**Taiwanese Credit Card Capstone Final Report**

[GitHub Project Link](https://github.com/CeyhunSahinkaya/Credit-Card-Payment-Prediction)

# Table of Contents

[**Table of Contents**](#_heading=h.30j0zll) **1**

[**Problem Statement**](#_heading=h.1fob9te) **2**

[**Data Collection**](#_heading=h.3dy6vkm) **3**

[**Data Wrangling**](#_heading=h.17dp8vu) **4**

[**Exploratory Data Analysis**](#_heading=h.1ksv4uv) **5**

Credit Card Limit vs Default Payment **6**

Marriage, Sex vs Credit Card Limit **7**

Sex vs Default Payment **8**

Marriage, Sex vs Age **10**

**Model** [**Prediction Analysis**](#_heading=h.2xcytpi) **15**

Logistic Regression Classifier  **16**

Confusion Matrix  **17**

Decision Tree Classifier **17**

Feature Importance **18**

Confusion Matrix **19**

Random Forest Classifier  **20**

Feature Importance **20**

Confusion Matrix **21**

Xgboost Classifier **22**

Feature Importance **22**

Confusion Matrix **23**

[**Conclusion**](#_heading=h.2u6wntf) **24**

# Problem Statement

How do the banks decide if a person will pay their next credit card payment? How do the banks assess the risk of giving out loans to individuals?

In the banking and investment world, this is the question that needs to be answered countless times in a day. It is critical for the banks to know the answer for each individual that they work with to be able to make financial gain (or to not have financial loss). Not only do they have to answer it right, also they have to answer it in a way that would be applied to many cases as quickly as possible.

Today, banks use data science and machine learning to tackle this problem. This capstone project will show and demonstrate how machine learning can be used in the finance and banking world.

We will look at the dataset in Taiwan from April 2005 to September 2005 that contains information about default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients. Then, we will create a model that predicts if the individual will default on card payment based on the dataset. By doing this, we will see which features are more important for the banks while making the decision whether the person will make the next payment or not. The project will include data visualization and exploratory data analysis and it will conclude with some metrics that measure the success of the predicted model.

# 

# Data Collection

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. There is no missing data on the dataset, the dataset has 3000 rows and 25 categories(columns).

## Content

There are 25 variables:

* 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 (-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)
* PAY\_2: Repayment status in August, 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)

The dataset is available at the below:

<https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>

# 

# Data Wrangling

### **Marriage Column**

There are 3 categories documented for the MARRIAGE column.

MARRIAGE: Marital status (1=married, 2=single, 3=others)

However, if we look into the columns, there is another category as **0** which is not documented. We can add this **0** to **3=others** since it will make more sense for our analysis.

### **Repayment Status Columns**

Repayment status columns (PAY\_0, PAY\_2, PAY\_3, PAY\_4, PAY\_5, PAY\_6) includes below categories:

(-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).

But, if we take a closer look at the columns, we can see that there are also -2 and 0 as categories and there is a good amount of data under these categories. ( 2759 for -2 and 14737 for 0). Since there is a category for duly paid incidents, we should collect all duly paid data together, meaning ( all data in category -2,-1,0 into category 0). Then, we apply this to all repayment status columns.

## Education Column

Let's check the EDUCATION category as well. We have 6 documented and 1 undocumented categories for EDUCATION, I think we should add categories 0, 5 and 6 to 4, thus we can have separate categories for clients who went to grad school, college (university) and high school. Rest is combined to 4=others.

EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)

