

Capstone Project 3


Credit Card Default Prediction

Supervised ML - Classification



Sayali Kamalapurkar

Table of Contents



| |
|------------------------------------|
| Problem Statement |
| Data Summary |
| EDA on features |
| Using SMOTE oversampling Technique |
| Feature Engineering |
| Model Implementation |
| Conclusions |

Problem Statement

In recent years, the credit card issuers in Taiwan faced the cash and credit card debt crisis and the delinquency is expected to peak in the third quarter of 2006 (Chou,2006). In order to increase market share, card-issuing banks in Taiwan over-issued cash and credit cards to unqualified applicants. At the same time, most cardholders, irrespective of their repayment ability, overused credit card for consumption and accumulated heavy credit and cash-card debts. The crisis caused the blow to consumer finance confidence and it is a big challenge for both banks and cardholders.

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.

The main objective is to build a predictive classification model, which could help in predicting the defaulters among the credit card users.

Data Summary

Attribute Information:

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable.

This study reviewed the literature and used the following 23 variables as explanatory variables:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit

X2: Gender (1 = male; 2 = female)

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others; 5 = unknown; 6 = unknown)

X4: Marital status (1 = married; 2 = single; 3 = others)

X5: Age (year)

X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows:

X6 = the repayment status in September, 2005

X7 = the repayment status in August, 2005

X8 = the repayment status in July, 2005

X9 = the repayment status in Jun, 2005

X10 = the repayment status in May, 2005

X11 = the repayment status in April, 2005

The measurement scale for the repayment status is: -2 = No consumption; -1 = pay duly; 0 = use of revolving credit (paid minimum only); 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.

Data Summary

X12-X17: Amount of bill statement (NT dollar)

X12 = amount of bill statement in September, 2005

X13 = amount of bill statement in August, 2005

X14 = amount of bill statement in July, 2005

X15 = amount of bill statement in June, 2005

X16 = amount of bill statement in May, 2005

X17 = amount of bill statement in April, 2005

X18-X23: Amount of previous payment (NT dollar)

X18 = amount paid in September, 2005

X19 = amount paid in August, 2005

X20 = amount paid in July, 2005

X21 = amount paid in June, 2005

X22 = amount paid in May, 2005

X23 = amount paid in April, 2005

This dataset contains 30,000 rows and 25 columns, including ID column.

Here, the target variable is “Default_Next_Month”

Data Summary(Continued...)

Let us see, how the data looks like -

| | 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 |
|---|----|-----------|-----|-----------|----------|-----|-------|-------|-------|-------|-----|-----------|-----------|-----------|----------|----------|----------|----------|
| 0 | 1 | 20000 | 2 | 2 | 1 | 24 | 2 | 2 | -1 | -1 | ... | 0 | 0 | 0 | 0 | 689 | 0 | 0 |
| 1 | 2 | 120000 | 2 | 2 | 2 | 26 | -1 | 2 | 0 | 0 | ... | 3272 | 3455 | 3261 | 0 | 1000 | 1000 | 1000 |
| 2 | 3 | 90000 | 2 | 2 | 2 | 34 | 0 | 0 | 0 | 0 | ... | 14331 | 14948 | 15549 | 1518 | 1500 | 1000 | 1000 |
| 3 | 4 | 50000 | 2 | 2 | 1 | 37 | 0 | 0 | 0 | 0 | ... | 28314 | 28959 | 29547 | 2000 | 2019 | 1200 | 1100 |
| 4 | 5 | 50000 | 1 | 2 | 1 | 57 | -1 | 0 | -1 | 0 | ... | 20940 | 19146 | 19131 | 2000 | 36681 | 10000 | 9000 |

5 rows × 25 columns

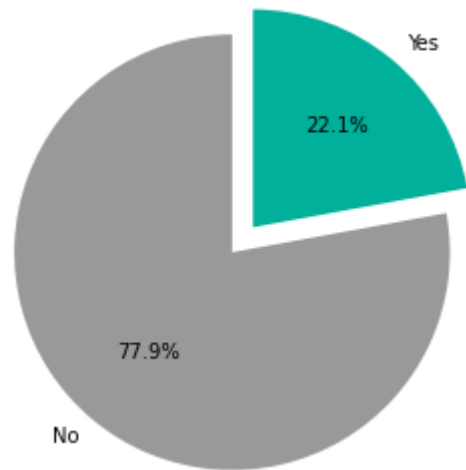
There are some categorical variables such as – sex, education, marriage

We will replace those values with their respective categories, so that it will be easy to explore and understand dataset.

There are no null values and duplicate values in the dataset.

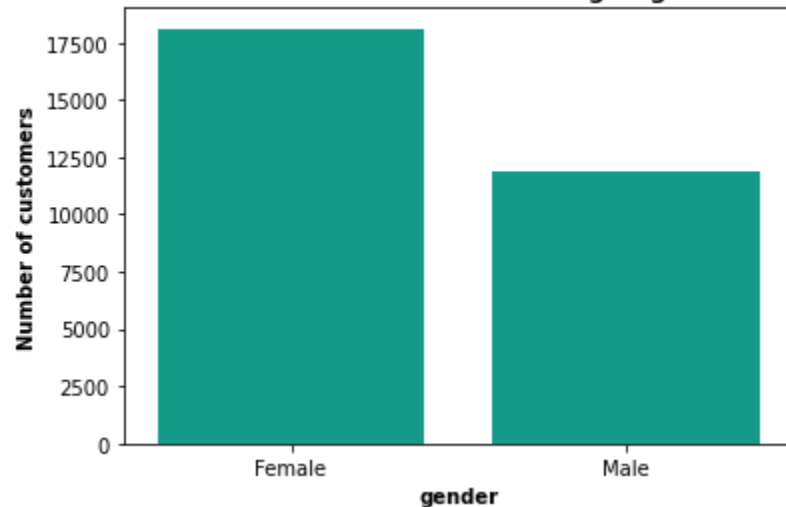
EDA on features : Univariate Analysis

Proportion of defaulters vs non-defaulters



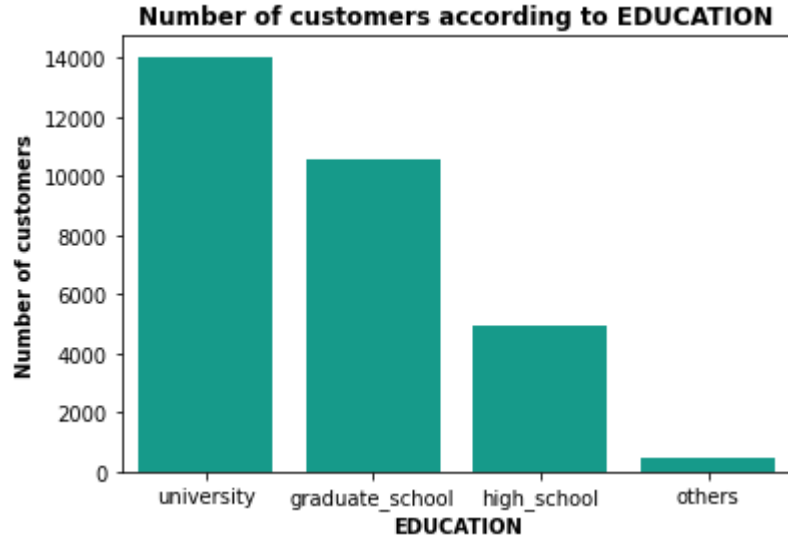
From the pie chart, we can easily identify that the credit card default dataset is highly unbalanced. We will deal with this after our EDA is done. By using SMOTE overfitting method, we will balance the dataset for defaulters and non defaulters.

