Capstone Project

Credit Card Default Prediction

Team- Data Minds

Team Members

Uday Kant

Sonu Kumar

Contents

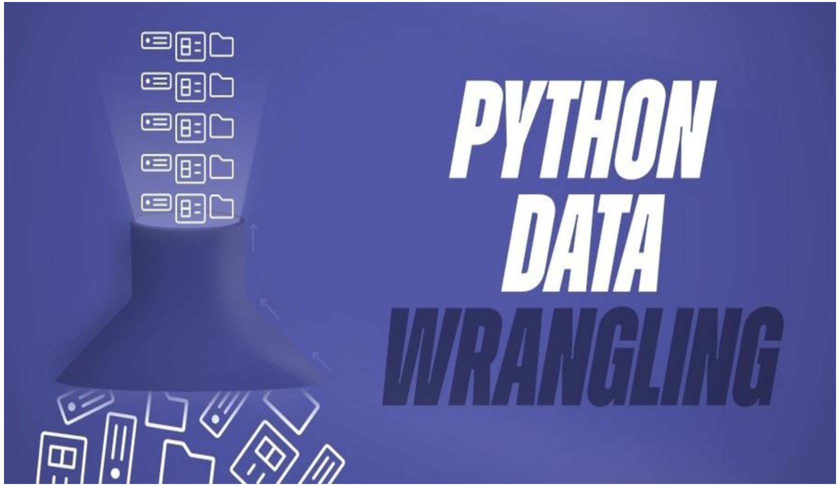
1. Introduction
2. Data Wrangling
3. Data Visualization
4. Pairplot
5. Correlation between features of the dataset
6. Sample Data
7. Data cleaning
8. Feature Scaling
9. Applying train test split
10. Applying Machine Learning Algorithm for Classification Problem
    a. Logistic regression
    b. Decision tree
    c. Random Forest
    d. Stochastic Gradient Descent
    e. K-Nearest Neighbour
    f. Support Vector Machine
11. Grid Search CV on all Models
12. Conclusion

1. Introduction

The credit card companies in Taiwan faced a cash and debt crisis in 2005, with a peak in delinquency anticipated for the third quarter of 2006. (Chou). Taiwan's card-issuing banks over issued cash and credit cards to unauthorised applicants in an effort to gain market dominance. In addition, most cardholders, regardless of their capacity to pay back, abused their credit cards for consumption and racked up substantial cash and credit card debt. Consumer financial confidence was damaged by this crisis, which also provided major challenges for cardholders and banks.

2. Data Wrangling

1. Data Wrangling is the process of gathering, collecting, and transforming Raw data into another format for better understanding, decision-making, accessing, and analysis in less time.
2. Data Wrangling is also known as Data Munging.Here in data wrangling part we observe our dataset and checking data-type, min, max value, mean, null values and many more things.



DATA MINDS

3. Data Visualization

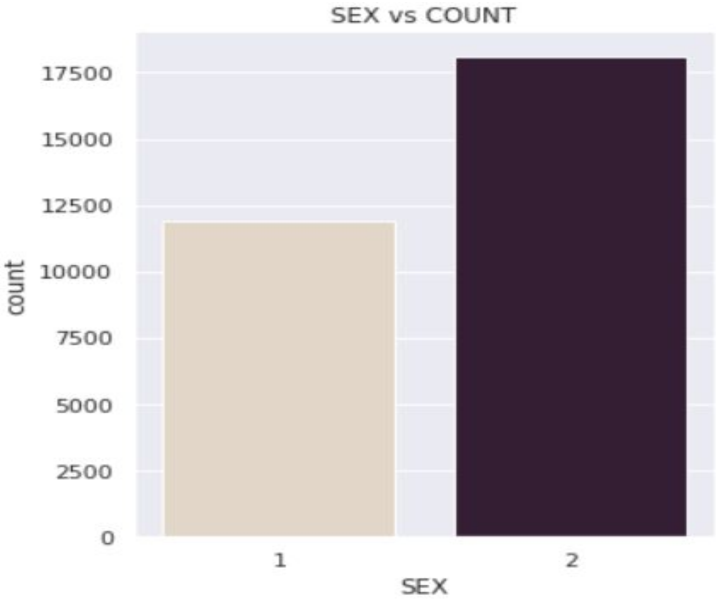


Figure 1

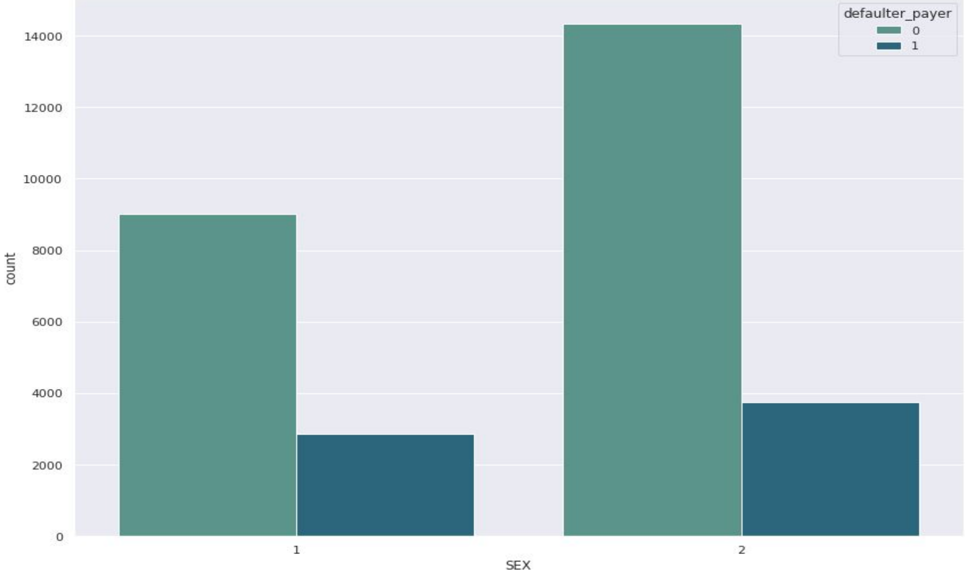


Figure 2

From the above figure-1, we can conclude that female credit card holders are greater than the male credit card holders.