# Exploratory Data Analysis

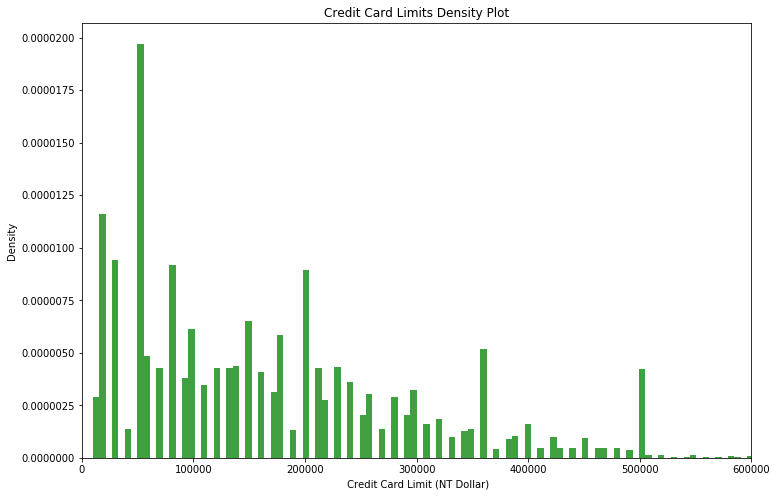
Exploratory Data Analysis (EDA) is a good way to start with your analysis and learn about the dataset. Now, we will look into some features and check how they are correlated with some other features. Also, we will create some charts and plots to visualise the data to make more sense about it. Visualization helps us to ask relevant questions to solve our problem.

## Credit Card Limit vs. Default Payment

Let's check if there is a correlation between credit card limit and Default Payment.

There are 81 different credit card limit values and most of the clients have 50,000 limits (3365 people).

We can see the density plot below:



Below plot verifies that 50,000 is the most frequent limit among the clients.It has the most density, followed by 20,000 ( 1976 people), 30,000 (1610 people) and 80,000 (1567 people)

## Marriage, Sex vs Credit Card Limit Balance

## Among married clients, male population has a bigger mean than the female population. Married population is the only population among Marriage features that male Q3 is bigger than female ones.

## Married men have the bigger max credit card limit, others (could be divorced or widowed) have the lowest credit card limit.

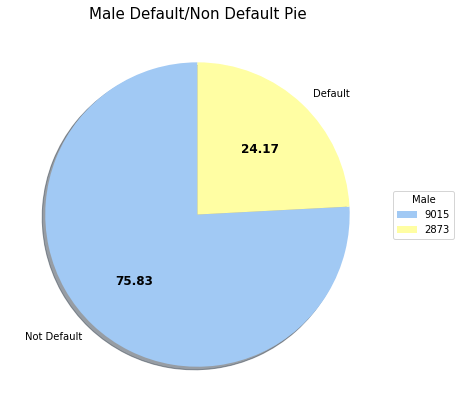
## 

## 

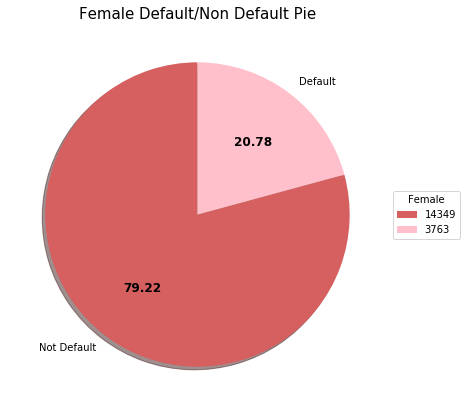
## Sex vs Default Payment

Which sex do default more payments? Is there any correlation between them? Let's answer these questions!

Let's visualize the density plot for sex grouped by default payment next month.

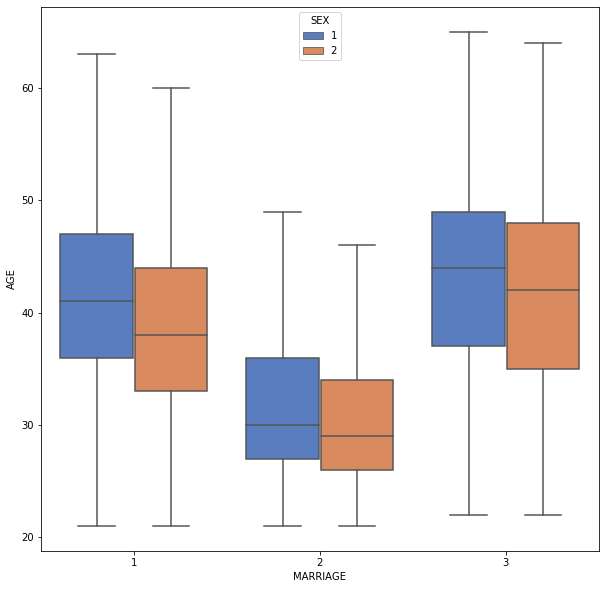


75% of the male clients make their next payment, 25% of them default the next payment.



