# Capstone Project-3

## Credit Card Default Prediction

# Agenda

- 1) Introduction
- 2) Problem Statement
- 3) Overview of the dataset
- 4) EDA and Feature engineering
- 5) Implementation of model
- 6) Model Performance
- 7) Hyperparameters tuning
- 8) Conclusions

## Indroduction

Credit Card risk plays a major role in the banking/finance 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. As such data analytics can provide solutions to tackle the current phenomenon and management credit risks. This project/paper provides a performance evaluation of credit card default prediction.

# 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. We can use the KS-chart to evaluate which customers will default on their credit card payments

## Overview of the dataset

This dataset consists of 30000 observations that represent distinct credit card clients. Each observation has 24 attributes that contain information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan.

1. ID: ID of each client, categorical variable

2. LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)

3. SEX: Gender, categorical variable (1=male, 2=female)

4. EDUCATION: level of education, categorical variable (1=graduate school, 2=university, 3=high school, 4=others)

5. MARRIAGE: Marital status, categorical variable (1=married, 2=single, 3=others)

6. AGE: Age in years, numerical variable

7. 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)

8. PAY_2: Repayment status in August 2005 (same scale as before)

9. PAY_3: Repayment status in July

10. PAY_4: Repayment status in June

11. PAY_5: Repayment status in May

12. PAY_6: Repayment status in April 2005

# Overview of dataset cont.

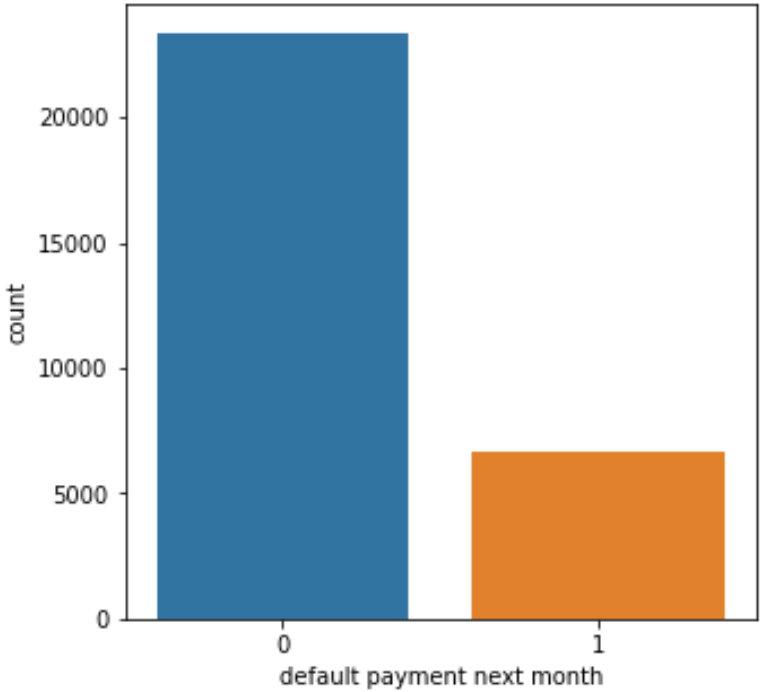
13. BILL_AMT1: Amount of bill statement in September
14. BILL_AMT2: Amount of bill statement in August
15. BILL_AMT3: Amount of bill statement in July
16. BILL_AMT4: Amount of bill statement in June
17. BILL_AMT5: Amount of bill statement in May
18. BILL_AMT6: Amount of bill statement in April
19. PAY_AMT1: Amount of previous payment in September
20. PAY_AMT2: Amount of previous payment in August
21. PAY_AMT3: Amount of previous payment in July
22. PAY_AMT4: Amount of previous payment in June
23. PAY_AMT5: Amount of previous payment in May
24. PAY_AMT6: Amount of previous payment in April

# EDA and Feature engineering

```
data['default payment next month'].value_counts()
```

```
0    23364
1     6636
Name: default payment next month, dtype: int64
```