From the above figure-2, we can observe that females have overall less default payments than male.

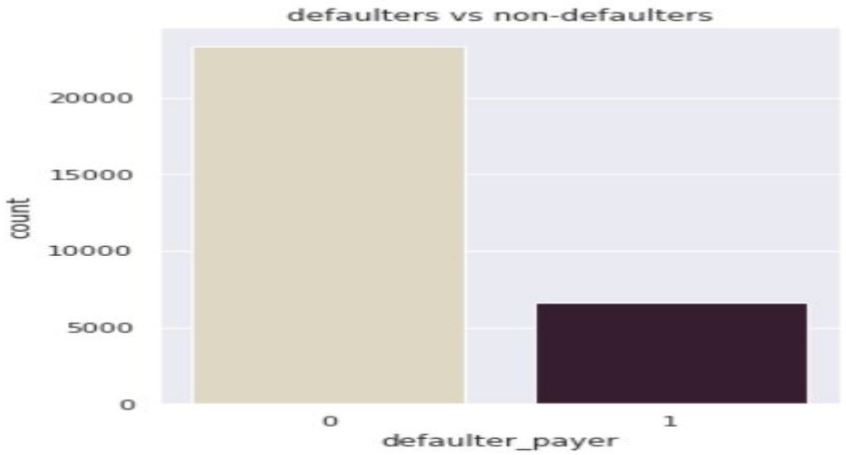


Figure 1

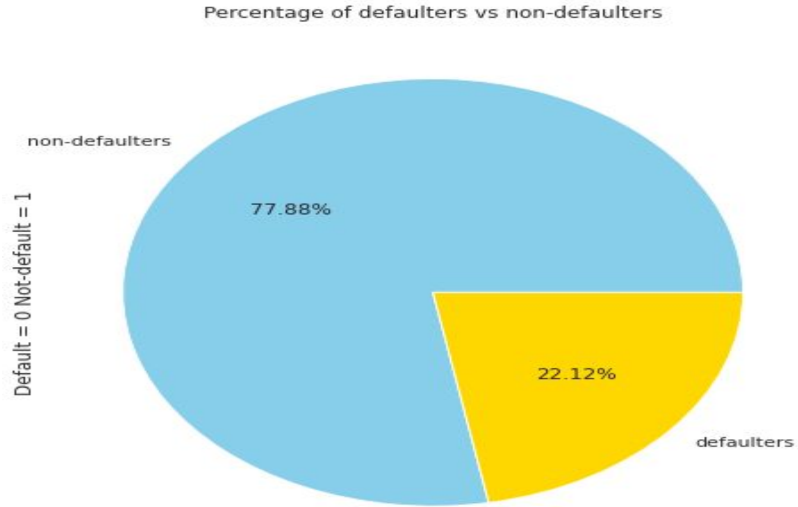
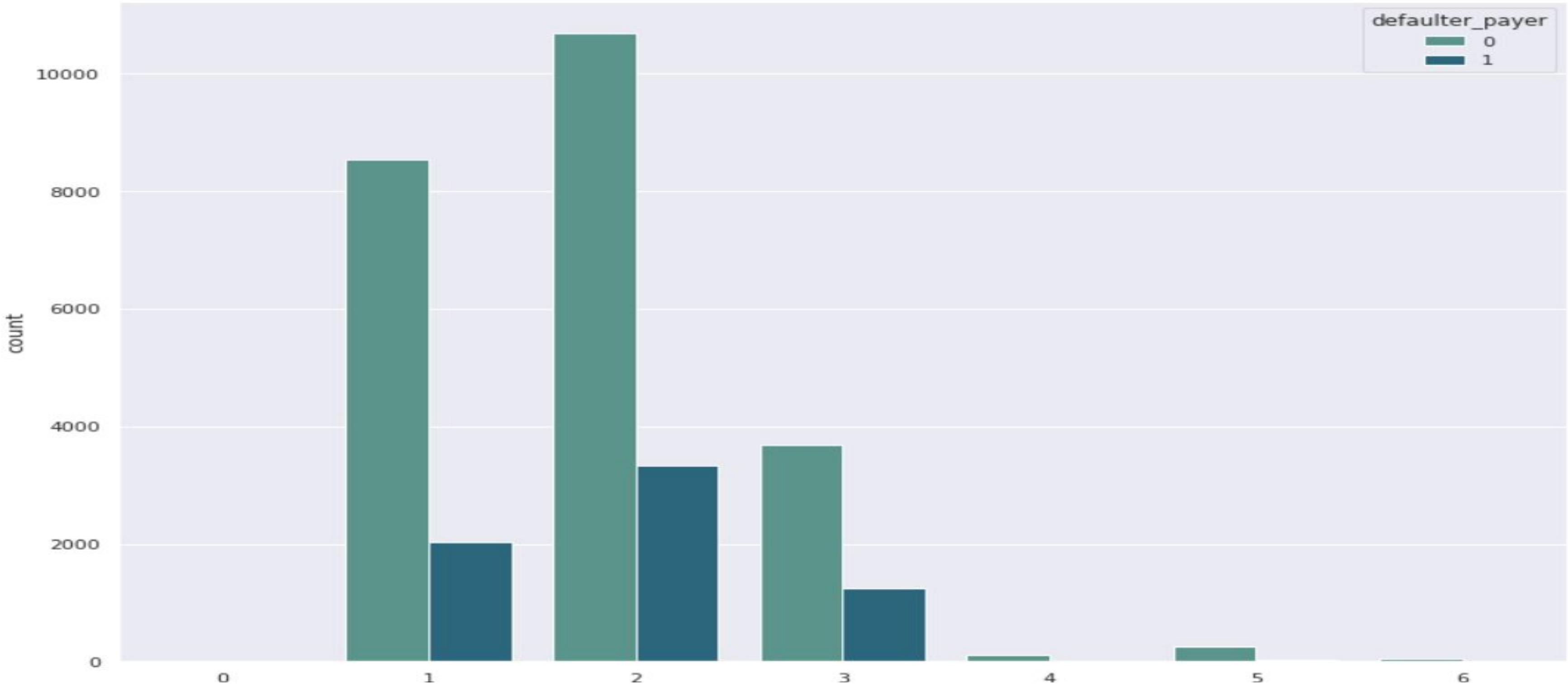


Figure 2

From the above figure-1, we can conclude that most of the credit card holders are non-defaulters. Around 6000 card holders come in the category of payment defaulters.