80% of the female clients make their next payment, 20% of them default the next payment.

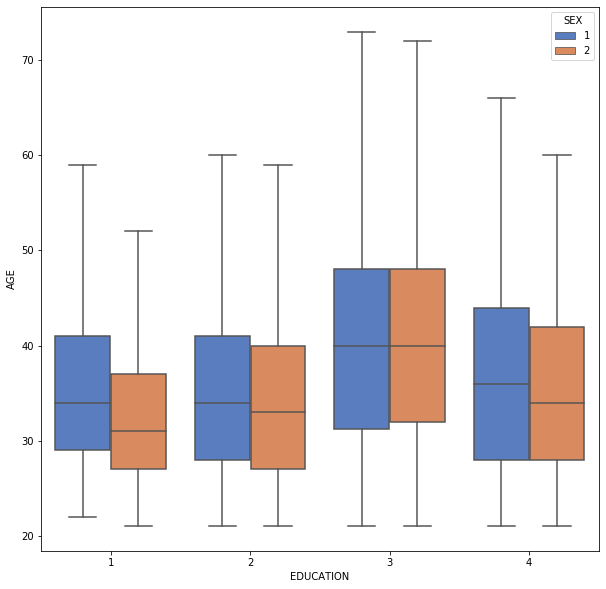
## Marriage, Sex vs Age



Others (could be divorced or widowed) have the higher mean of age as it could be guessed. Single female clients have the lowest mean age.

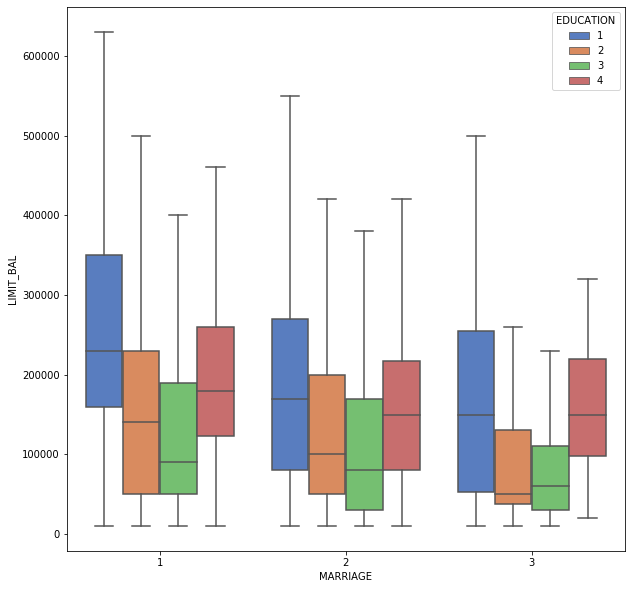
In general, the female population in all categories has lower mean age.

**Education, Sex vs Credit Card Limit Balance**

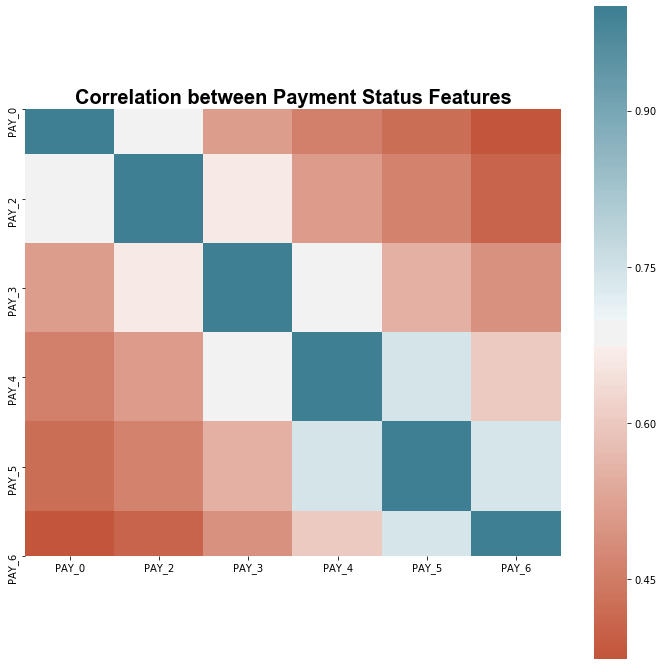
****

Lower age mean group is the grad school graduate females, and higher mean group is high school educated clients. Also, high school groups have the maximum age.

