# Credit Card Default Prediction

By

ALOK KUMAR

**Data science trainee, Alma Better, Bangalore**

## Abstract

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



## 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 K-S chart to evaluate which customers will default on their credit card payments.

## Dataset 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).

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; . . .;

X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -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.

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

X12 = amount of bill statement in September, 2005;

X13 = amount of bill statement in August, 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; . . .;
X23 = amount paid in April, 2005.

## Steps involved:

1. Importing important libraries to be used.
2. Mounting the drive and loading the dataset.
3. Exploratory data Analysis
4. Data extraction and manipulation.
5. Splitting the data using Train-test-Split.
6. Fitting the dataset over various models like logistic regression, random forest, KNN, Stochastic Gradient Descent , SMOTE.

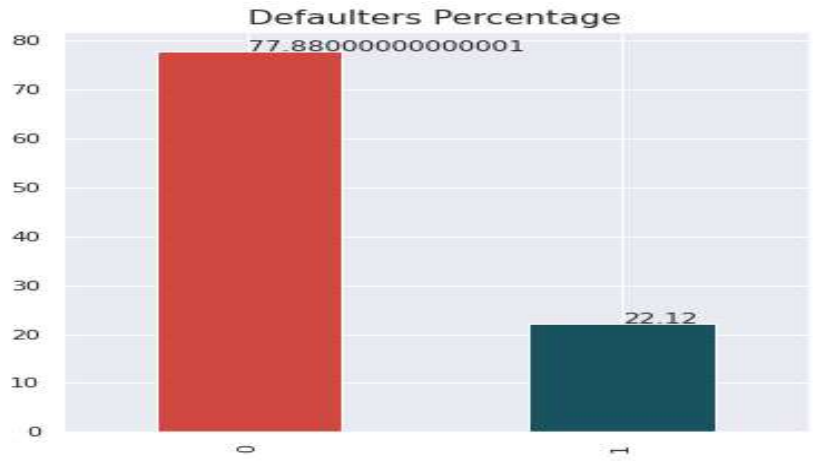
## Exploratory Data Analysis

1. In this Analysis Part, Firstly I viewed few insights of the dataset and checked if there are any null values or not.

2. Checked the summary of dataset by using describe function.

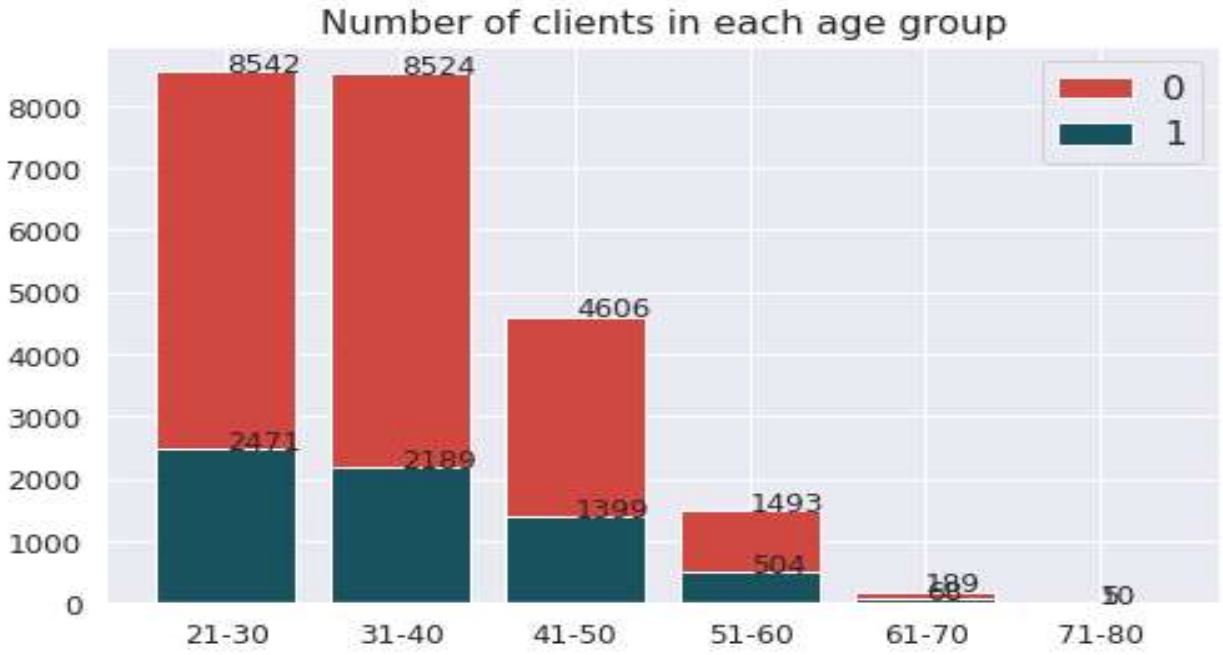
3. Renamed PAY_0 to PAY_1 and default payment next month to Default.
4. Also Dropped ID column which is not of our use here.