In figure-2, we can see the percentage of the defaulter and non defaulter credit card holders.Here, defaulters are less than non-defaulters. Non-defaulters are 77.88% and defaulters are 22.12%.

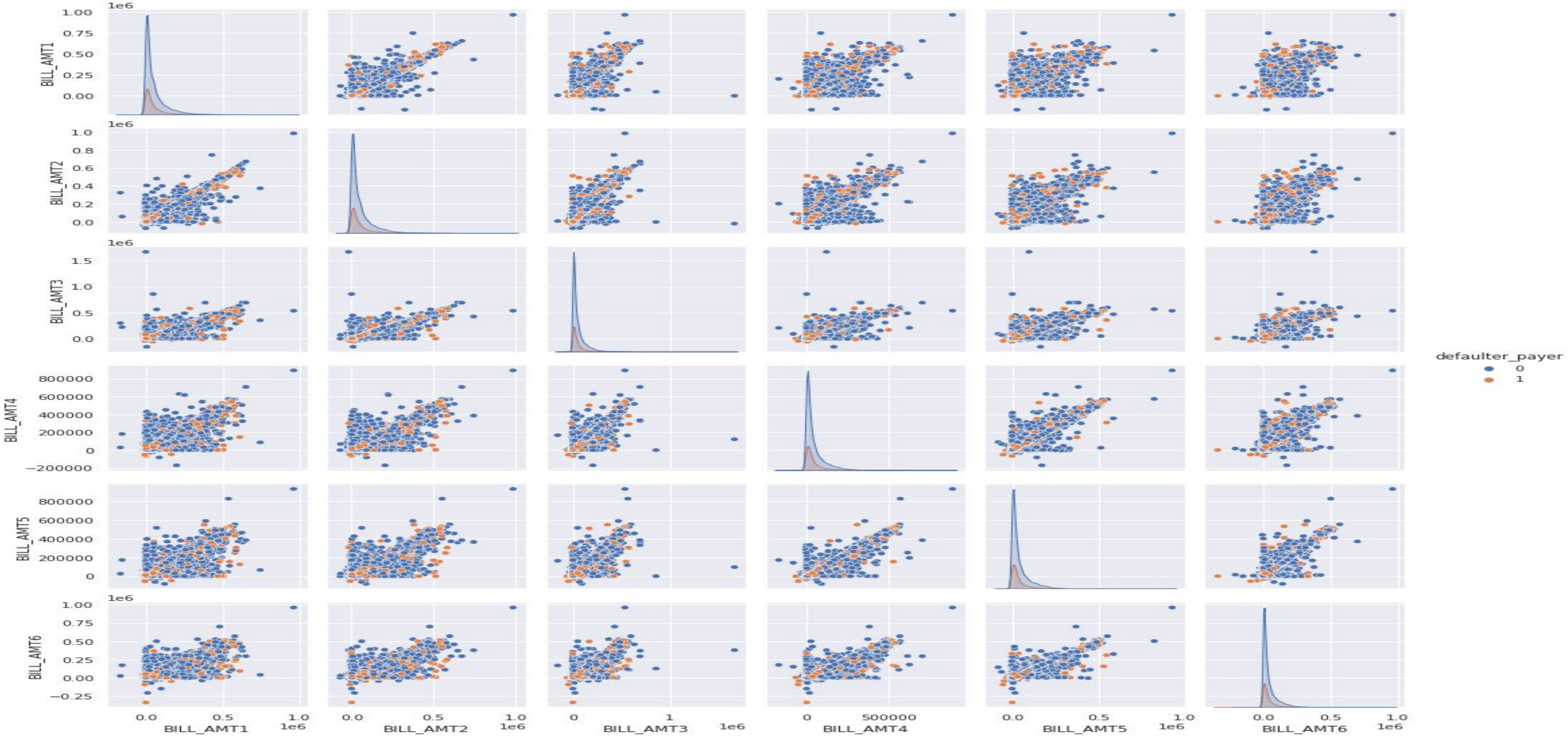


- In this graph, we can see that the university category has 15000 non defaulter credit card holders and around 3700 defaulters in payments.
- For the graduate, the non defaulter payments are around 8400 and the defaulter payments are 1900.
- For the category of high school, around 3800 are non defaulter and 1000 are defaulters.