**Education, Marriage vs Credit Card Limit Balance**

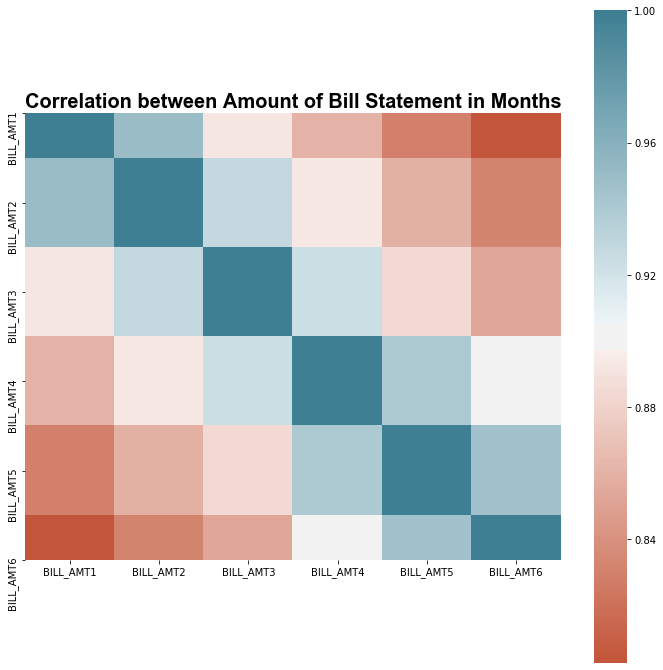
****

**Correlation Between Payment Status Features**

****

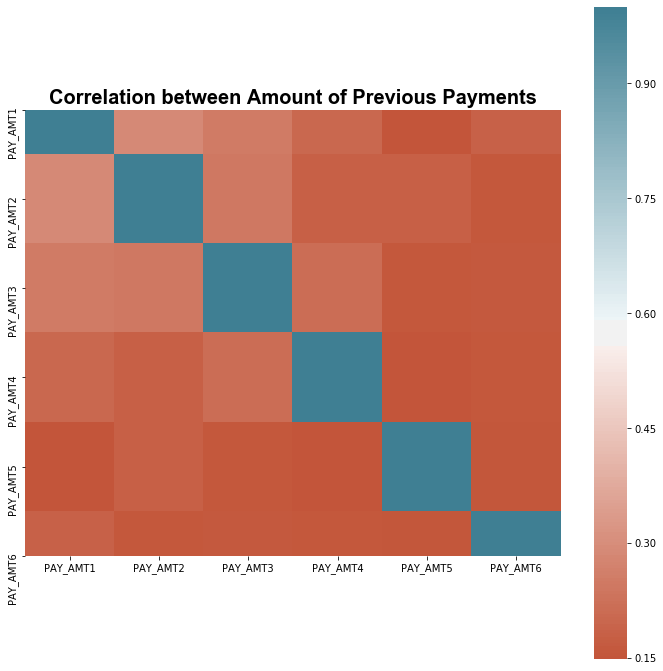
Correlation is not strong with features. Correlations decrease with the distance between months. Biggest correlations are between PAY\_5 - PAY\_4 (May-June) and PAY\_6 - PAY\_5 (April - May).

## Correlation between Amount of Bill Statement in Months



Strong correlation between consecutive months, correlation decreases between months in distance.

## Correlation between Amount of Previous Payments

****

Mostly, there is no correlation between amounts of previous payments.

# Model Prediction Analysis

Now, we will start building our model to predict whether the client will default the next payment or not by the features we worked on so far. Since the question that we are looking to answer has only two categories (default/non default), we can use binary algorithms such as Logistic regression classifier, Decision Tree classifier, Random Forest classifier or Xgboost classifier.