Number of customers according to gender

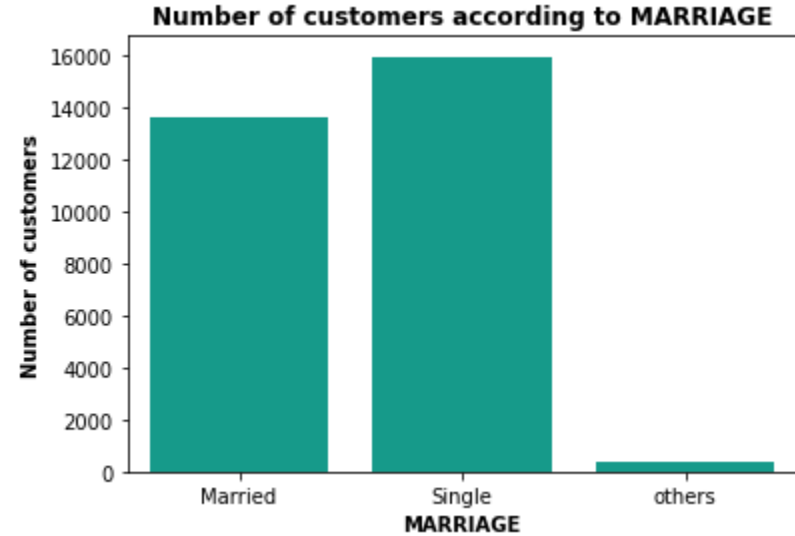


There are more female credit card holders than male credit card holders.

Continued...

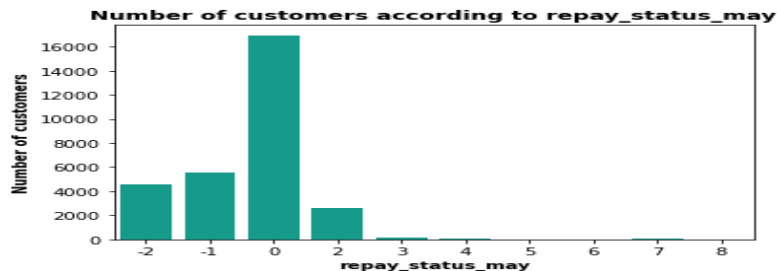
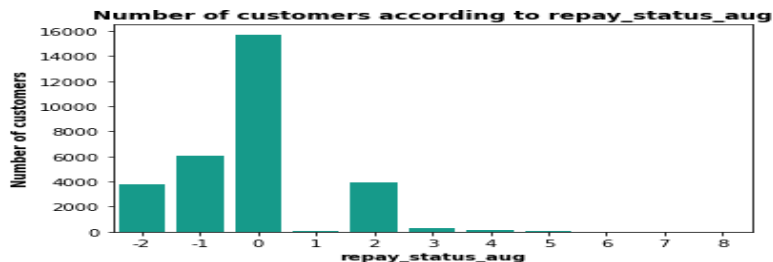
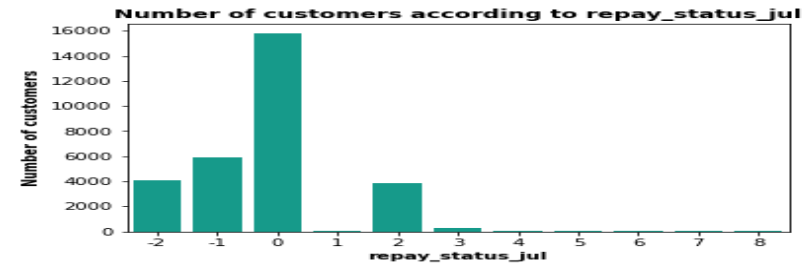
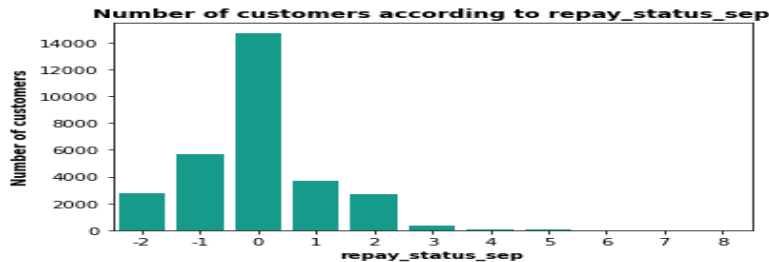


Most of the customers have completed university level, followed by graduate level and then high school level. Very few customers are below that level.



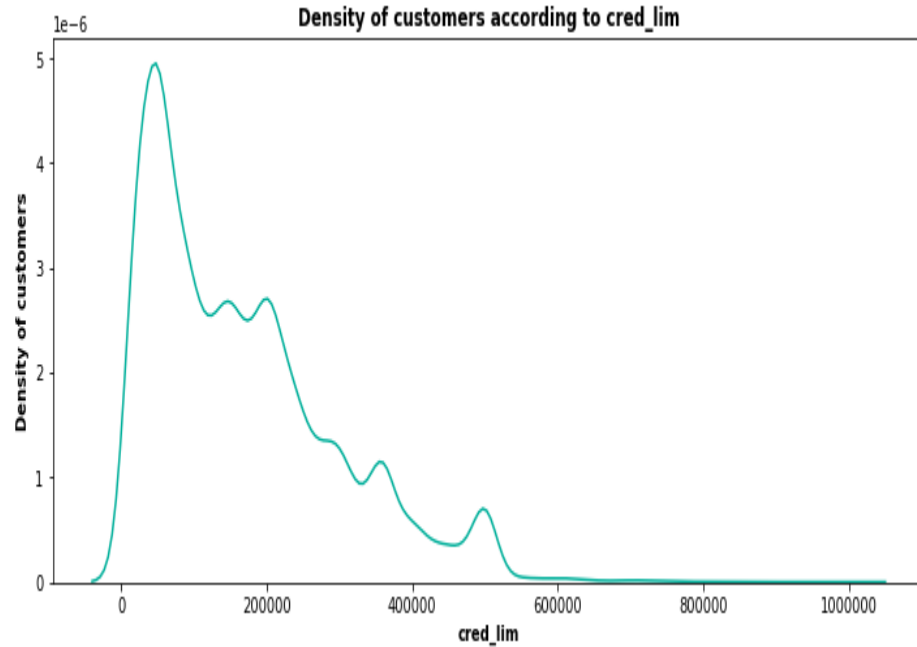
The customers who have taken credit contains mostly Single people, followed by married people.

Continued...

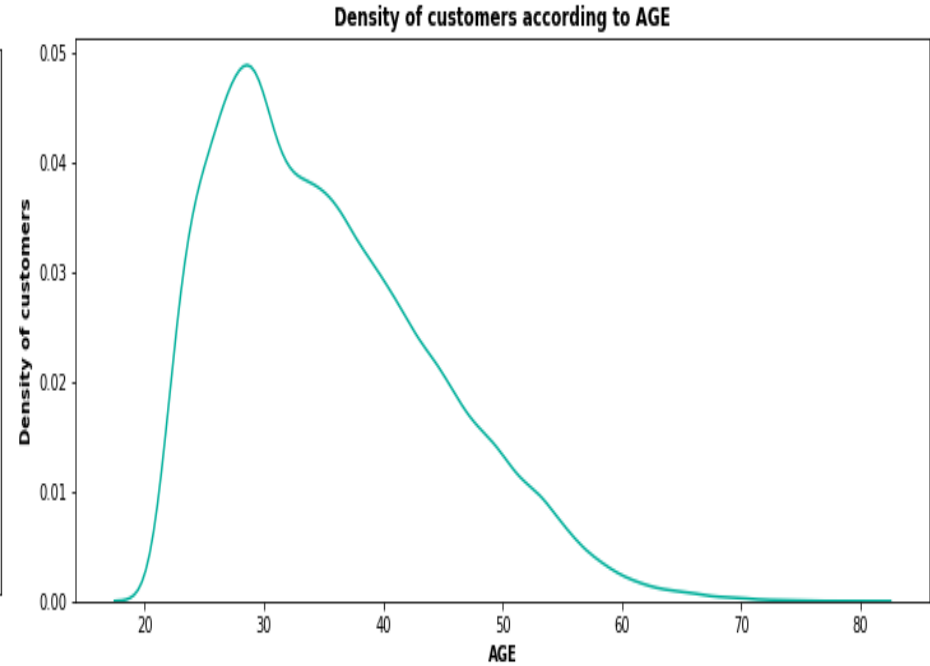


Approx 6000 customers are repaying every month (not necessary same customers, count is overall), and maximum customers are repaying the minimum credit amount only. Further, there are very few customers who are delayed in payment for 3 and more than 3 months.