* There 6636 people who defaulted and 23364 who have not defaulted

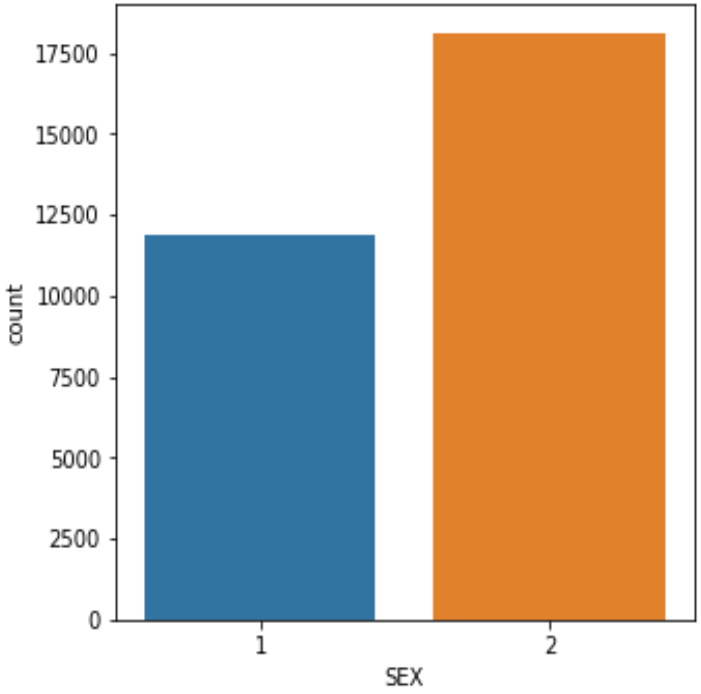


# What gender tells us?

There are 11888 males and 18112 females
We can say that there are 60% females
And 40% males

```
data['SEX'].value_counts()
```

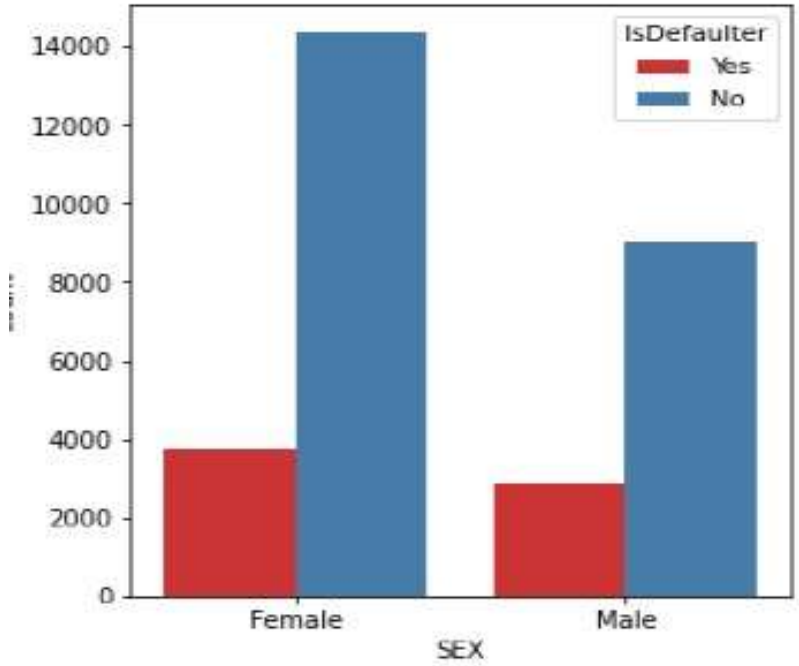
```
2    18112
1    11888
Name: SEX, dtype: int64
```



# Cont.

- Even though there are less male  then
- Female, the default count for male
- Are higher then female