## Defining Features and Target Value

Before jumping the regression models, we need to define our target value and features. Our features will X and our target value will be y as per below:

X = All features except ID and default.payment.next.month columns (predictors)

y = default.payment.next.month column (target value)

**Training and Test Datasets**

We set up our datasets, now we will split our sets into training data and test data. The reason behind is that we would like to build a strong model that predicts well. By splitting our data into test and training data, we can use cross validation on our model later. We split our data as 70% training and 30% test data.

## Variance Inflation Factor (VIF)

Collinearity is when two or more variables are highly correlated with one another, in other words, duplicate information within a dataset. Ideally, features in a dataset display different information. Collinearity can cause inflation of the variance of a regression coefficient which may cause predictions with large errors. A VIF value is calculated for each feature/column of data:

After removing all columns of data where the VIF is greater than 5, the dataset is left with 14 columns:   
(LIMIT\_BAL', 'PAY\_0', 'PAY\_2', 'PAY\_3', 'PAY\_4', 'PAY\_5', 'PAY\_6', 'PAY\_AMT1', 'PAY\_AMT2', 'PAY\_AMT3', 'PAY\_AMT4', 'PAY\_AMT5', 'PAY\_AMT6', 'SEX')

## 

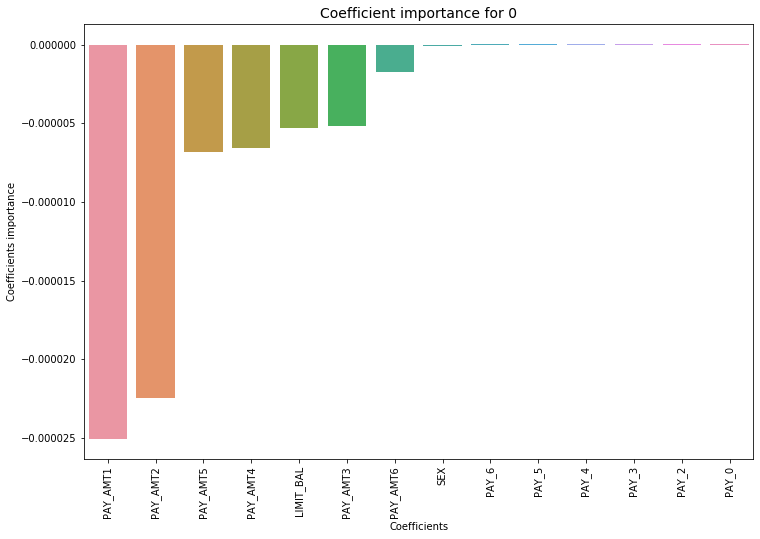
## Logistic Regression Classifier

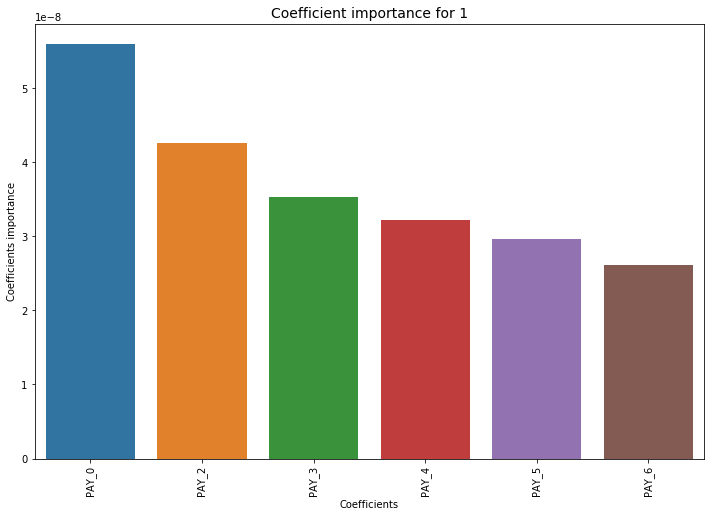
We will start a basic Logistic regression classifier on our dataset. When fitting models, we would like to ensure two things:

- We have found the best model (in terms of model parameters).

- The model is highly likely to generalize i.e. perform well on unseen data

We fit our training data on Logistic Regression Classifier, and predict the default payment. We can see the coefficients importance below;





Now, we will check our model accuracy. We will check accuracy, AUC and Confusion Matrix.

