70 years.

5. Has the repayment status changed in the 6 month from April 2005 (PAY_6) to September 2005(PAY_1)?

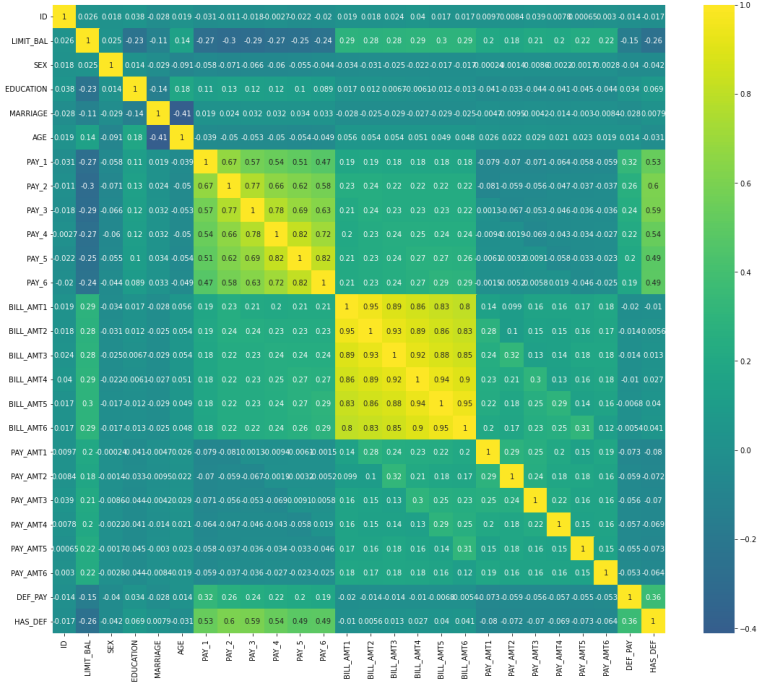


There was a huge jump from May, 2005 (PAY_5) to July, 2005 (PAY_3) when delayed payment increased significantly, then it peaked at August, 2005 (PAY_2), things started to get better in September, 2005 (PAY_1).

BIVARIATE ANALYSIS



Heat Map for Multicollinearity



"age" is inversely correlated with "Marriage".

the bill amounts from April to September are highly positively correlated.

Multicollinearity exists between the payment status variables.

Moreover, the bill amounts of each month are weakly positively correlated or even NOT correlated to the payment amount of the same month, meaning that the customers are not paying the exact amount of their bills.

The payment status(PAY_1 TO PAY_6) and credit limit(LIMIT_BAL) are negatively correlated, i.e. the later the repayment, the lower the credit limit.

The credit limit is inversely correlated with the customer having a default payment next month.

MODEL BUILDING

Train test data split(80%-20%)