DATA MINDS

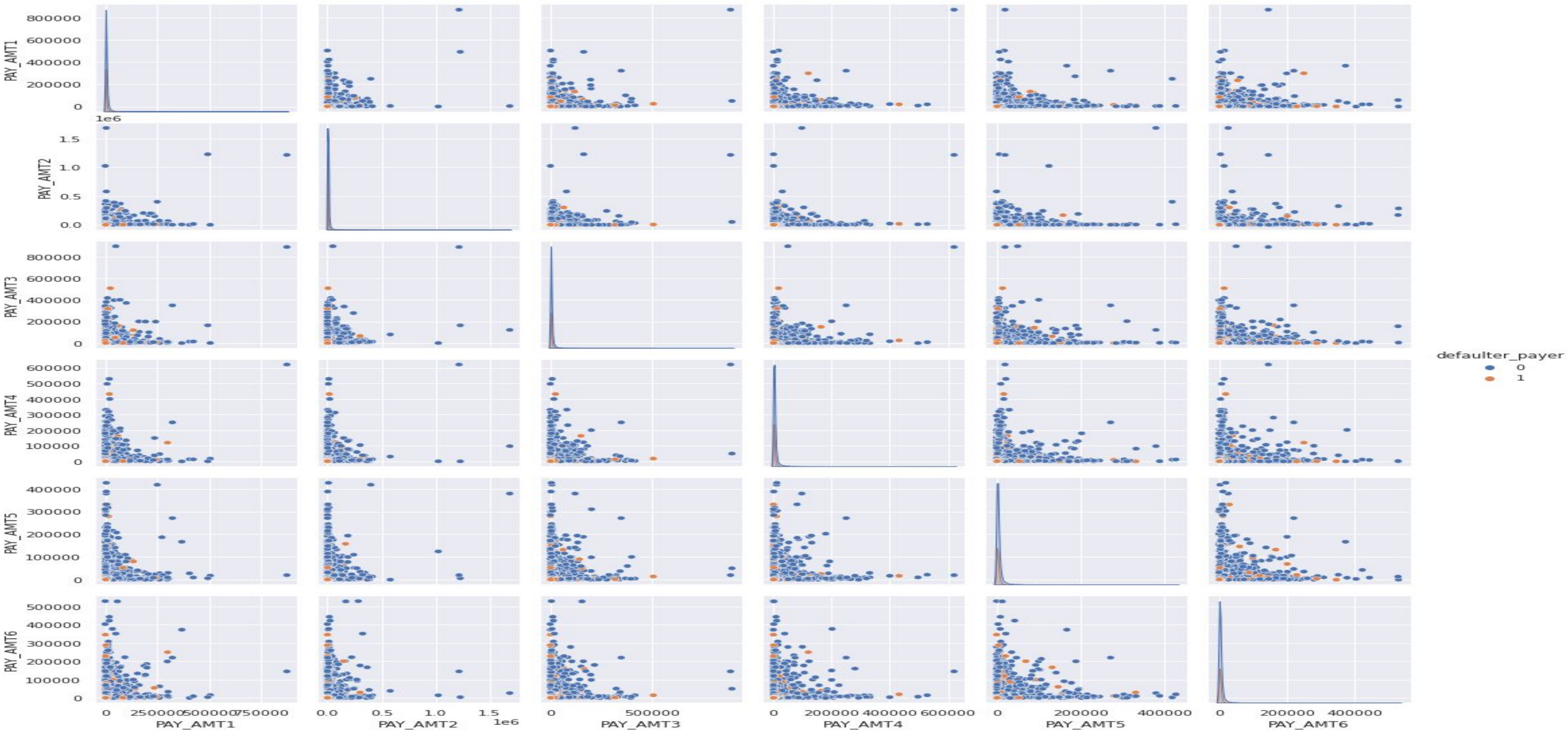
4. Pairplot

a.Pairplot between all bill amount



b. Pairplot between all payment amount

DATA MINDS



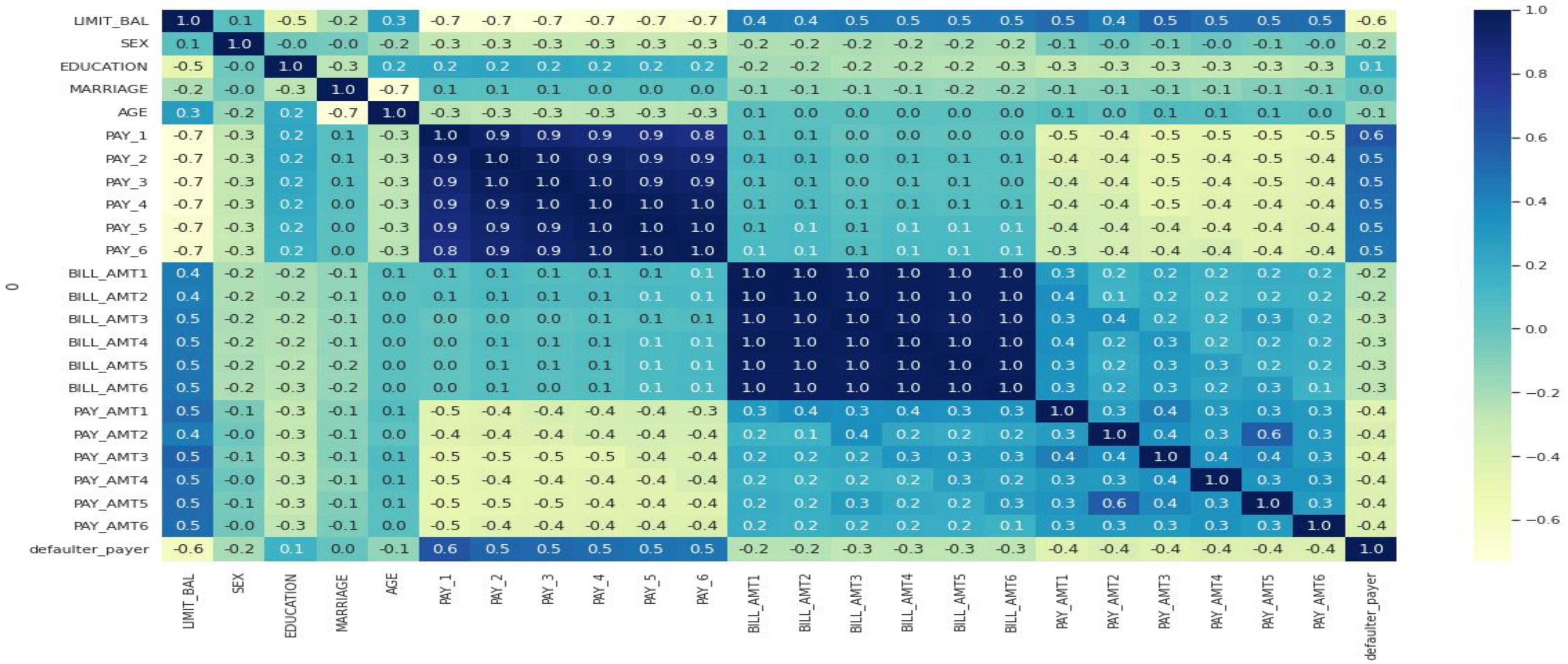
- Pairplot visualizes given data to find the relationship between them where the variables can be continuous or categorical.
- Plot pairwise relationships in a data-set.
- The distribution of the bill amounts and pay amounts are right skewed. For models requirement we have to do normalisation of data , a log transformation or standardisation can be used.