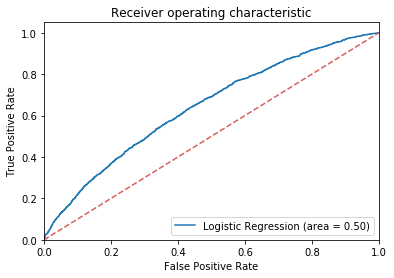
### Confusion Matrix

Accuracy of logistic regression classifier on test set: 0.78

Confusion matrix:

[[6978 0]

[2022 0]]



As we can see above, our models can’t predict 1 label (default payments). That means the dataset doesn’t fit the linear model and we need to try some other classifier for our dataset.

## Decision Tree Classifier

Now, we will build a Decision Tree Classifier and see if our data will fit better on this classifier. We use max depth as 10 while creating our classifier. Then, we train our training data and test the model on test data.

To improve our score, we will use GridSearch to test different parameters and apply the best parameters for the model. Therefore we will feed different parameters into the model and let GridSearch decide the best parameter to use in this model. The parameters can be seen below:

Max depth: between 3,4,5,6,7,8,9,10

criterion: between ‘gini' and 'entropy'

Max leaf nodes: between 5, 10, 20, 100

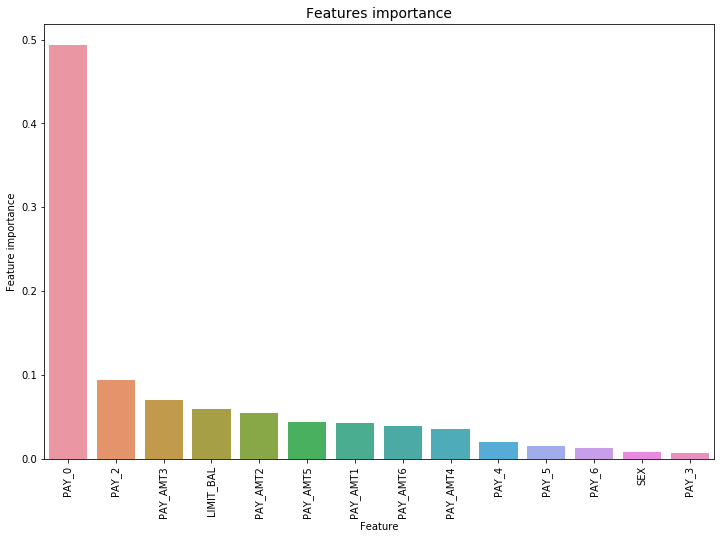
Min samples split: between 2, 5, 10, 20

Gridsearch determines the best parameters to use in this model as criterion: gini, max\_depth: 3, max\_leaf\_nodes: 5, min\_samples\_split: 2.

After using Gridsearch, we can see 3% improve on our accuracy score. Now let’s check the feature importance, the confusion matrix and AUC scores.

### Feature Importance

Below is the feature importance graph for our decision tree classifier. As it can be seen, the most important feature for our model is PAY\_0 (Repayment status in September, 2005 column).



### Confusion Matrix

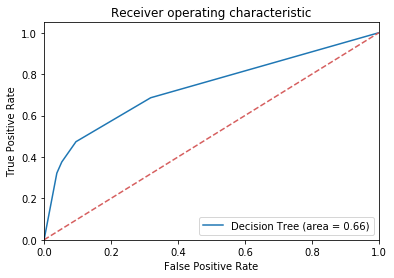
Below is the confusion matrix of our model. We can easily say that our data fits better on decision tree classifier rather than Logistic Regression.

Accuracy of Decision Tree classifier on test set: 0.82

Confusion matrix:

[[6609 369]

[1262 760]]



AUC score is 0.66 which is a great improvement after Logistics Regression.

## 

## 

## Random Forest Classifier

We will try the Random Forest Classifier on our dataset. Random Tree Classifier solves the overfitting problem that Decision Tree Classifiers often fail. Therefore, we will try random forest classifier and see if we can get better scores than our previous model.

We will build our random classifier model with criterion as 'gini', number of estimators as 100, number of parallel jobs as 4. Once again, we train our training data (X\_train and y\_train) and we test out test data (X\_test) to see the predictions.

Let’s check the most important feature of this model first.