- *Male has hight default count
- *Female has low default count

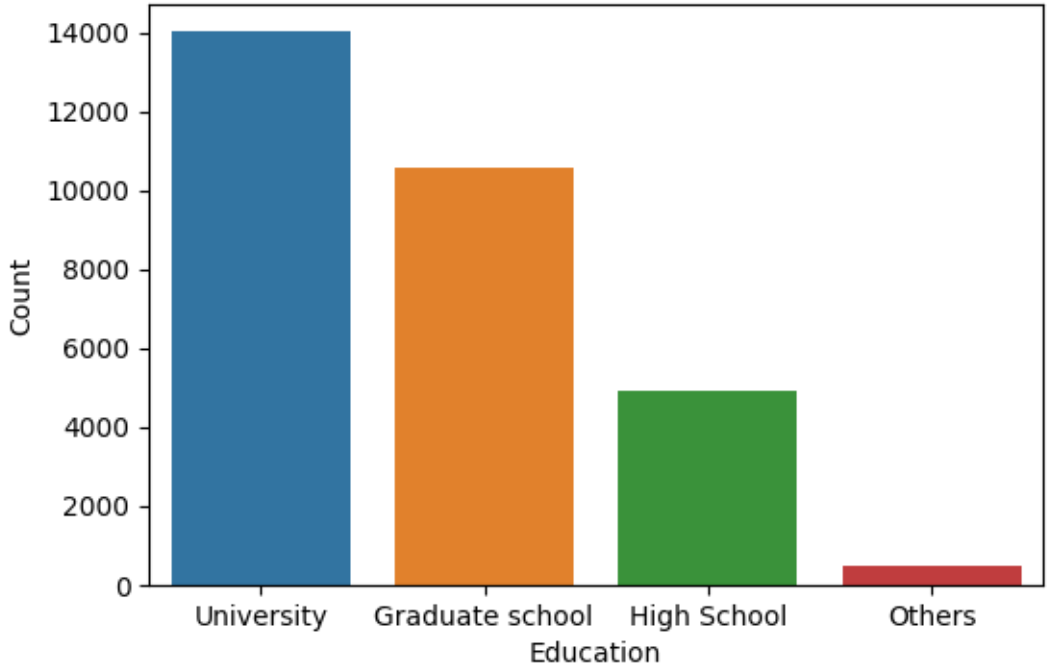


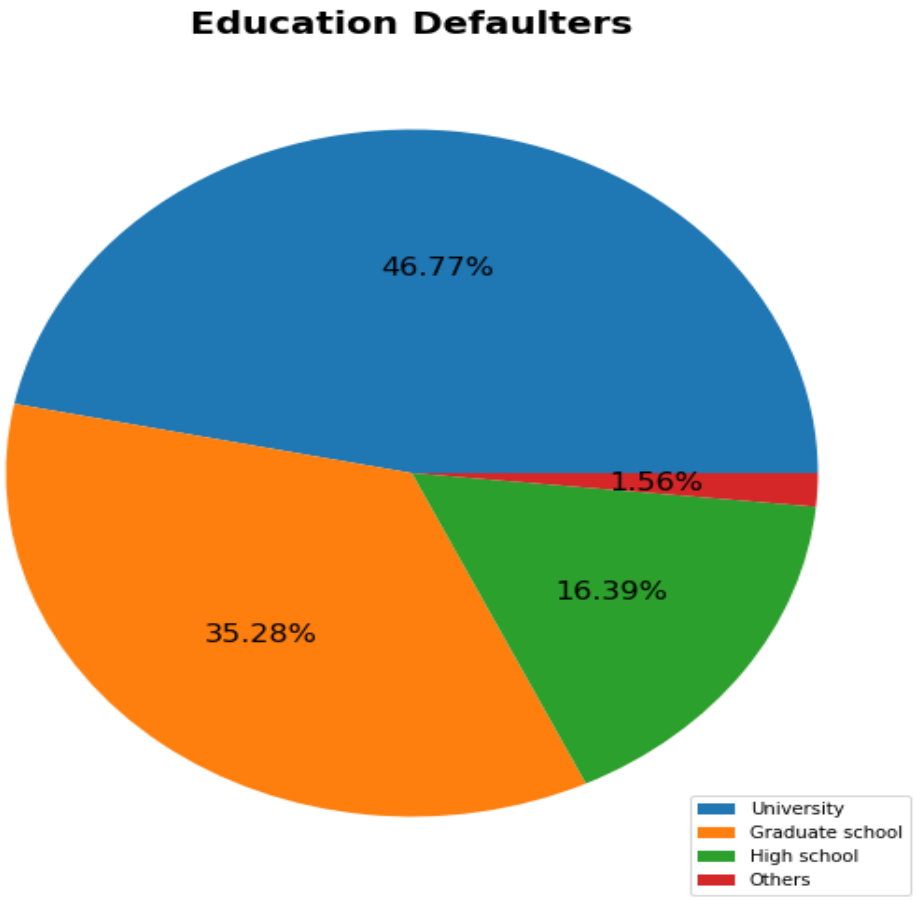
# What Education can tell us?

```
data['EDUCATION'].value_counts()
```