5. Then calculated percentage of **defaulters** present in our dataset.



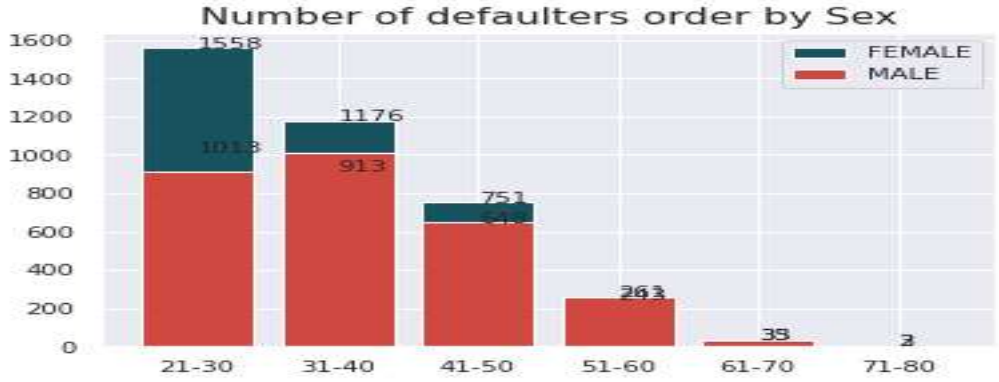
Clearly we have **22% defaulters** in our dataset and 77% persons are non defaulters

6. Then I checked the defaulters by **age**, **sex**, limit balance.




We have maximum clients from 21-30 age group followed by 31-40. Hence with increasing age group the number of clients that will default the payment next month is decreasing.