Continued...

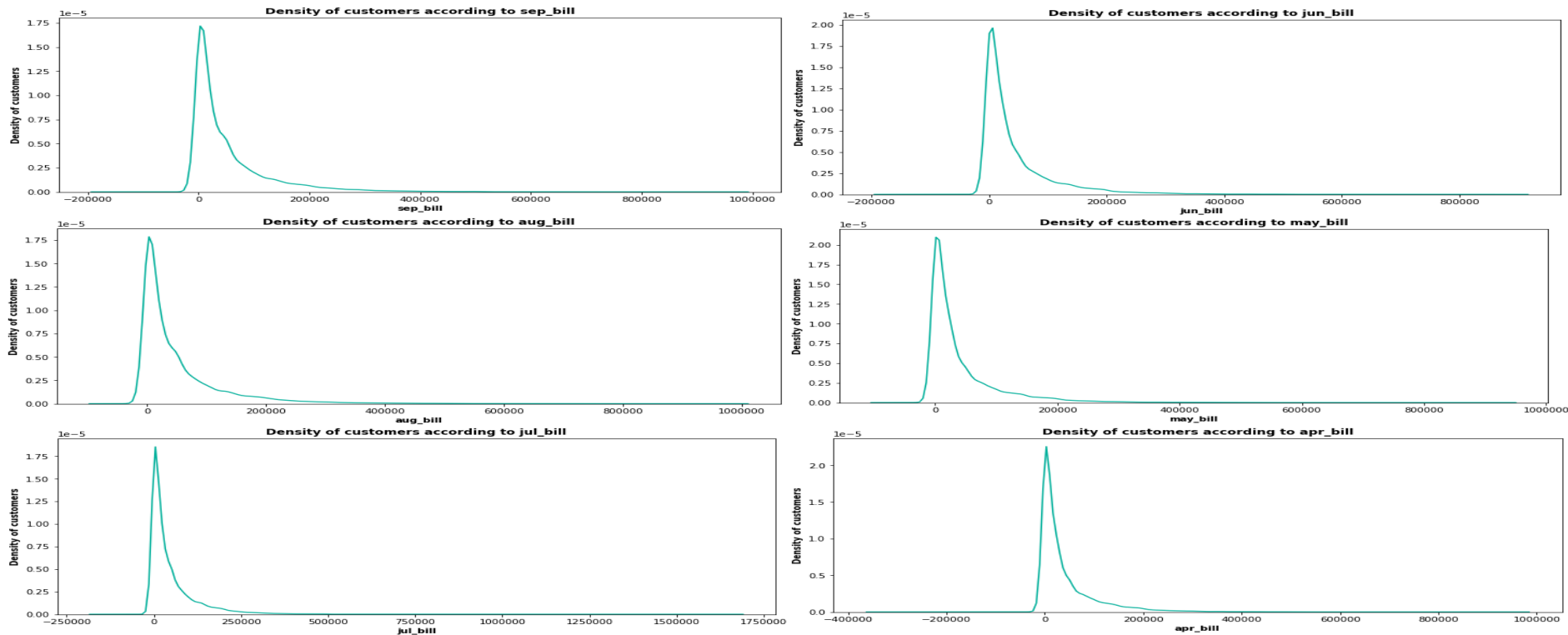


Maximum customers (more than 70%) are in the range of credit limit 20,000 to 2,50,000.



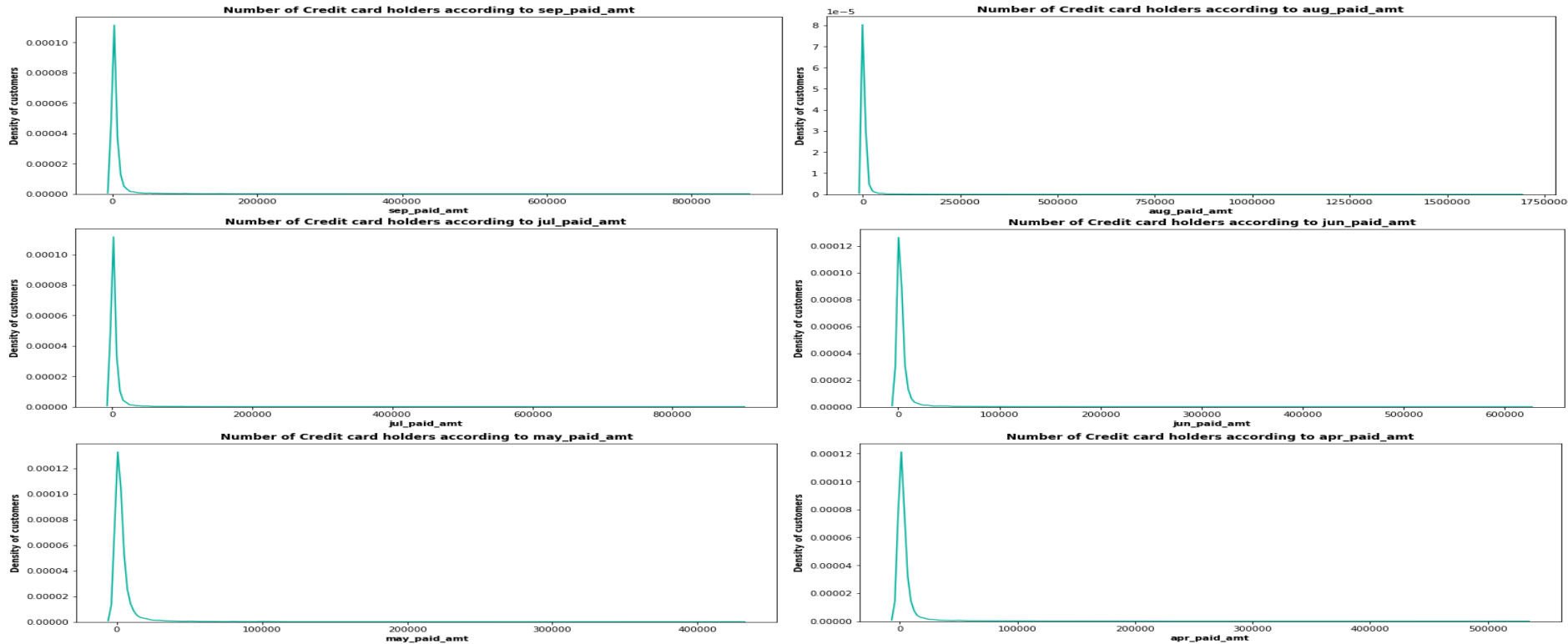
It is observed that, generally customers taking credit facility are in the age group 25 to 45 years.

Continued...



Most of the people spend between 0 to 200000 from their credit cards. Very few people have higher credit limits, and hence they are spending little higher too. That's why, all these density plots are highly right skewed.