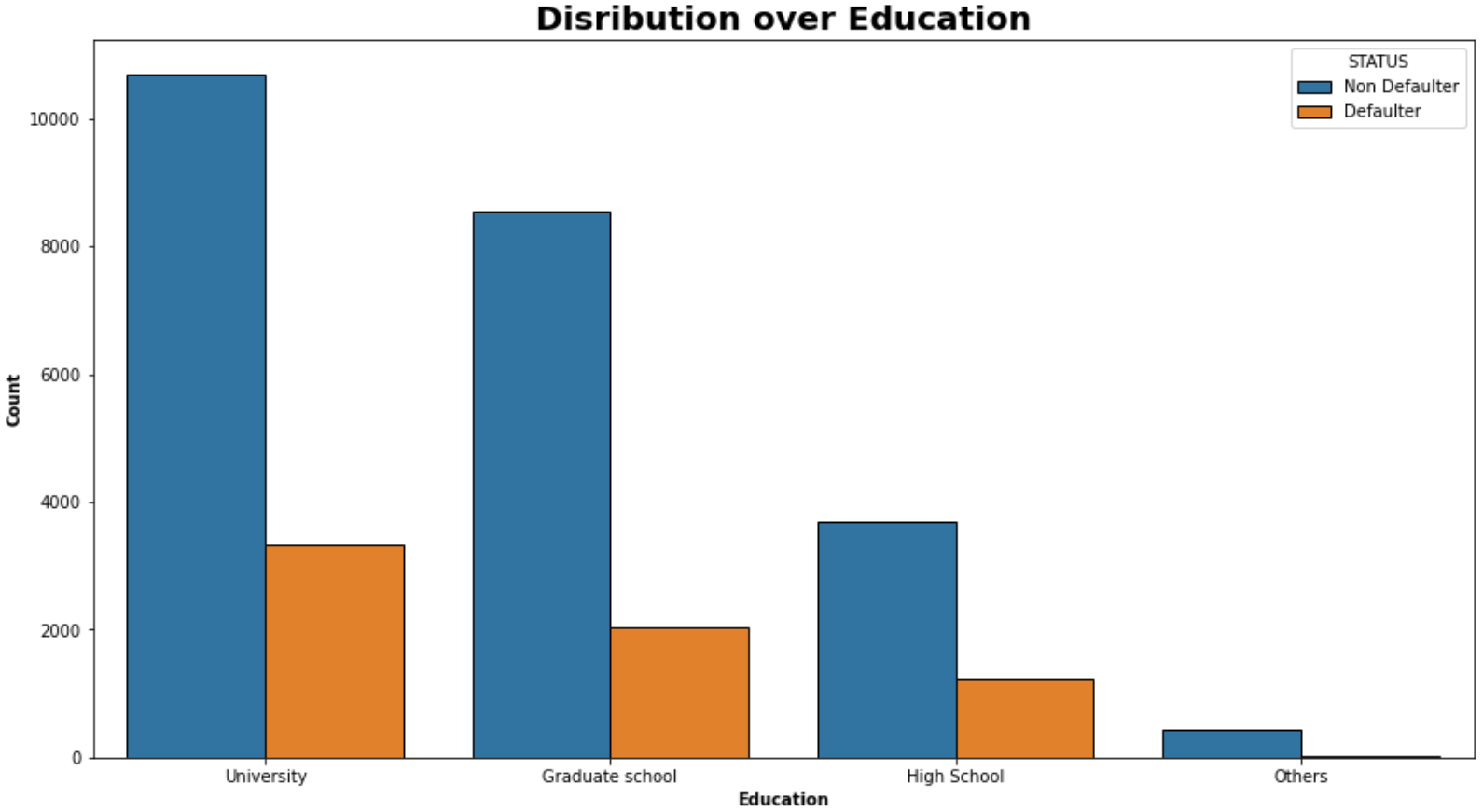
```
2    14030
1    10585
3     4917
5      280
4      123
6       51
0       14
Name: EDUCATION, dtype: int64
```

University-14030
Graduate School-10585
High School-4917
Others-3454




Education Defaulters
46.77%
35.28%
16.39%
1.56%
University
Graduate school
High school
Others