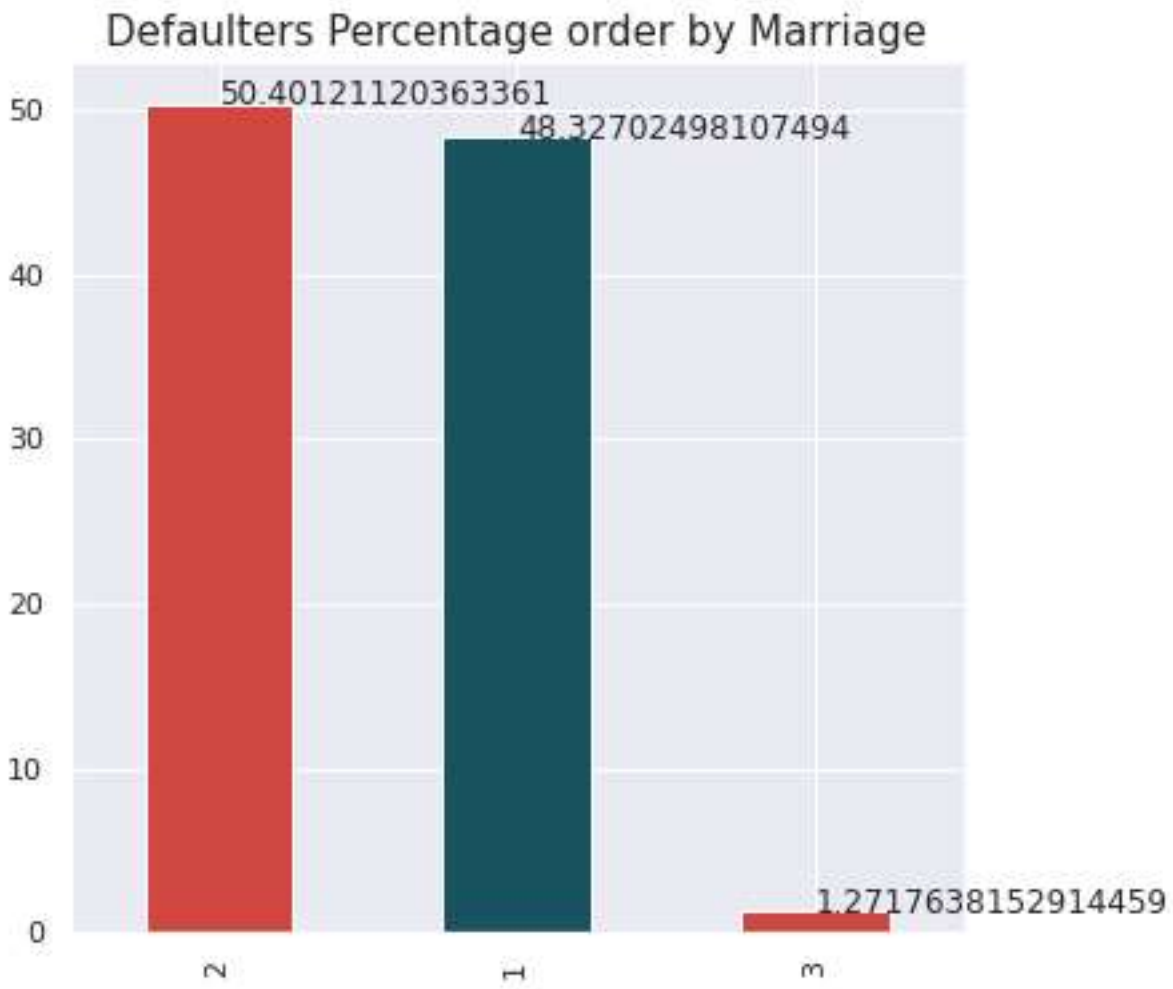
7. Then calculated **number of defaulters** with respect to **Gender** and their respective **AGES**