5. Sample Data

- We have 30,000 of rows so sometimes it takes huge amount of time for doing gridsearchcv. so, we sampled the dataset so that all features distribution are preserved. this is done mainly to speed up the computation.
- We take a small sample here instead of running experiments, feature engineering, and training models on all the dataset.Typically, 10–20% sampled data from original dataset is enough.
- Here we are taking 20% as a sample data from original dataset.

6. Correlation between features of the dataset

DATA MINDS



	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_1	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	defaulter_payer
LIMIT_BAL	1.0	0.1	-0.5	-0.2	0.3	-0.7	-0.7	-0.7	-0.7	-0.7	-0.7	0.4	0.4	0.5	0.5	0.5	0.5	0.5	0.4	0.5	0.5	0.5	0.5	-0.6
SEX	0.1	1.0	-0.0	-0.0	-0.2	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.1	-0.0	-0.1	-0.0	-0.1	-0.0	-0.2
EDUCATION	-0.5	-0.0	1.0	-0.3	0.2	0.2	0.2	0.2	0.2	0.2	0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	0.1
MARRIAGE	-0.2	-0.0	-0.3	1.0	-0.7	0.1	0.1	0.1	0.0	0.0	0.0	-0.1	-0.1	-0.1	-0.1	-0.2	-0.2	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	0.0
AGE	0.3	-0.2	0.2	-0.7	1.0	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.1	0.0	-0.1
PAY_1	-0.7	-0.3	0.2	0.1	-0.3	1.0	0.9	0.9	0.9	0.9	0.8	0.1	0.1	0.0	0.0	0.0	0.0	-0.5	-0.4	-0.5	-0.5	-0.5	-0.5	0.6
PAY_2	-0.7	-0.3	0.2	0.1	-0.3	0.9	1.0	1.0	0.9	0.9	0.9	0.1	0.1	0.0	0.1	0.1	0.1	-0.4	-0.4	-0.5	-0.4	-0.5	-0.4	0.5
PAY_3	-0.7	-0.3	0.2	0.1	-0.3	0.9	1.0	1.0	1.0	0.9	0.9	0.1	0.1	0.0	0.1	0.1	0.0	-0.4	-0.4	-0.5	-0.4	-0.5	-0.4	0.5
PAY_4	-0.7	-0.3	0.2	0.0	-0.3	0.9	0.9	1.0	1.0	1.0	1.0	0.1	0.1	0.1	0.1	0.1	0.1	-0.4	-0.4	-0.5	-0.4	-0.4	-0.4	0.5
PAY_5	-0.7	-0.3	0.2	0.0	-0.3	0.9	0.9	0.9	1.0	1.0	1.0	0.1	0.1	0.1	0.1	0.1	0.1	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	0.5
PAY_6	-0.7	-0.3	0.2	0.0	-0.3	0.8	0.9	0.9	1.0	1.0	1.0	0.1	0.1	0.1	0.1	0.1	0.1	-0.3	-0.4	-0.4	-0.4	-0.4	-0.4	0.5
BILL_AMT1	0.4	-0.2	-0.2	-0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	1.0	1.0	1.0	1.0	1.0	1.0	0.3	0.2	0.2	0.2	0.2	0.2	-0.2
BILL_AMT2	0.4	-0.2	-0.2	-0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.1	1.0	1.0	1.0	1.0	1.0	1.0	0.4	0.1	0.2	0.2	0.2	0.2	-0.2
BILL_AMT3	0.5	-0.2	-0.2	-0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.1	1.0	1.0	1.0	1.0	1.0	1.0	0.3	0.4	0.2	0.2	0.3	0.2	-0.3
BILL_AMT4	0.5	-0.2	-0.2	-0.1	0.0	0.0	0.1	0.1	0.1	0.1	0.1	1.0	1.0	1.0	1.0	1.0	1.0	0.4	0.2	0.3	0.2	0.2	0.2	-0.3
BILL_AMT5	0.5	-0.2	-0.2	-0.2	0.0	0.0	0.1	0.1	0.1	0.1	0.1	1.0	1.0	1.0	1.0	1.0	1.0	0.3	0.2	0.3	0.3	0.2	0.2	-0.3
BILL_AMT6	0.5	-0.2	-0.3	-0.2	0.0	0.0	0.1	0.0	0.1	0.1	0.1	1.0	1.0	1.0	1.0	1.0	1.0	0.3	0.2	0.3	0.2	0.3	0.1	-0.3
PAY_AMT1	0.5	-0.1	-0.3	-0.1	0.1	-0.5	-0.4	-0.4	-0.4	-0.4	-0.3	0.3	0.4	0.3	0.4	0.3	0.3	1.0	0.3	0.4	0.3	0.3	0.3	-0.4
PAY_AMT2	0.4	-0.0	-0.3	-0.1	0.0	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	0.2	0.1	0.4	0.2	0.2	0.2	0.3	1.0	0.4	0.3	0.6	0.3	-0.4
PAY_AMT3	0.5	-0.1	-0.3	-0.1	0.1	-0.5	-0.5	-0.5	-0.5	-0.4	-0.4	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.4	1.0	0.4	0.4	0.3	-0.4
PAY_AMT4	0.5	-0.0	-0.3	-0.1	0.1	-0.5	-0.4	-0.4	-0.4	-0.4	-0.4	0.2	0.2	0.2	0.2	0.3	0.2	0.3	0.3	0.4	1.0	0.3	0.3	-0.4
PAY_AMT5	0.5	-0.1	-0.3	-0.1	0.1	-0.5	-0.5	-0.5	-0.4	-0.4	-0.4	0.2	0.2	0.3	0.2	0.2	0.3	0.3	0.6	0.4	0.3	1.0	0.3	-0.4
PAY_AMT6	0.5	-0.0	-0.3	-0.1	0.0	-0.5	-0.4	-0.4	-0.4	-0.4	-0.4	0.2	0.2	0.2	0.2	0.2	0.1	0.3	0.3	0.3	0.3	0.3	1.0	-0.4
defaulter_payer	-0.6	-0.2	0.1	0.0	-0.1	0.6	0.5	0.5	0.5	0.5	0.5	-0.2	-0.2	-0.3	-0.3	-0.3	-0.3	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	1.0

From the above collinearity, we can observe that PAY_0, PAY_X variables are the strongest predictors of default, followed by the LIMIT_BAL and PAY_AMT_X variables.