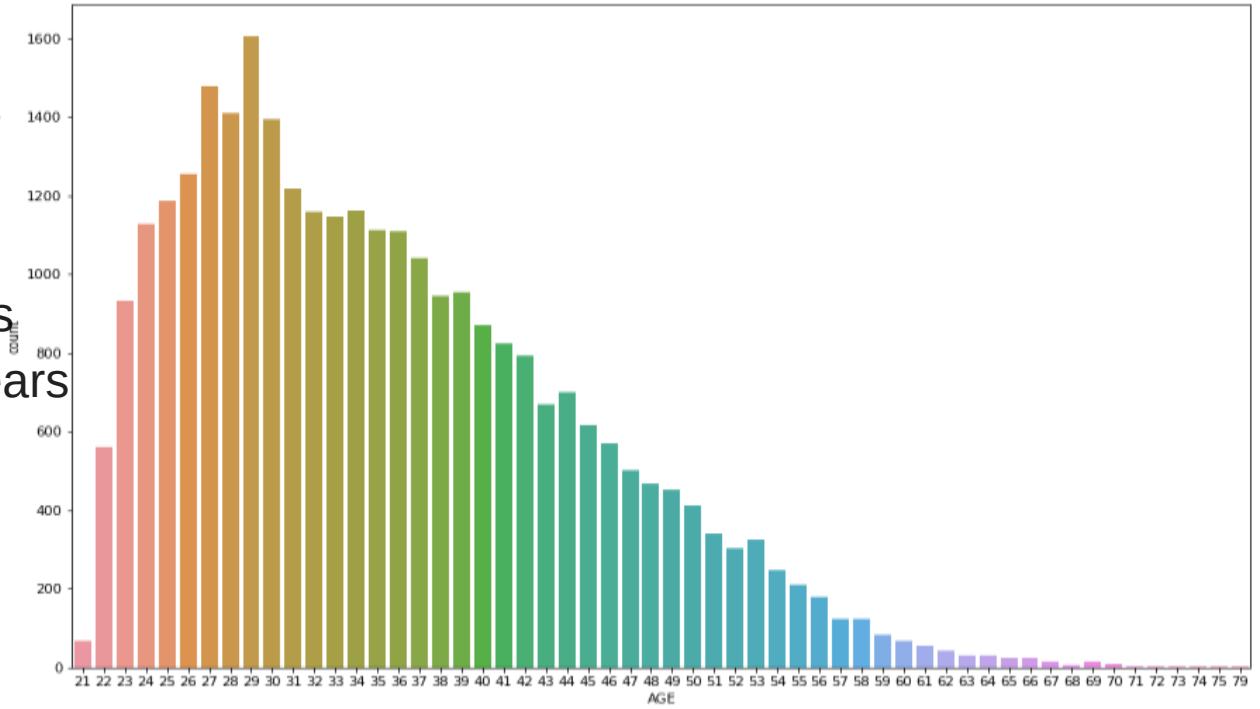
# Distribution of defaulters in education



# What age tells us?

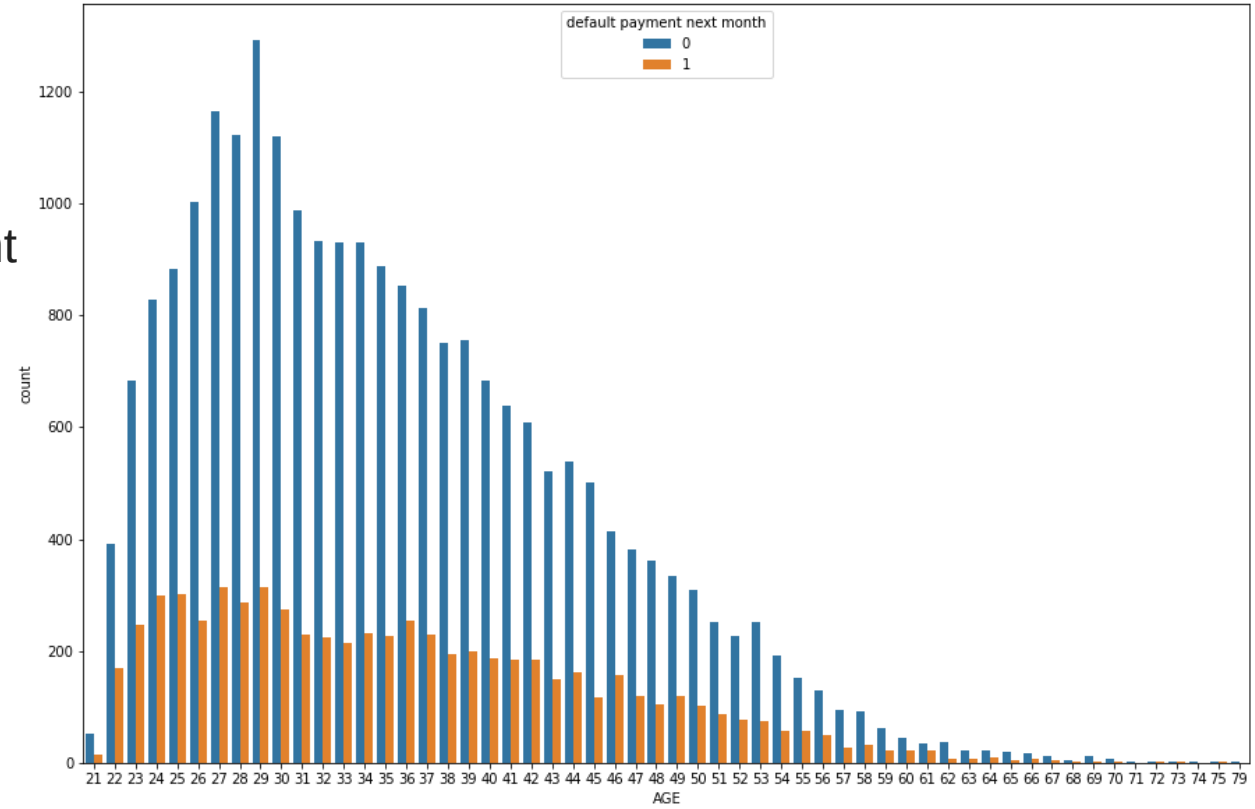
Most of the people who
Have credit cards range
Between 23-40 years