- Start with default model parameters
- Hyperparameter tuning

- Model testing
- Precision, Recall Score
- Compare with the other models

LOGISTIC CLASSIFICATION

Accuracy on train data: 0.81025

Accuracy on test data: 0.80883

Precision: 0.7163461538461539

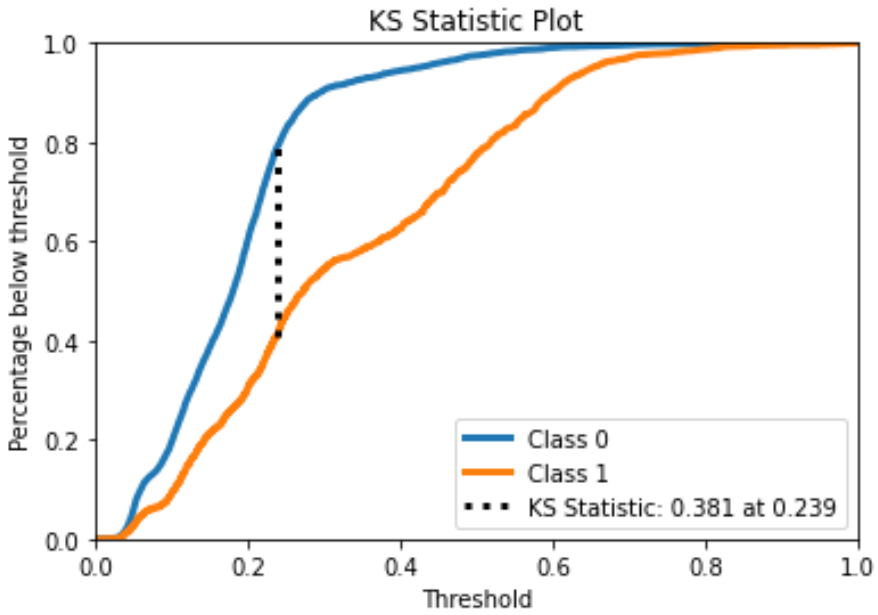
Recall: 0.22456669178598343

Parameters

'C': 1.0,

'penalty': 'l2',

'solver': 'newton-cg'



SUPPORT VECTOR CLASSIFICATION

PARAMETERS

C': 1.0,

'gamma': 'auto',

'kernel': 'rbf

Accuracy:
0.7886666666666666
Precision:
0.7322834645669292
Recall:
0.07008289374529013