### Feature Importance

The best features for our Random Forest Classifier are PAY\_AMT1 (Amount of previous payment in September, 2005), PAY\_0 (Repayment status in September, 2005 column), LIMIT\_BAL ( Amount of given credit).

### Confusion Matrix

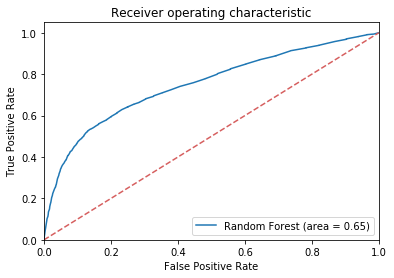
Below is the confusion matrix and accuracy scores of our random forest classifier.

Accuracy of Random Forest classifier on test set: 0.81

Confusion matrix:

[[6599 379]

[1296 726]]



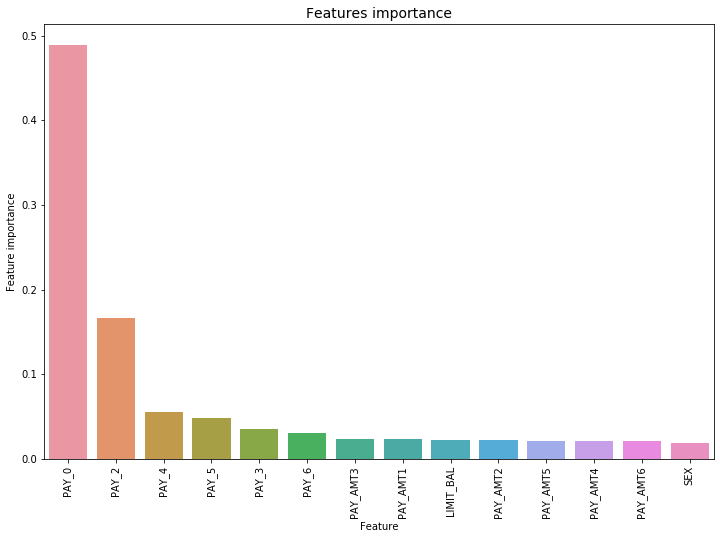
AUC score is 0.65, surprisingly slightly lower than the decision tree’s AUC score (0.66).

## XGBoost Classifier

The last classifier we will try our dataset on is Xgboost classifier. Xgboost classifier is one of the most popular algorithms in the modern machine learning world due to its computing speed and performance.

We will fit our data into this algorithm, check feature importance, confusion matrix and accuracy score with Xgboost classifier.

### Feature Importance



Most important feature for Xgboost classifier is PAY\_0 (Repayment status in September, 2005 column).

### 

### Confusion Matrix

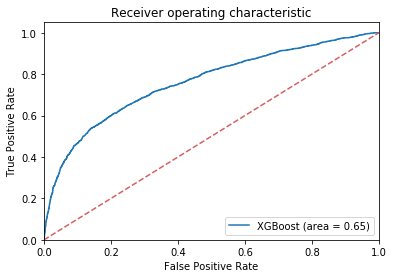
The confusion matrix and accuracy score of Xgboost Classifier is shown below;

Accuracy of XGBoost on test set: 0.81

Confusion matrix:

[[6616 362]

[1309 713]]



AUC score is 0.65, similar to random forest’s AUC, but again slightly lower than decision tree’s AUC score.

# Conclusion

In this project, we cleaned the data, looked for correlation between the predictor variables and four three different models to predict if the person will default the next credit card payment or not. While working on the project, we asked questions and got answers like below:

There are 30,000 bank clients' data in this dataset.

There are more women clients than men.

Average credit card limit is 167,484 NT Dollar .

Most of the clients are college educated.

Most of the clients are married.

Average age of the clients is 35.5 and standard deviation is 9.21.

75% of the male clients make their next payment, 25% of them default the next payment.

80% of the female clients make their next payment, 20% of them default the next payment.

About the prediction models, Decision Tree Classifier gives us the best AUC score among the four models we tried in our project. Below is the list of all classifiers are used in this project with AUC scores:

Logistic Regression Classifier has 0.50 AUC score.

Decision Tree Classifier has 0.66 AUC score.

Random Forest Classifier has 0.65 AUC score.

XGBoost Classifier has 0.65 AUC score.