7. Data cleaning

1.EDUCATION has category 5 and 6 that are unlabelled, moreover the category 0 is undocumented.
The 0 (undocumented), 5 and 6 (label unknown) in EDUCATION can also be put in a 'Other' cathegory (thus 4)

2.MARRIAGE has a label 0 that is undocumented.

The 0 in MARRIAGE can be safely categorized as 'Other' (thus 3).

3.PAY_1 to PAY_6 has a label 0 and -2 that is undocumented but 0 has maximum frequency. So as -1 declared as properly payment at time. We merge -1 and -2 to 0 as properly payment at time.

8. Feature Scaling

- It refers to putting the values in the same range or same scale so that no variable is dominated by the other.

- It is mostly used in the categorical data where the categorical data where the categories are assigned simple integers such as 0,1,2,…. Which might represent different categories.

- Here , we are using Z score normalisation. It calculates the Z score of each value and replaces the value with the calculated Z score.

9. Applying train test split

We have divided the entire data in two category. We have divided 70 percent of data for training dataset and 30 percent data for test dataset.

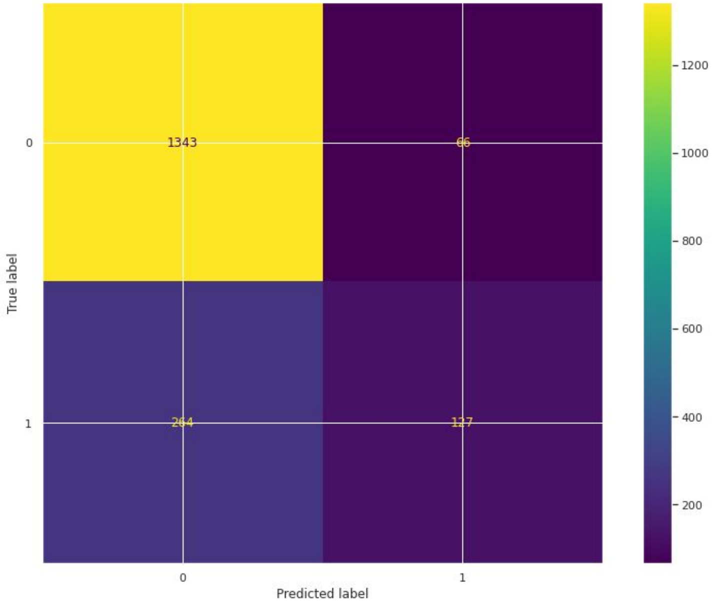
10. Applying Machine Learning Algorithm for Classification Problem

We have applied here different supervised machine learning.

a. Logistic regression
b. Decision tree
c. Random Forest
d. Stochastic Gradient Descent
e. K-Nearest Neighbour
f. Support Vector Machine

a. Logistic regression

In Logistic Regression, we wish to model a dependent variable(Y) in terms of one or more independent variables(X). It is a method for classification. This algorithm is used for the dependent variable that is Categorical. Y is modeled using a function that gives output between 0 and 1 for all values of X. In Logistic Regression, the Sigmoid Function is used.



By confusion matrix, we can conclude our accuracy, precision, recall, f1 score, and ROC.

- Accuracy-0.81
- Precision-0.658031
- recall-0.324808
- f1 score-0.434932
- ROC-0.638983

b. Decision tree

The idea of a decision tree is to divide the data set into smaller data sets based on the descriptive features until you reach a small enough set that contains data points that fall under one label.

Advantages of Decision Trees

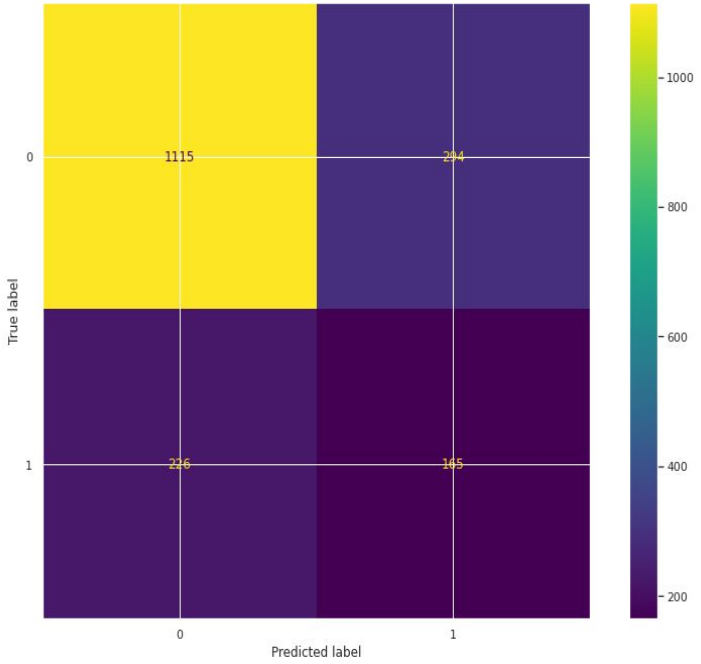
Decision trees are easy to interpret. To build a decision tree requires little data preparation from the user- there is no need to normalize data

Disadvantages of Decision Trees

Decision trees are likely to overfit noisy data. The probability of overfitting on noise increases as a tree gets deeper

By confusion matrix, we can conclude our accuracy, precision, recall, f1 score, and ROC.