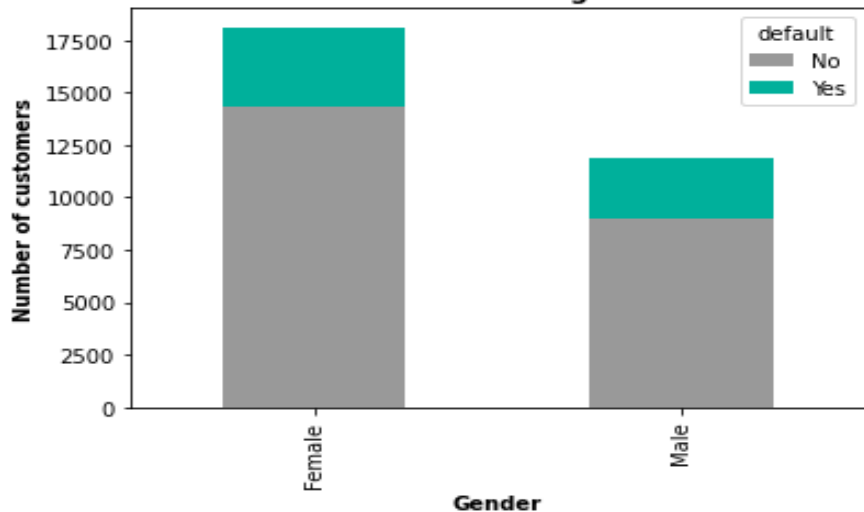
Continued...



Maximum people repaying the minimum credit amount can be seen in density plots for density of customers according to monthwise paid amount. Very few people are repaying total billed amount for every month. Hence, we can see that these monthly paid amount plots are highly right skewed.

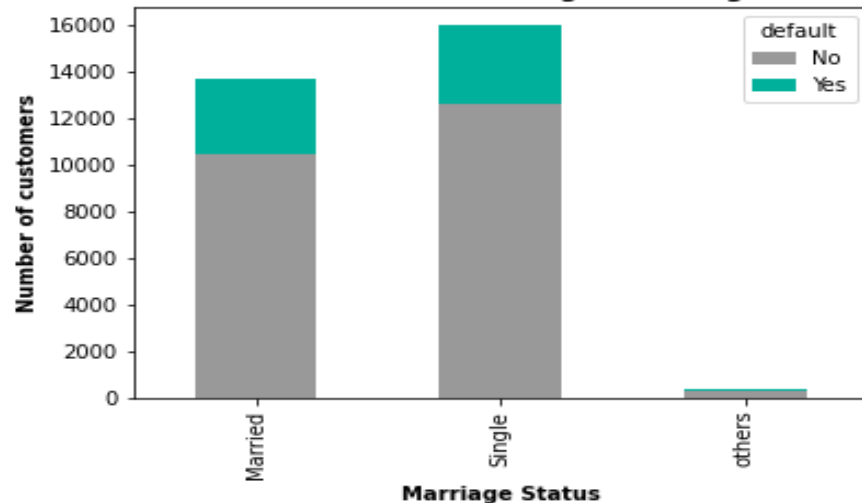
Bivariate Analysis

Number of defaults - genderwise



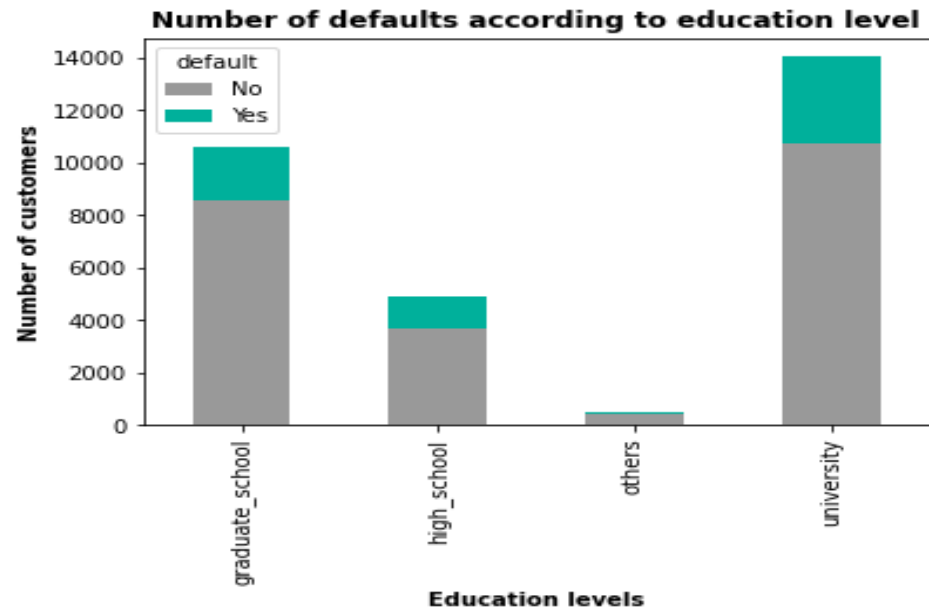
From above stacked bar chart, it can easily be understood that female customers are more than male customers in taking credit cards. Also, if we consider percentages of male and female customers who default in next month, then we see that female customers are more likely to default compared to male customers.

Number of defaults according to marriage status

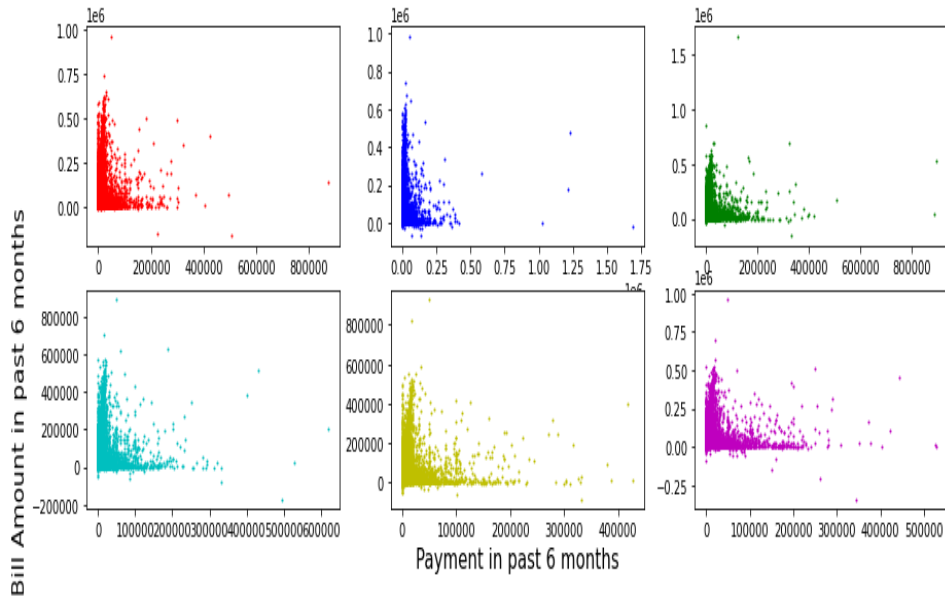


As we can see the same trend in marital status. There are more single customers than married ones, who are taking credit cards, and the defaulter percentage is also higher in single customers.

Continued...

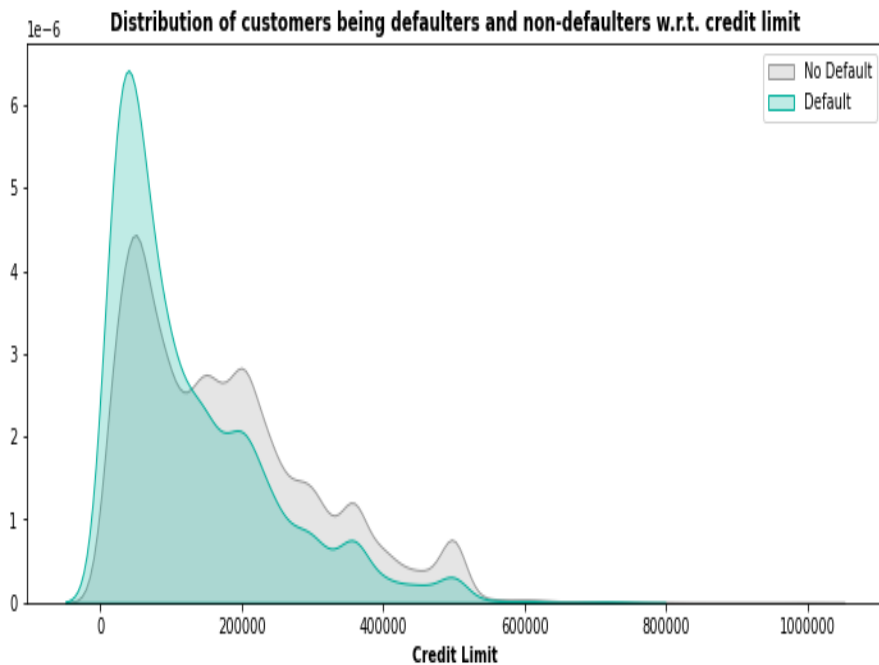


With respect to education level, more educated people are more likely to default in next month.
Strange but true!