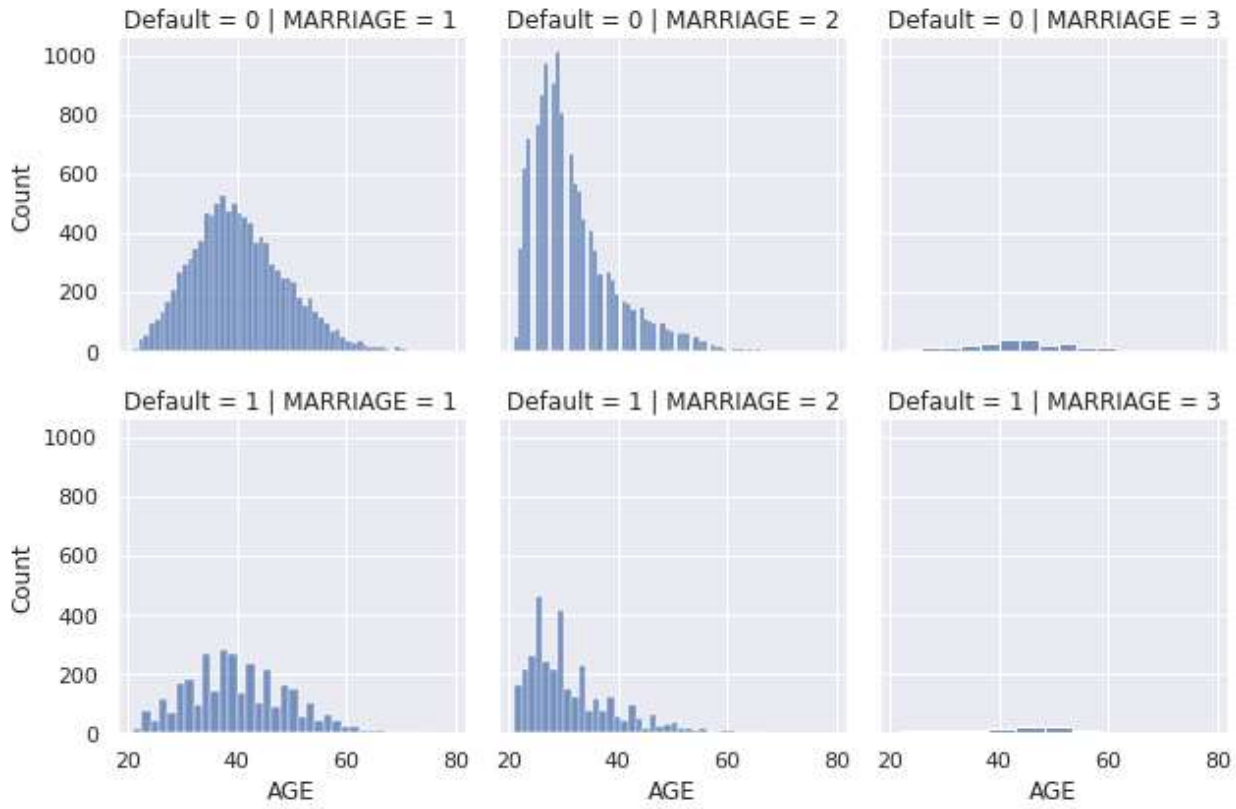


Number of defaulters order by Sex we have female defaulters more than males in 20-30 age group

8. Checking for Marriage and Education perspective of defaults. Here I did some data manipulation work of removing unwanted rows with inappropriate values.

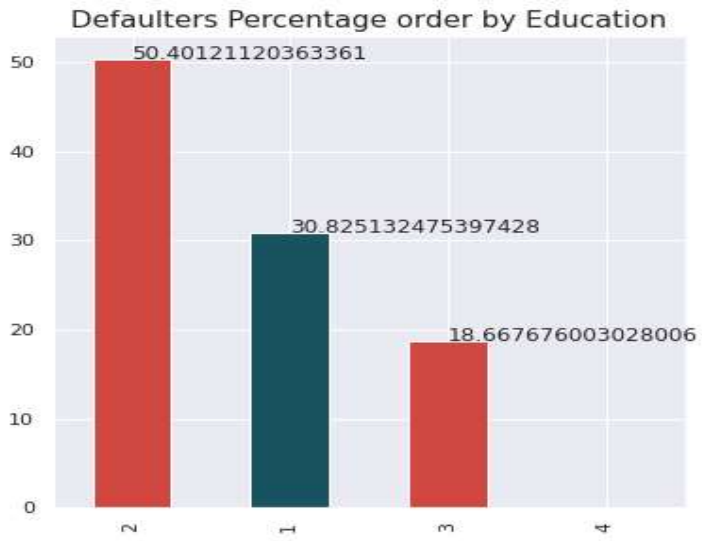
9. Plotting defaulters by their Marriage.