Accuracy-0.711111
Precision-0.359477
recall-0.421995
f1 score-0.388235
ROC-0.606668

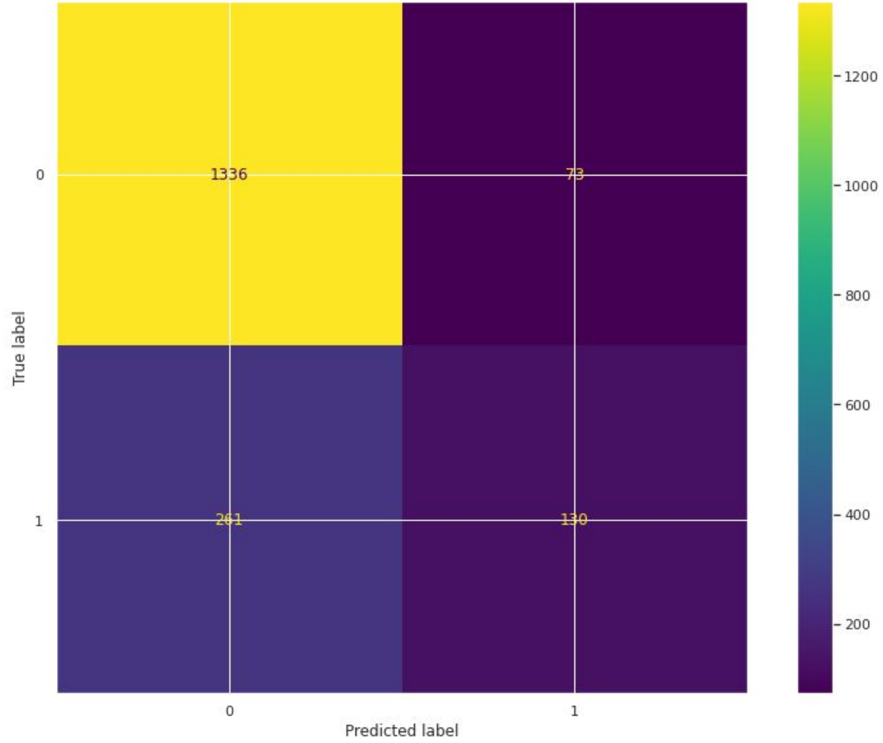


c. Random Forest

Random Forest is a supervised learning algorithm, it creates a forest and makes it somehow random. The "forest“ it builds, is an ensemble of Decision Trees.

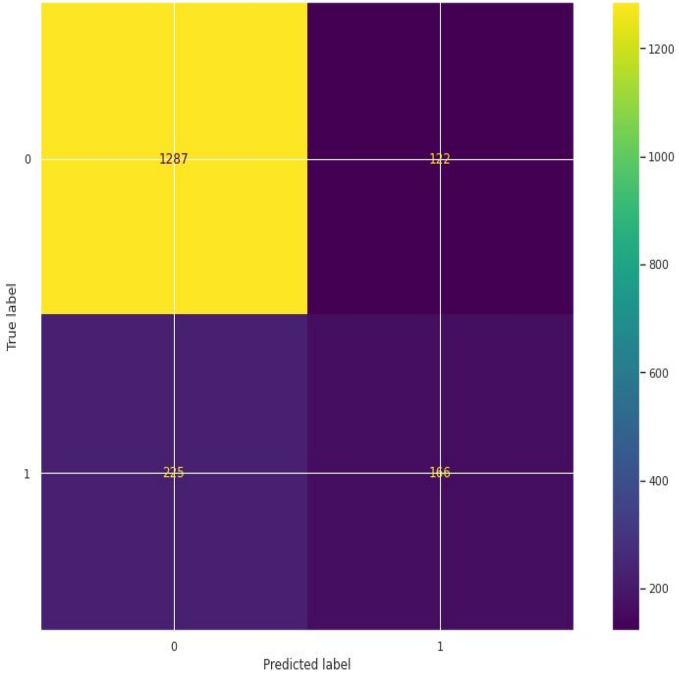
By confusion matrix, we can conclude our accuracy, precision, recall, f1 score, and ROC.

- Accuracy-0.814444
- Precision-0.640394
- Recall-0.332481
- f1 score-0.43771
- ROC-0.640336



d. Stochastic Gradient Descent

Stochastic gradient descent is an optimization algorithm often used i machine learning applications to find the model parameters that correspond to the best fit between predicted and actual outputs



By confusion matrix, we can conclude our accuracy, precision, recall, f1 score, and ROC.

- Accuracy-0.807222
- Precision-0.576389
- Recall-0.424552
- f1 score-0.488954
- ROC-0.668983

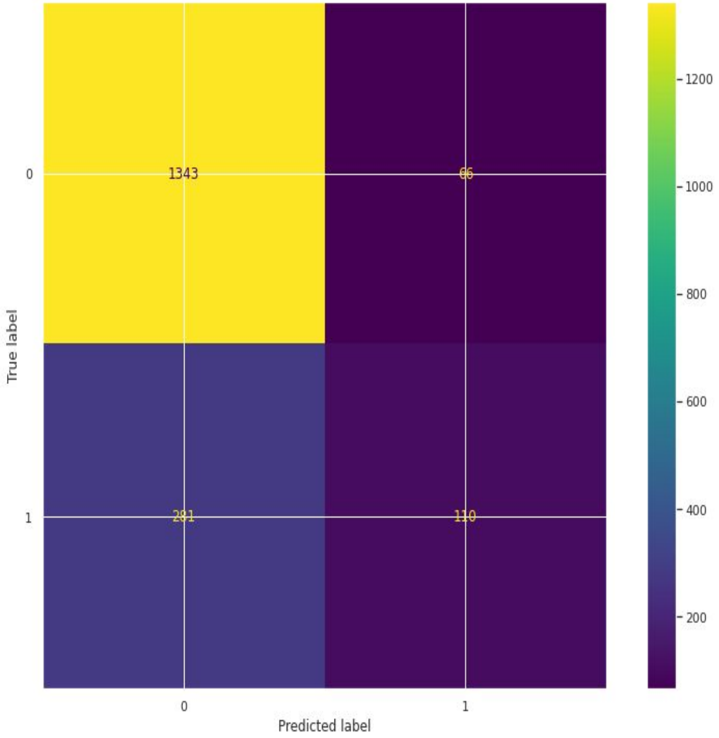
e. K-Nearest Neighbour

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry.

KNN focuses on easy implementation and good performance at the cost of computational time, but in our case the size of the dataset is considerably small so we can apply KNN.