And few people possess
Credit card above 50 years

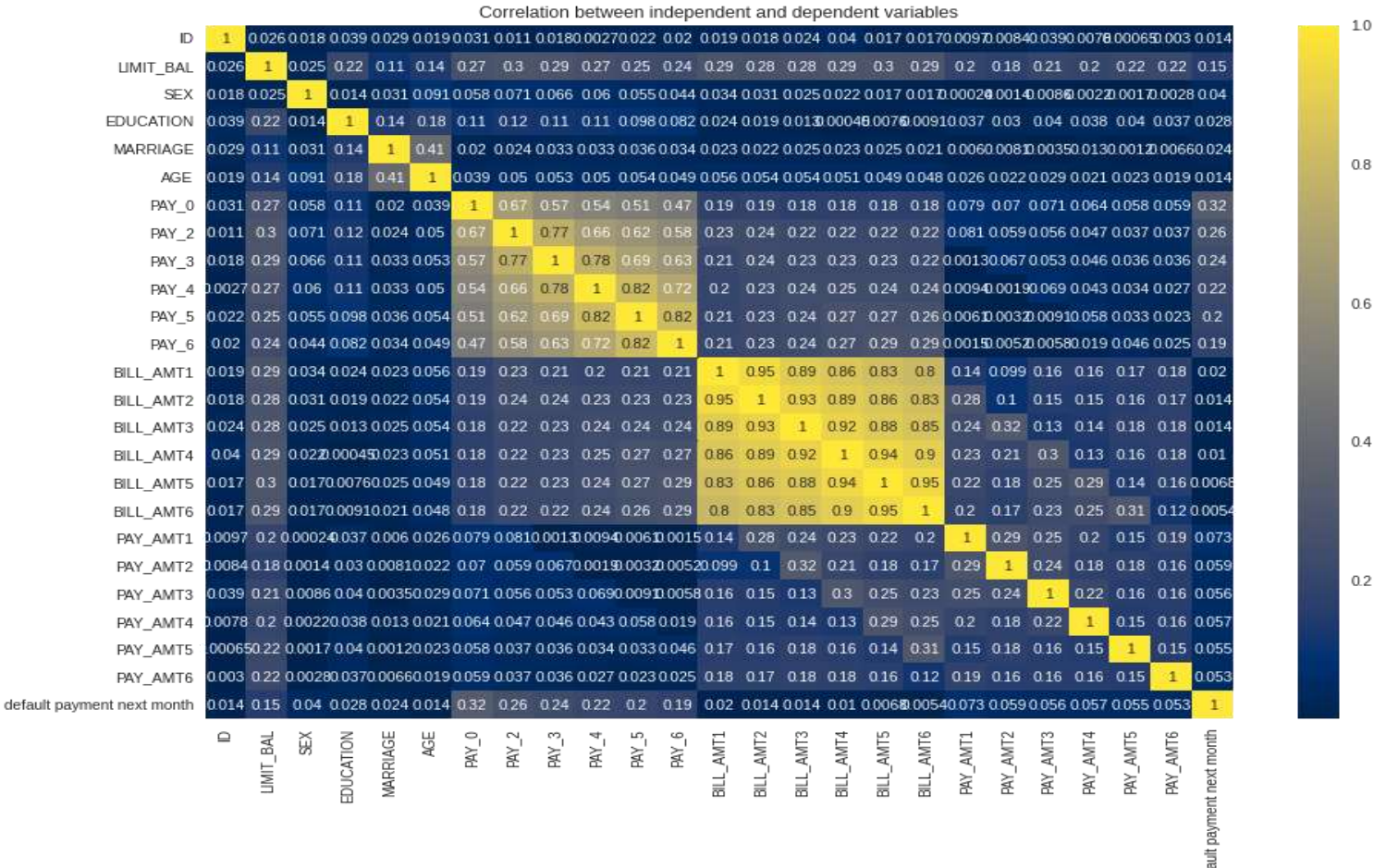


# Age distribution of both defaulters and non defaulters

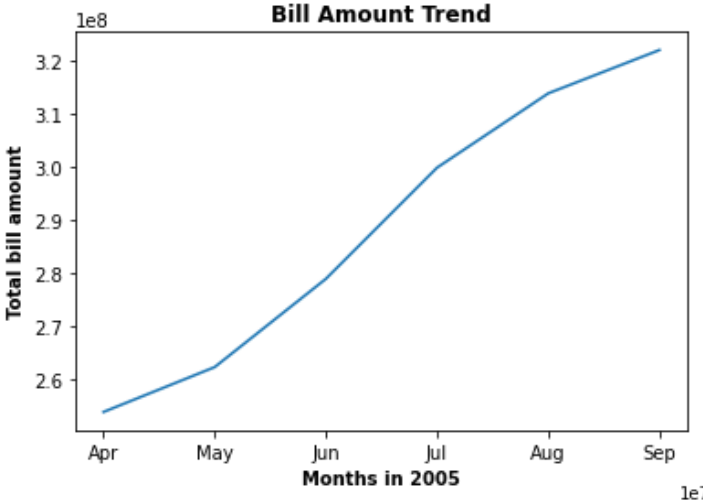
We can observe that