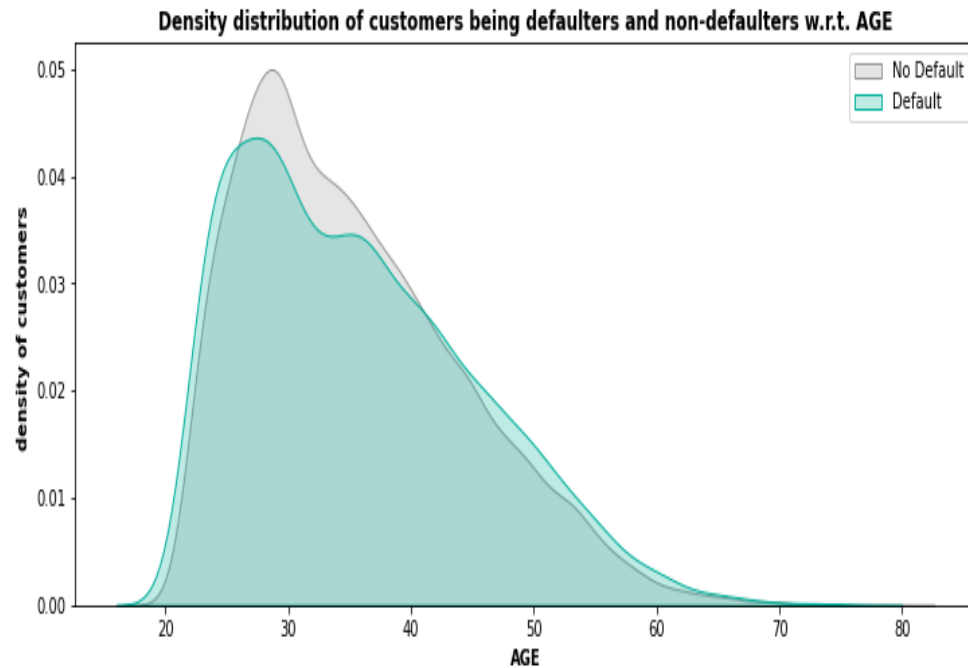


Above plot indicates that there is higher proportion of clients for whom the bill amount is high but payment done against the same is very low. This we can infer since maximum number of datapoints are closely packed along the Y-axis near to 0 on X-axis.

Continued...

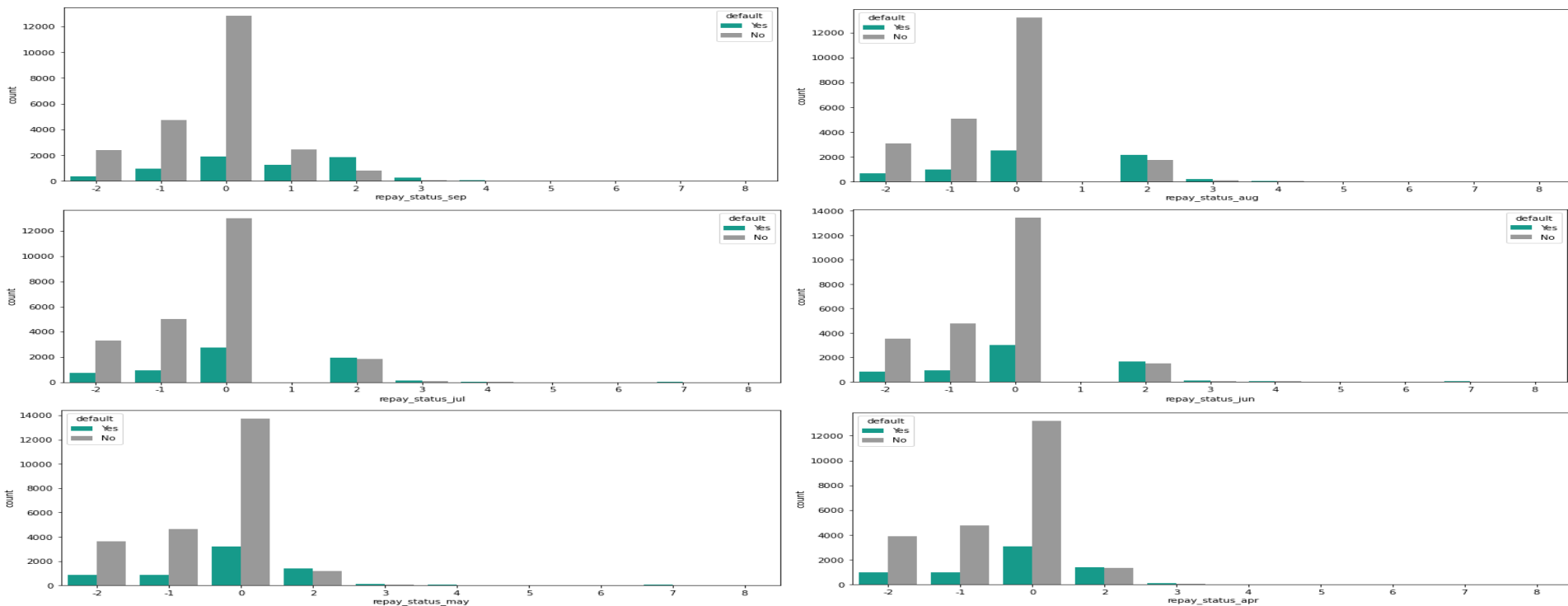


It is observed that defaulter's density is higher for the credit limit between 10,000 to 2,00,000.



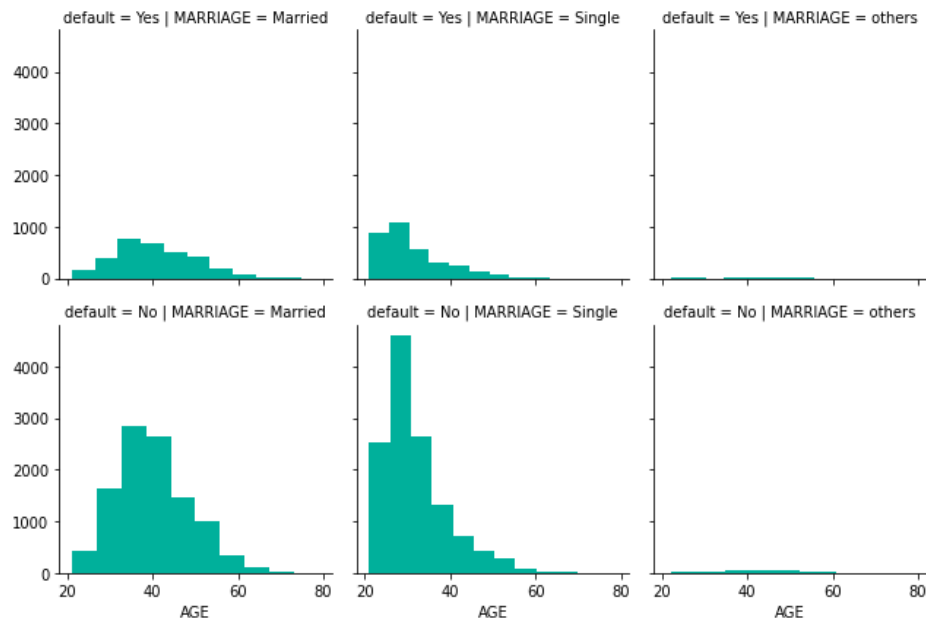
In the age group of 22 to 45 years, there is a higher chance that customer will default in the next month payment.