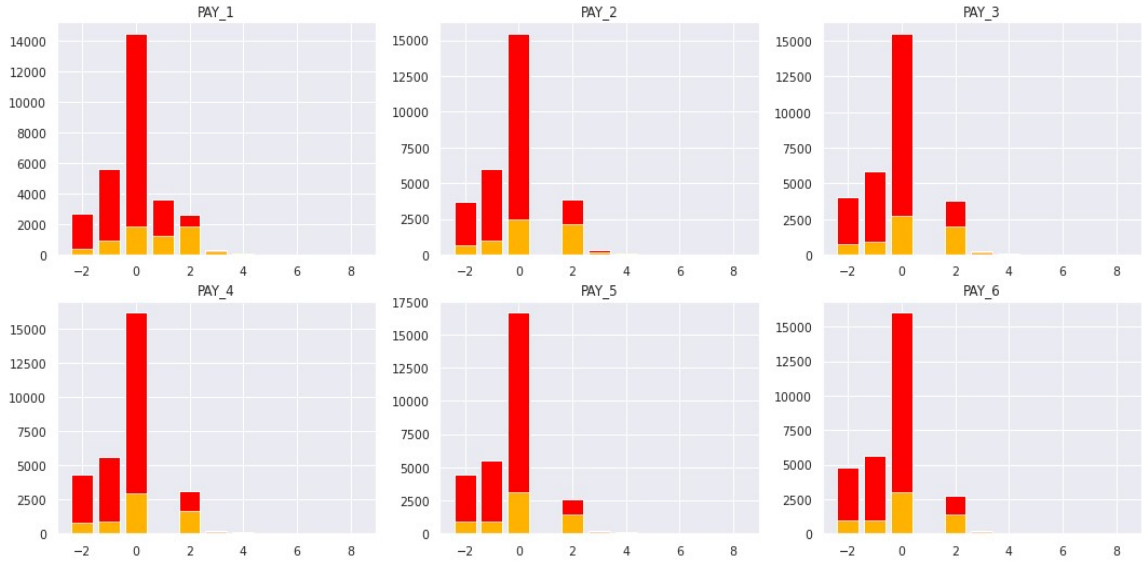

We can see there is **no trend or behaviour** of married or unmarried people as a defaulter.

## 10. Plotting defaulters by their Education



**University level** student tend to **default** more followed by **graduate** and **high school** students.

**11.** Plotting **bar plot** for **each month payment status** which show the count of **defaulters and non-defaulter**.




From the below **FaceGrid Plot** we can see that **NonDefaults** have a higher proportion of people **30-40** years.

12. Converting the columns of **SEX**, **EDUCATION** into **continuous variable** from object type with pandas **get_dummies** encoding.

# Split the data into train and validation dataset

I have used Train Test Split to split the data to find out which type give the best model results.

# Model Fitting

Here I fitted models over Logistic regression, logistic regression, random forest, KNN, Stochastic Gradient Descent , SMOTE and used evaluation matrices to check the performance of fitted models.

## Model Evaluation Metrics

I have checked for Accuracy, Precision, Recall and F1 Score for evaluating our model performance.

## 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 (aka Logistic).

Here we found the accuracy to be 0.7768.

## Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to discriminative learning of linear classifiers under convex loss functions such as (linear) Support Vector Machines and Logistic Regression.

It is one of the Gradient Descent Algorithm. It uses only a single example (a batch size of 1) per iteration. Given enough iterations, SGD works but is very noisy. The term "stochastic" indicates that the one example comprising each batch is chosen at random.

| | Model | Accuracy | Precision | Recall | F1 Score | ROC |
|---|---|---|---|---|---|---|
| 0 | Stochastic Gradient Descent | 0.573552 | 0.282042 | 0.589705 | 0.381582 | 0.579309 |

## Random Forest

Testing Accuracy is- **0.8170917074818442**

roc_auc_score is- **0.6521385314155942**

recall_score is - 0.3542770628311885

## KNN

Accuracy score is- 0.7714913021449079

roc_auc_score - 0.5286285916466444

recall_score is - 0.09008327024981075

## SMOTE

Random Forest with SMOTE

Accuracy score is- 0.8091216216216216

roc_auc_score is- 0.6523221343873519

recall_score is - 0.3693181818181818

KNN with SMOT

Accuracy score is- 0.7658361486486487
roc_auc_score - 0.5380393610013176
recall_score is - 0.1268939393939394

**Random Under Sampler:**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.84 | 0.62 | 0.72 | 4600 |
| 1 | 0.31 | 0.60 | 0.41 | 1321 |
| accuracy | | | 0.62 | 5921 |
| macro avg | 0.58 | 0.61 | 0.57 | 5921 |
| weighted avg | 0.73 | 0.62 | 0.65 | 5921 |

## Conclusion

Our best prediction accuracy was around 88-89%, our lowest measured prediction accuracy was about 13%. This is not a particularly large spread, especially considering the disparity in the time it takes for some models to train with the Grid Search to find the best hyper-parameters.

we streamlined the training and prediction process very well, but despite the feature engineering, and hyper parameter tuning we didn't see a particularly significant change in predictive power of our models.

If we were to make recommendation to a crediting agency about the granting of credit, with respect to the debtors likelihood to default, it would be:

- Consider the applicants marital status. Married people seem to default more often.

- Consider the age of the applicant. Younger people are at higher risk of defaulting.

- Establish a lower limit balance for applicants that would be considered risky.

- Once granted credit, pay special attention to the ratio of payments to the amount owed on their balance. There will be some threshold were it's very unlikely that they will be able to pay you back.

- Limit the number of months a debtor may be late on their payments to 3, like most credit agencies. At this point it's more likely that the person will default than pay you back. At this point you're only giving more money away.

.