People who are in
Between ages 22-36
Have high default count

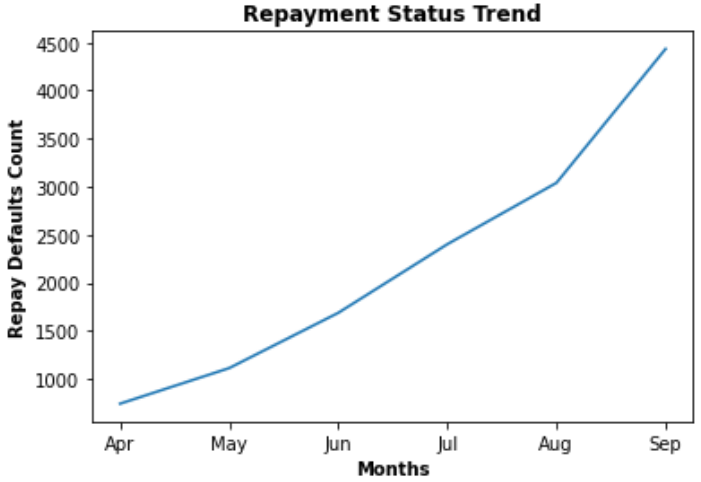


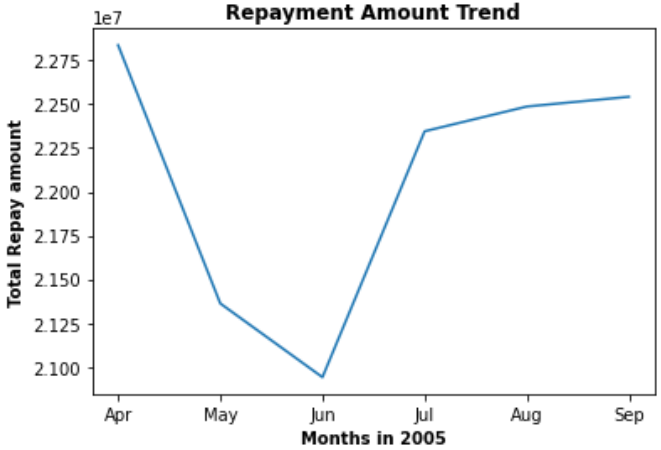
# Correlation between dependent and independent variables