Defaulters w.r.t. repay status of past 6 months

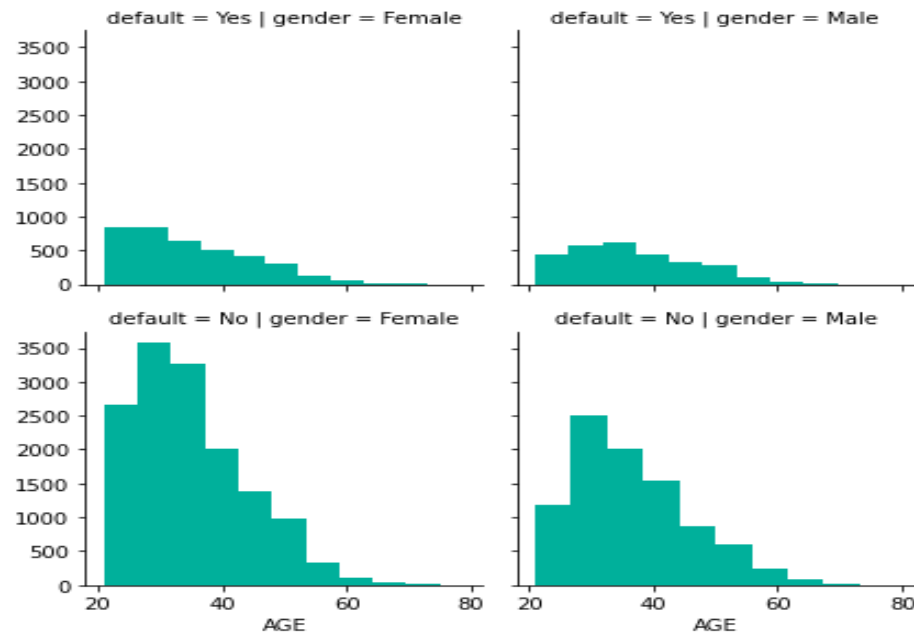


- From April to August, there are almost no customers with payment delay by one month, but there is sudden growth is seen in the month of September for customers delaying by one month.
- Also, defaulters have increased in September with one and two months payment delay.
- Those who are repaying minimum amount in all the previous months are more likely to default in the next month's payment.

Multivariable Analysis

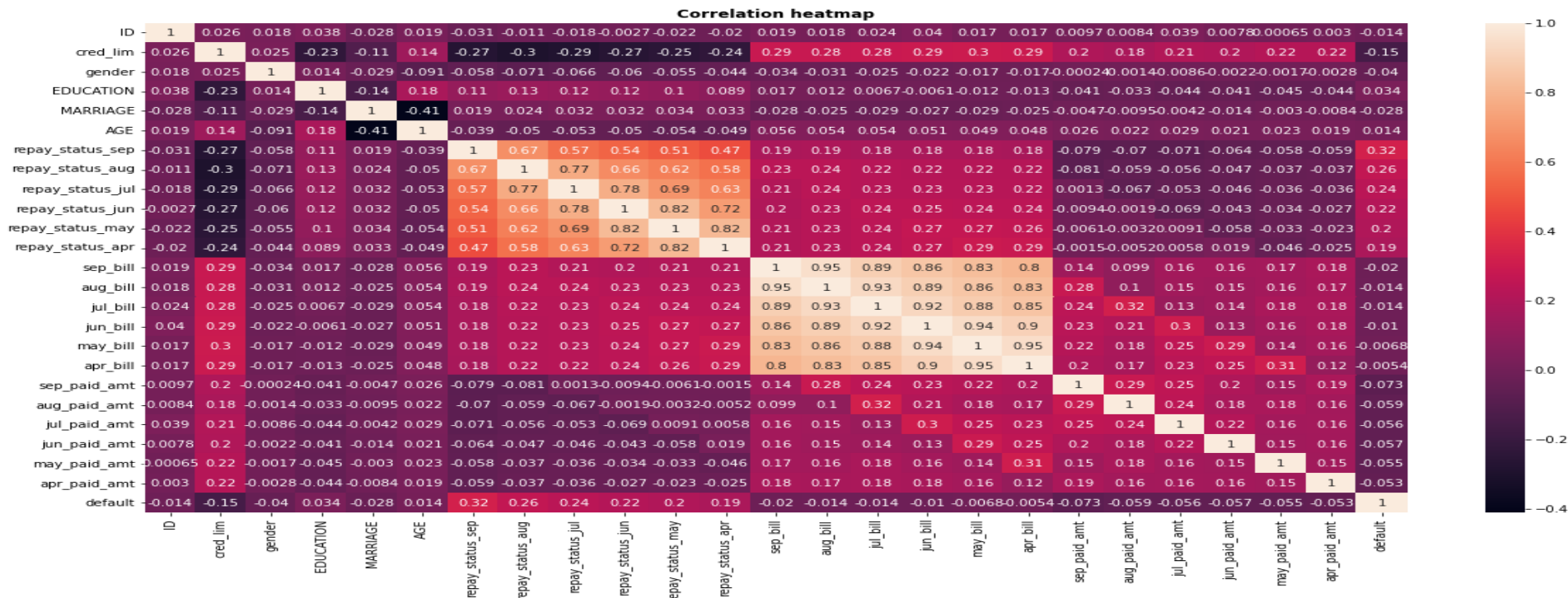


From above plot we can infer that married customers between age bracket of 30 to 50 and unmarried customers of age 20 to 30 tend to default payment with unmarried customers higher probability to default payment. Hence we can include MARRIAGE feature of customers to find probability of defaulting the payment next month.



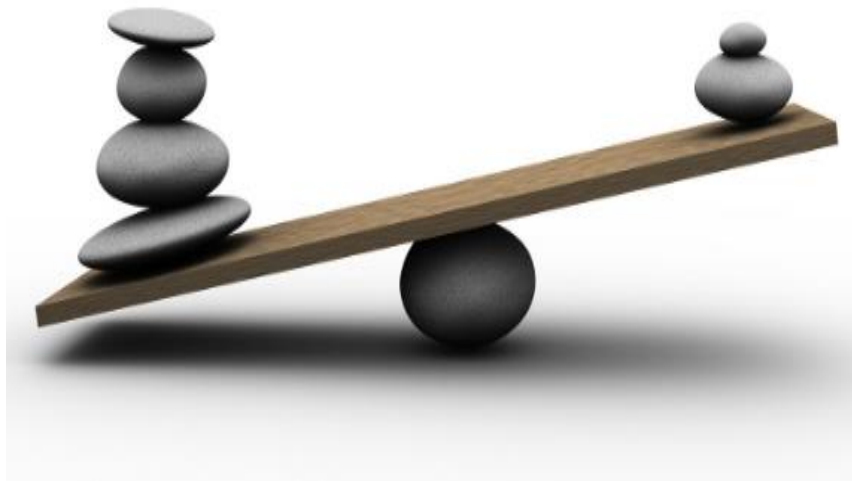
It can be seen that females of age group 20-30 have very high tendency to default payment compared to males in all age brackets. Hence we can keep the 'gender' column of customers to predict probability of defaulting payment.

Correlation among features



We can see very high correlation among bill amounts for the past 6 months and also in repay status of the all 6 months. We have removed this correlation by taking total bill amount as one column, where we have added all the bill amounts for past 6 months.

Handling Imbalanced Data



In our data set we have Imbalanced Data Distribution in our dependent variable, it generally happens when observations in one of the class are much higher i.e not defaulter or lower than the other classes i.e defaulter.

As Machine Learning algorithms tend to increase accuracy by reducing the error, they do not consider the class distribution.

SMOTE (Synthetic Minority Oversampling Technique) – Oversampling is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them.