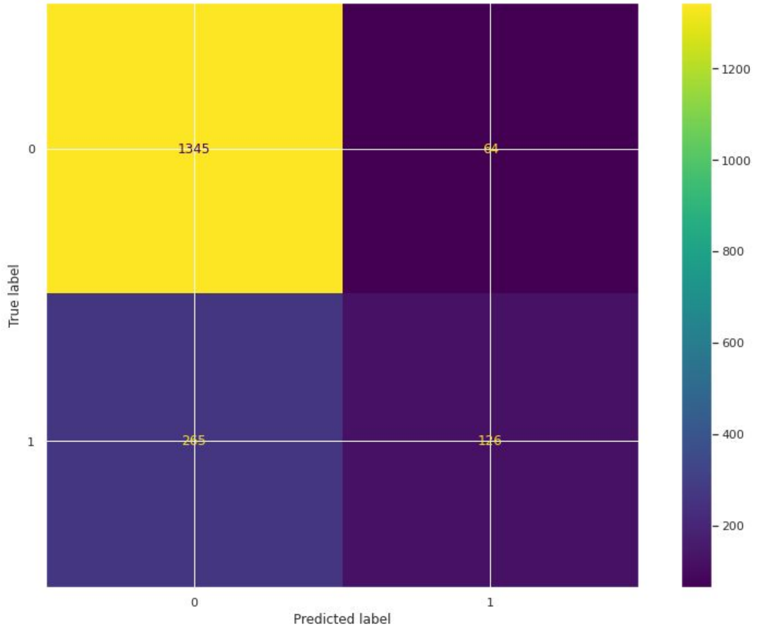
By confusion matrix, we can conclude our accuracy, precision, recall, f1 score, and ROC.

- Accuracy-0.807222
- Precision-0.625
- Recall-0.28133
- f1 score-0.388007
- ROC-0.617244



f. Support Vector Machine

In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

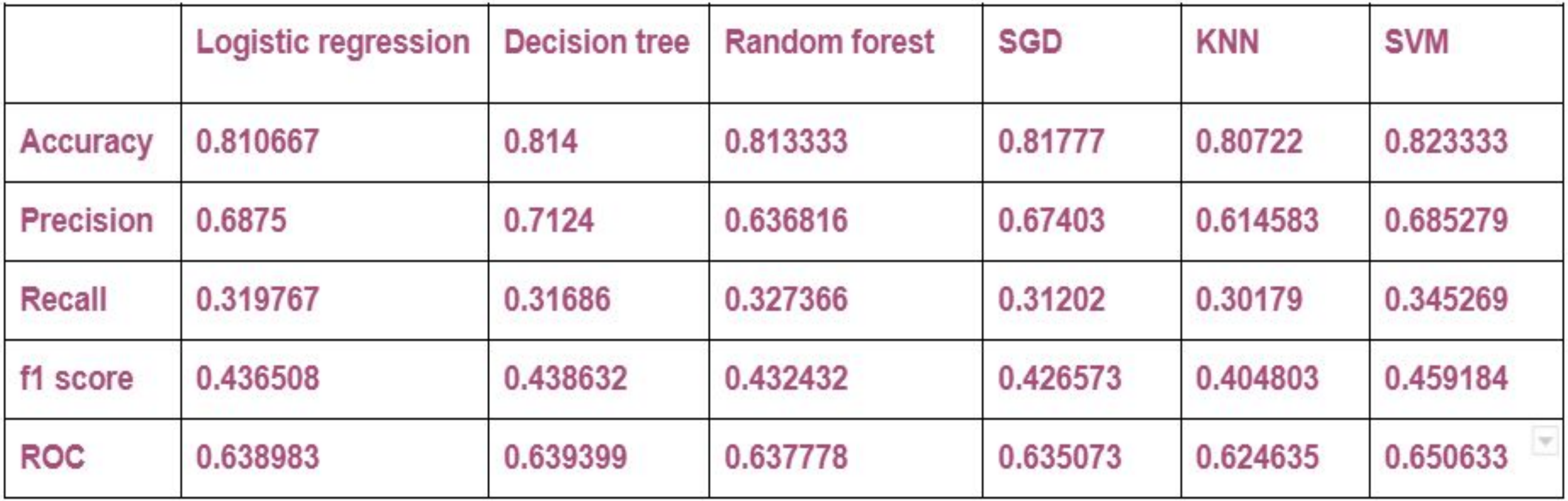


By confusion matrix, we can conclude our accuracy, precision, recall, f1 score, and ROC.

From the above confusion matrix, we can conclude our accuracy, precision, recall, f1 score, and ROC.

- Accuracy-0.817222
- Precision-0.66315
- Recall-0.32225
- f1 score-0.43373
- ROC-0.638414

11. Grid Search CV on all Models



	Logistic regression	Decision tree	Random forest	SGD	KNN	SVM
Accuracy	0.810667	0.814	0.813333	0.81777	0.80722	0.823333
Precision	0.6875	0.7124	0.636816	0.67403	0.614583	0.685279
Recall	0.319767	0.31686	0.327366	0.31202	0.30179	0.345269
f1 score	0.436508	0.438632	0.432432	0.426573	0.404803	0.459184
ROC	0.638983	0.639399	0.637778	0.635073	0.624635	0.650633

12. Conclusion

1)Using a Logistic Regression classifier, we can predict with 81.6% accuracy, whether a customer is likely to default next month.

2)Using a Decision Tree classifier, we can predict with 81.5% accuracy, whether a customer is likely to default next month.

3)Using a Random Forest classifier, we can predict with 81.33% accuracy, whether a customer is likely to default next month.

4)Using a Stochastic Gradient Descent classifier, we can predict with 81.7% accuracy, whether a customer is likely to default next month.

5)Using a K-Nearest Neighbour classifier, we can predict with 80.7% accuracy, whether a customer is likely to default next month.

6)Using a Support Vector Machine classifier, we can predict with 82.33% accuracy, whether a customer is likely to default next month.

- The strongest predictors of default are the PAY_X (ie the repayment status in previous months), the LIMIT_BAL & the PAY_AMTX (amount paid in previous months).
- We found that we are getting best results from SVM and then Stochastic Gradient Descent and then Logistic regression.
- The credit limit is a good indicator of financial stability. Whatever mechanism the bank is currently using works well and some of the features that go into choosing the credit line can be used directly in the model for default prediction.

Demographics:- we see that being Female, More educated, Single and between 30-40 years old means a customer is more likely to make payments on time.

THANK YOU !