# Trends


Bill Amount Trend
1e8
3.2
3.1
3.0
2.9
2.8
2.7
2.6
Total bill amount
Apr
May
Jun
Jul
Aug
Sep
Months in 2005


Repayment Status Trend
4500
4000
3500
3000
2500
2000
1500
1000
Repay Defaults Count
Apr
May
Jun
Jul
Aug
Sep
Months


Repayment Amount Trend
1e7
2.275
2.250
2.225
2.200
2.175
2.150
2.125
2.100
Total Repay amount
Apr
May
Jun
Jul
Aug
Sep
Months in 2005

# Models Implemented

There are 4 models implemented

- Decision Tree 83%
- Support Vector Machine 78%
- Logistic Regression 78%
- K Nearest Neighbour 74%

Clearly we can see that Decision tree has the 83% accuracy which is the highest of all these models

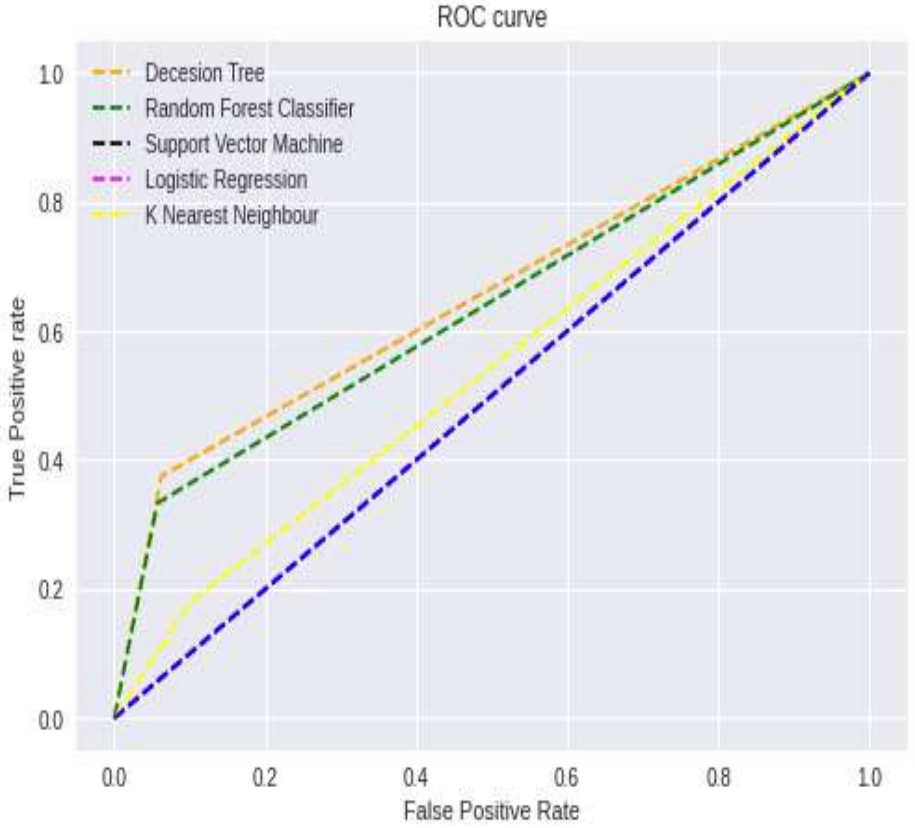
# Hyperparameter tuning

With decision tree, we were able to achieve 83% accuracy on our training data and 88.66 % accuracy on our testing data, we can say our model is overfitting The algorithm may overfit if we use this model on unknown data As a result, we must overcome this challenge, and we are doing so by implementing Hyperparameter tuning on decision tree classifier

We got 82% accuracy on our training data and testing data after implementing Hyperparameter tuning on decision tree classifier

# ROC curve

Receiver Operating Characteristic summarizes the model's performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1-specificity).
For plotting ROC, it is advisable to assume p > 0.5 since we are more concerned about success rate.



# Conclusions

* Males are likely to default on the next payments then women
* Defaults are higher in University and Graduate level in terms of education
* Ages from 22-36 has the highest defaults in all ages
*From all baseline model,  decision tree shows highest test accuracy
* Baseline model of decision tree shows huge difference in train and test accuracy which shows overfitting.
* After cross validation and hyperparameter tunning, Decision Tree shows highest test accuracy score of 83%.

Thank You