RANDOM FOREST CLASSIFIER

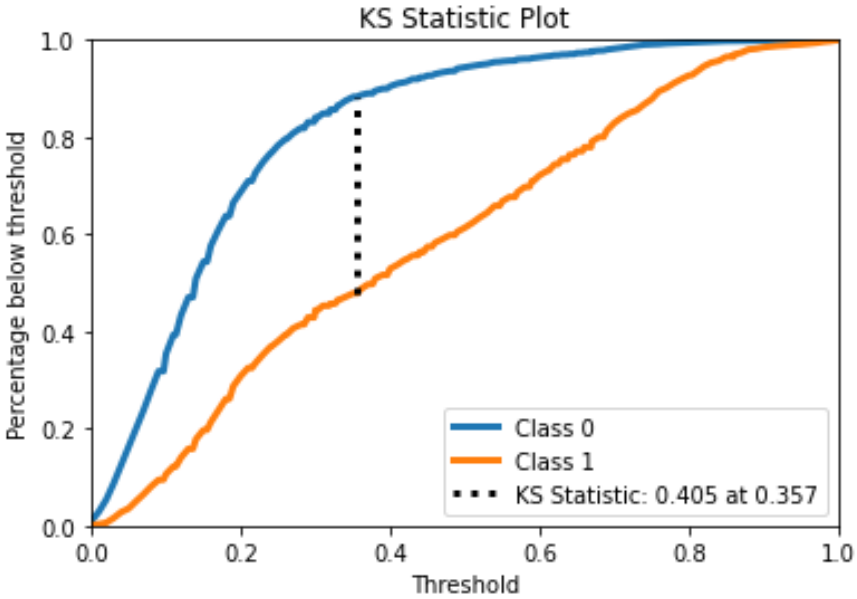
Accuracy: 0.8135

Precision: 0.6357702349869452

Recall: 0.36699321778447624

BEST PARAMETER

n_estimators': 1000



GRADIENT BOOSTING ALGORITHM

Accuracy: 0.8186666666666667

Precision: 0.665742024965326

Recall: 0.3617181612660136

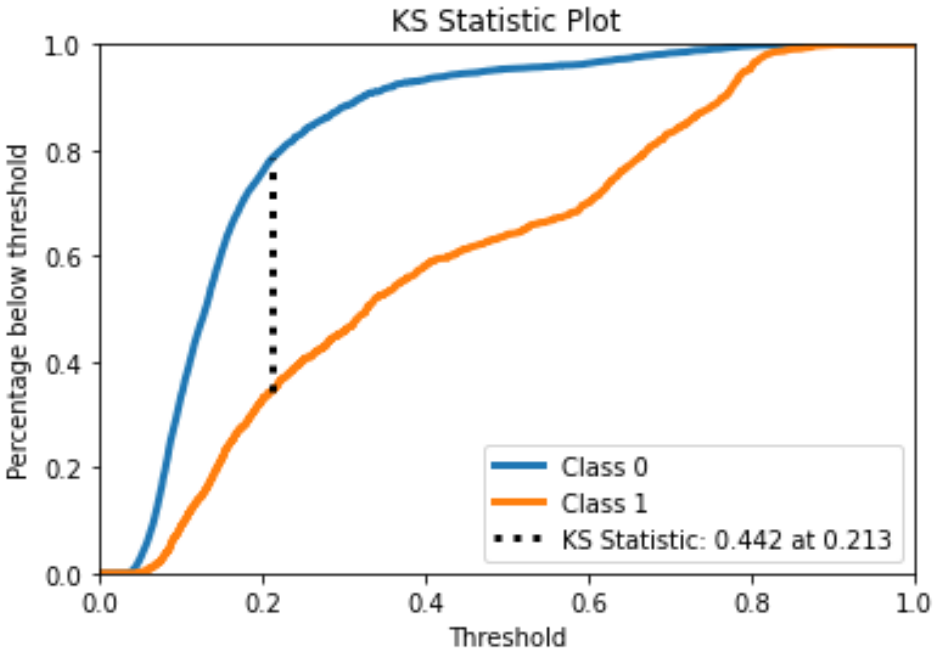
Parameters

learning_rate: 0.1,

max_depth: 3,

n_estimators: 50,

subsample: 0.5



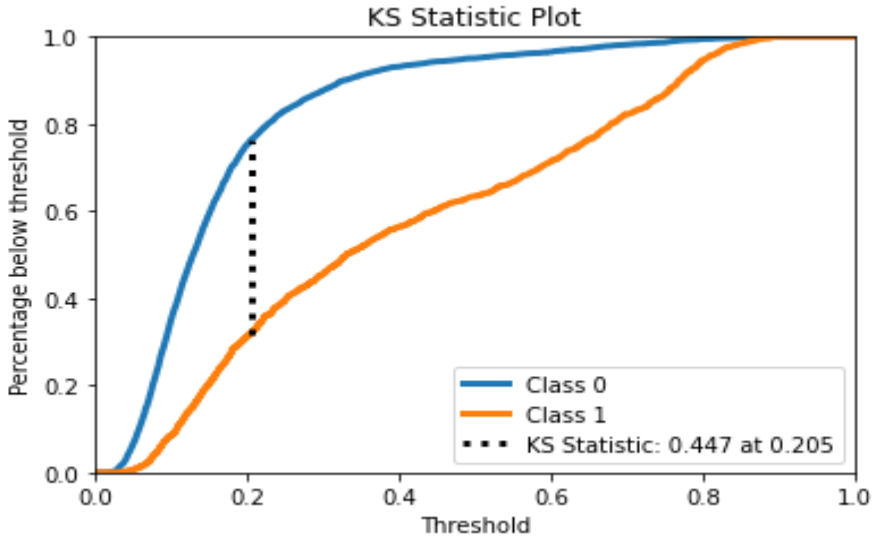
XGBOOST

The accuracy on Trained data-
0.8256666666666667

The accuracy on test data is
0.8213333333333334

The precision on test data is
0.36623963828183875

The recall on test data is
0.6778242677824268



BAGGING CLASSIFIER

Performance Metrics

Accuracy: 0.8125

Precision: 0.6268844221105527

Recall: 0.3760361718161266

PARAMETERS

'n_estimators': 45

Comparisons on Models

ML Algorithms	Accuracy on Trained Data	Accuracy on Test Data	Precision	Recall
Logistic classification	0.81025	0.8088	0.7163	0.2245
Support vector Classification	0.7886	0.7734	0.7322	0.0700
Random Forest classifier	0.8126	0.8135	0.6570	0.3669
Gradient Boosting Classifier	0.8102	0.8226	0.6886	0.3617
XGBoost	0.825	0.8213	0.3662	0.6778
Bagging Classifier	0.7996	0.8125	0.6268	0.3760

CONCLUSION

The dataset was not normally distributed and mostly it was right skewed

There is NO significant difference in the proportion of default payment across different education levels. But clients who have lower than high school level education tend to have default payment more.

Married clients have a higher default payment rate than single or other marital status clients.

People who have the payment delay for two months have a high ratio of default next month (October). In September, a quarter of customers who repay one month later have default payment next month in October. This situation does not exist in other months as almost no one repay one month later

The developed models took into account all possible factors and data. This final chosen model would benefit the bank before they make any decisions against that customers.

XGBOOST has the highest accuracy of 82.5 % with a recall of 67.7 % and KS chart value of 0.447.