Feature Engineering

As we saw in the correlation heatmap, there is a high correlation among billed amount of past 6 months, as well as repay status of the creditors. So, we have created new columns such as –

- `total_bill` – sum of bills from April to September
- `paid_total` – sum of amount paid from April to September
- `Dues` – difference between `total_bill` and `paid_total`

One Hot Encoding :

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.

Here we have performed one hot encoding on following columns –

| | | | |
|--------------------|--------------------|--------------------|--------------------|
| 'EDUCATION' | 'MARRIAGE' | 'gender' | 'repay_status_sep' |
| 'repay_status_aug' | 'repay_status_jul' | 'repay_status_jun' | 'repay_status_may' |
| 'repay_status_apr' | | | |

Train Test Split and standardization

```
#standardise the x value by using satandardscaler  
scaler = StandardScaler()  
X = scaler.fit_transform(X)
```

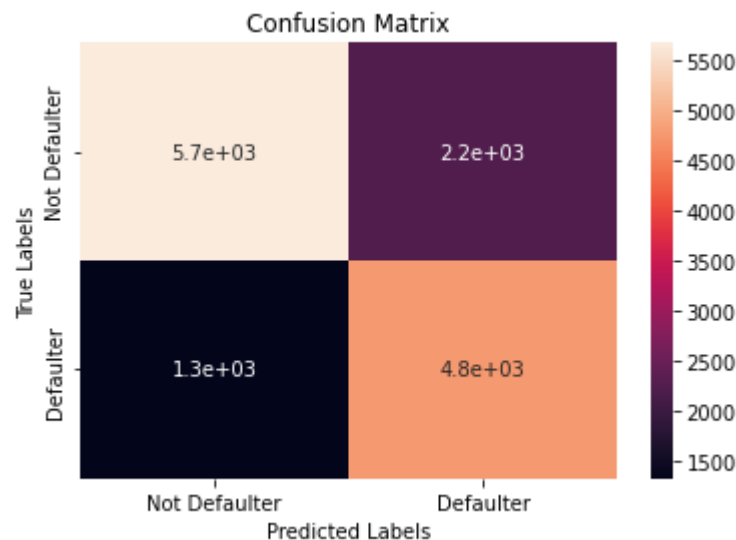
```
X_train, X_test, y_train, y_test = train_test_split( X,y , test_size = 0.3, random_state = 42)  
print(X_train.shape)  
print(X_test.shape)
```

```
(32709, 74)
```

```
(14019, 74)
```

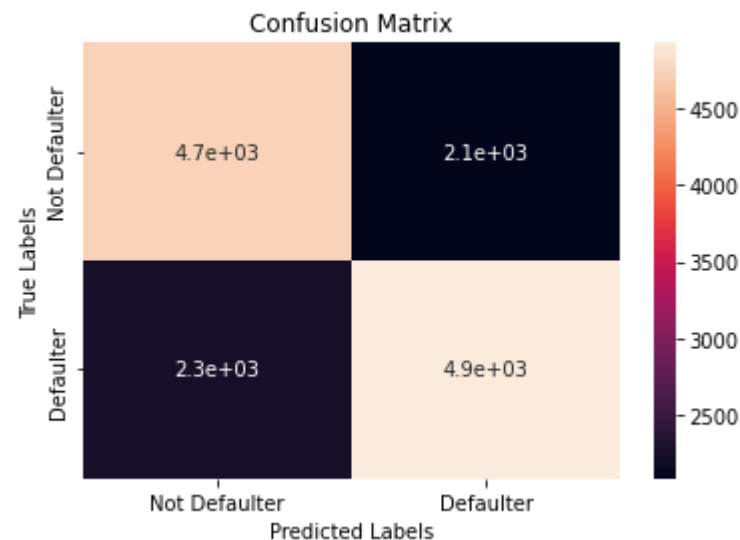
Models Implemented

1. Logistic Regression Model



The accuracy on test data is 74.57022612169199
The precision on test data is 68.07812945537496
The recall on test data is 78.26585805605639

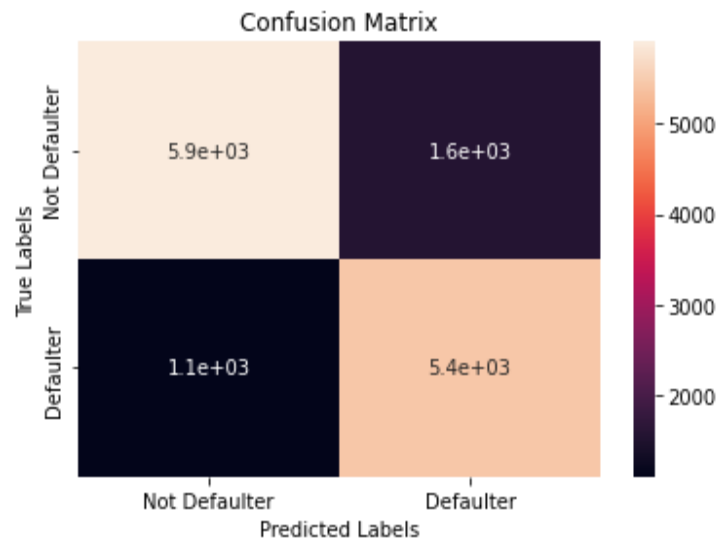
2. Decision Tree Classifier Model



The accuracy on test data is 68.99921535059562
The precision on test data is 70.25948103792416
The recall on test data is 68.55870895937673

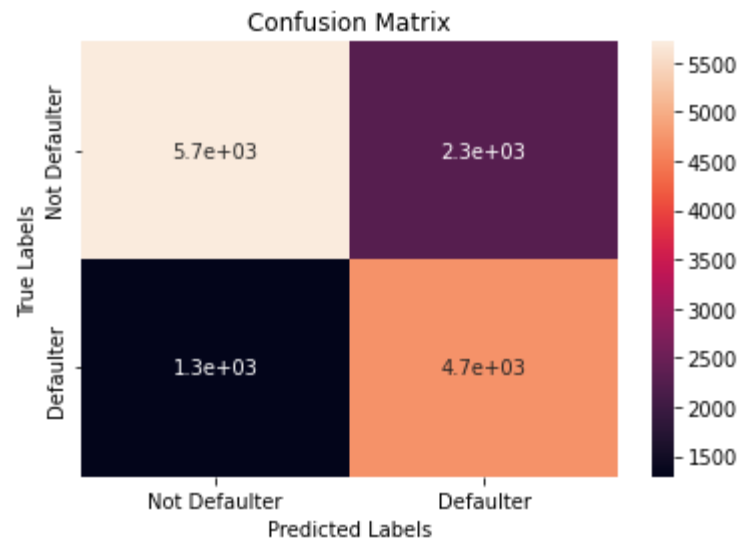
Models Implemented

3. Random Forest Classifier Model



The accuracy on test data is 80.98295170839575
The precision on test data is 77.68748217850015
The recall on test data is 83.19083969465649

4. Support Vector Machine (SVM)



The accuracy on test data is 74.4632284756402
The precision on test data is 67.37952666096379
The recall on test data is 78.53107344632768

Continued...

| | Name | Accuracy_train | Accuracy_test | Precision_train | Precision_test | Recall_train | Recall_test | F1_score_train | F1_score_test |
|---|------------------------------|----------------|---------------|-----------------|----------------|--------------|-------------|----------------|---------------|
| 0 | Logistic Regression | 75.236174 | 74.563093 | 68.733945 | 68.078129 | 78.996204 | 78.253032 | 73.508634 | 72.811833 |
| 1 | Decision Tree | 70.252836 | 68.999215 | 70.990826 | 70.259481 | 69.946969 | 68.558709 | 70.465032 | 69.398676 |
| 2 | Random Forest | 99.302944 | 80.676225 | 98.844037 | 77.445110 | 99.759259 | 82.817503 | 99.299539 | 80.041258 |
| 3 | Support Vector Machine (SVM) | 75.138341 | 74.463228 | 67.963303 | 67.379527 | 79.337427 | 78.531073 | 73.211227 | 72.529159 |

Let us understand these metrics of evaluation of the models -

| Metric | Definition | Meaning in this context |
|------------------|--|---|
| Accuracy | The proportion of the total number of predictions that are correct | Overall how often the model predicts correctly defaulters and non-defaulters |
| Precision | The proportion of positive predictions that are actually correct | When the model predicts default, how often is correct? |
| Recall | The proportion of positive observed values correctly predicted as such | The proportion of actual defaulters that the model will correctly predict as such |

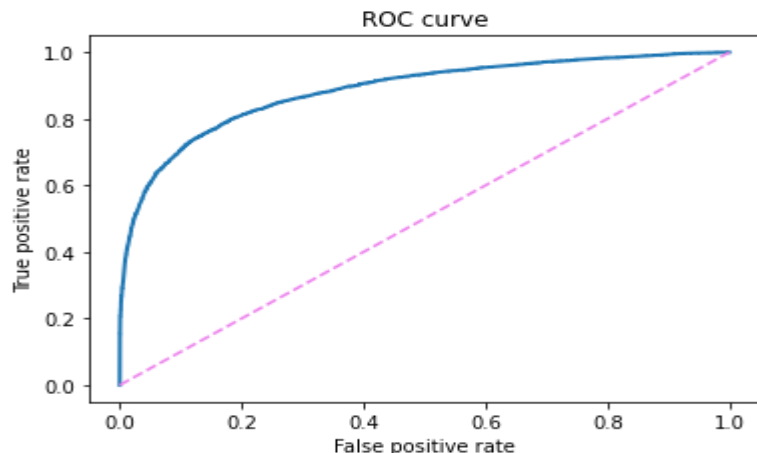
Model Selection

Which Metric should we use?

- **False Positive:** A person who will pay, predicted as defaulters
- **False Negative:** A person who default, predicted as payer

In this case, false negatives are worse ➡ look for a better recall

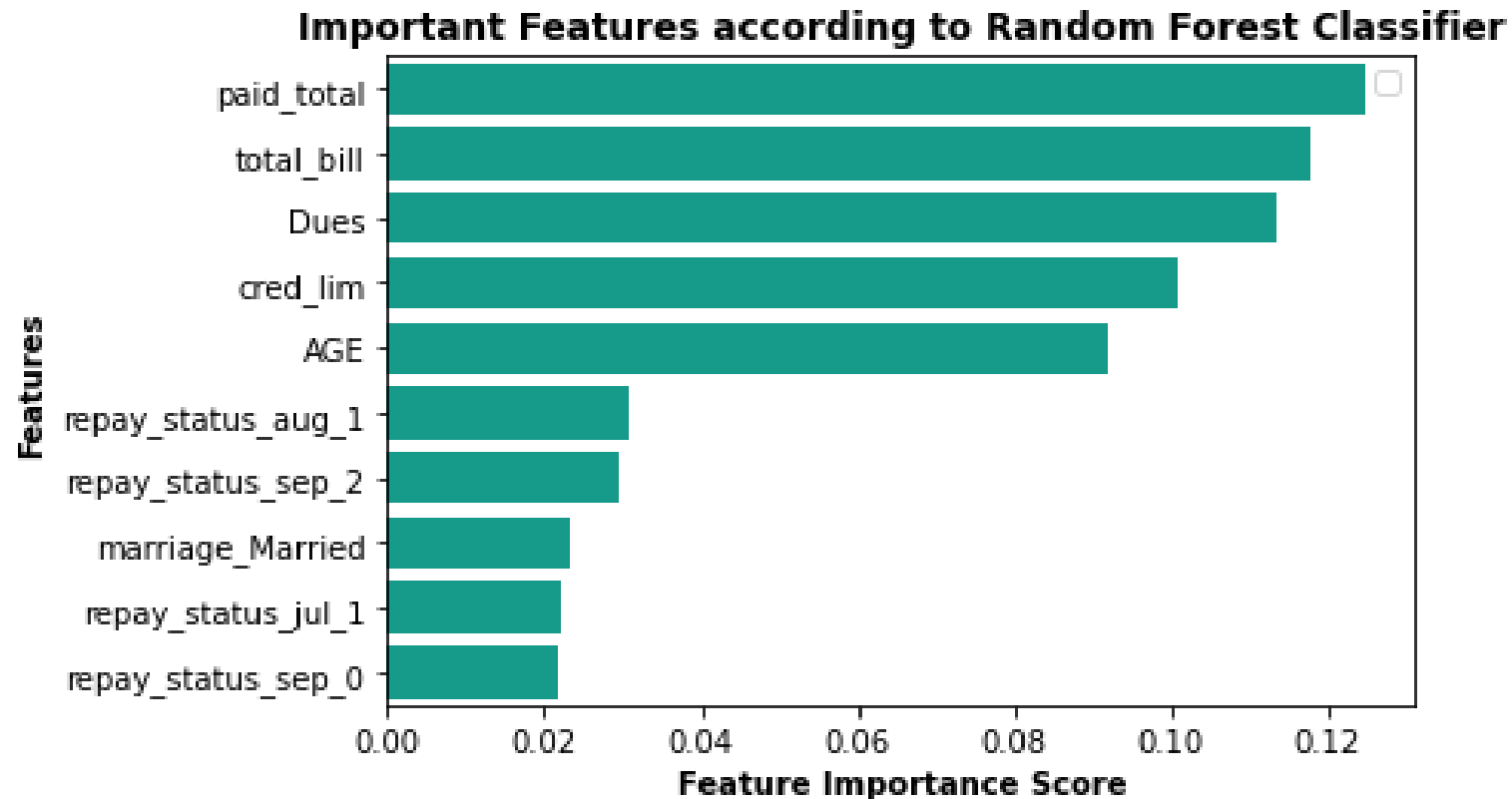
Among all 4 models which we tried to build here to predict defaulters correctly, Random Forest is the best and robust model with 82% recall score, which is very good.



We're calculating the area between the blue curved line and pink dotted line. This area is a number between 0 and 1, zero meaning the model predicted all of the data incorrectly, and one meaning the model predicted all of the data correctly. Our model is pretty good at 0.8851.

ROC AUC score = 0.8851132139717709

Feature Importance



Conclusions

- According to RandomForest model, features like paid_total, total_bill, dues, cred_lim, age, repay_status_sep_2, repay_status_aug_1 and marriage status as married are found to be most important features to predict future defaulters.
- If the credit card holder has paid minimum credit amount for past 6 months, also for every month if the dues of that customers are increasing, then it is obvious that the customer will have default for sure!
- While this age and limit_bal are the other two important predictors as we discussed in the data preparation part. This also makes sense because as one gets older, one is more likely to accumulate more resource and cares more about his reputations, which makes credit default less likely. Also, if one and one's family get more given credits, the person is more likely to live in a wealthier environment which also makes credit default less likely.
- Further, repay status is also playing a vital role in predicting future defaulters. If the creditor is not repaying for the past one or more months, he is more likely to default in upcoming months.

Continued...

At the end of the day, having the ability to predict 82% (recall score) of potential defaults would save a-lot of money on credit card charge-offs. Obviously, real-world application is more nuanced, but this modeling process is a step in the right direction.

Reflection

One important way that I think the predictive model could be improved is the enhancement of the data source. There are still lots of information that the data didn't cover. For example, the current economic conditions of one person, like incomes and jobs of creditors; the amount of non-liquid assets owned by the creditors, and so on. This imperfection of the dataset determined that the model would lose some predicted power and are facing more uncertainty.

Challenges

- ❑ Deciding on which visualizations to use while doing EDA.
- ❑ Dealing with imbalanced data was a task, as we do not want to lose important data if we have used underfitting.
- ❑ Deciding on classification models to be used and its hyperparameter tuning to prevent overfitting.

Thank You