#### Credit Card Default Prediction

**Shubham Narendra Kadu**

**Data science trainees**

**AlmaBetter**

**Abstract:**

Now a day it is one of the biggest threats faced by commercial banks is the risk prediction of credit clients. Recent studies mostly focus on enhancing the classiﬁer performance for credit card default prediction rather than an interpretable model. In classiﬁcation problems, an imbalanced dataset is also crucial to improve the performance of the model because most of the cases lied in one class, and only a few examples are in other categories. Traditional statistical approaches are not suitable to deal with imbalanced data. Data level resampling techniques are employed to overcome the problem of data imbalance. Various under-sampling and oversampling methods are used to resolve the issue of class imbalance. Different machine learning models are also employed to obtain efficient results. This model will help commercial banks, financial organizations, loan institutes, and other decision-makers to predict the loan defaulter earlier.

**1. 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.

**2. Goal**

The objective of our project is to predict which customers might default in the upcoming months. Before going any further let's have a quick look at the definition of what is actually meant by Credit Card Default.

### **3. Introducti****on**

In today’s world credit cards have become a lifeline to a lot of people so banks provide us with credit cards. Credit card is a commonly used transaction method in modern society and one of the main business of banks. Credit card fraud is a huge ranging term for theft and fraud committed using or involving at the time of payment by using this card. The purpose may be to purchase goods without paying or to transfer unauthorized funds from an account. Credit card fraud is also an add-on to identity theft. also For banks, it helps the bank to generate interest revenue but at the same time, it raises the liquidity risk and credit risk to the bank. In order to control the cash flow and risk, detecting the customers with default payments next month could play an important role in estimating the potential cash flow and risk management.

**4. Feature Description**

1) ID: ID of each client

2) LIMIT\_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)

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

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

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

6) AGE: Age in years

7) PAY\_0-6: History of past payments from April to September

8) BILL\_AMT1-6: Amount of bill statement from April to September 2005 (NT dollar)

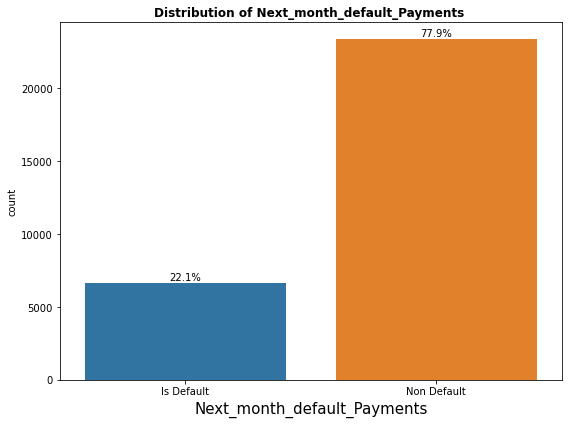
9) PAY\_AMT1-6: Amount of previous payment from April to September 2005 (NT dollar)

10) default\_payment\_next\_month: Default payment (1=yes, 0=no)

**4. Exploratory Data Analysis**

Exploratory Data Analysis (EDA) plays a vital role in the analysis of the data variables which are important from the aspect of feature engineering. It will help us to distribute and relate between dependent and independent variables. We have gone through an analysis of every independent as well as the dependent variable to check which independent factor affects the dependent factor.

**4.1)** **Dependent Variable: Defaulter**



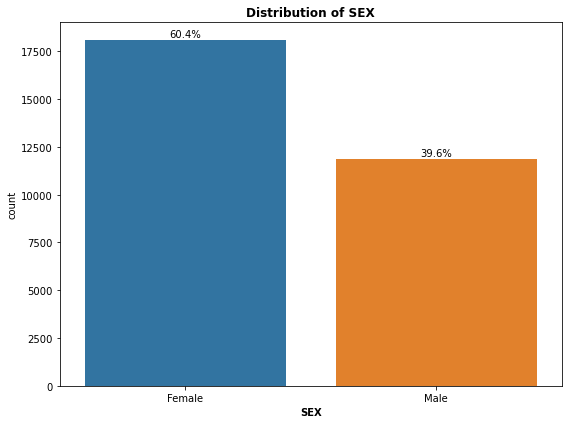
As we can see from the graph Here, there is a huge difference between non-defaulter (0) and defaulter (1).

Defaulters are less than the non-Defaulters.

Approx 78% are non-Defaulters and 22% are Defaulters respectively.

**4.2) Univariate Analysis**

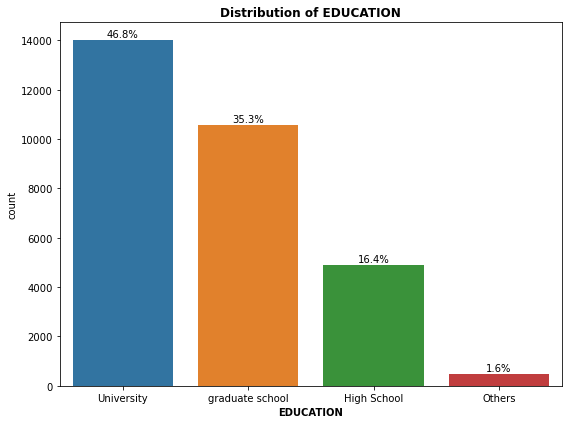
**SEX**

****

From the below graph, we can see that the Number of Male credit holders is less than females.

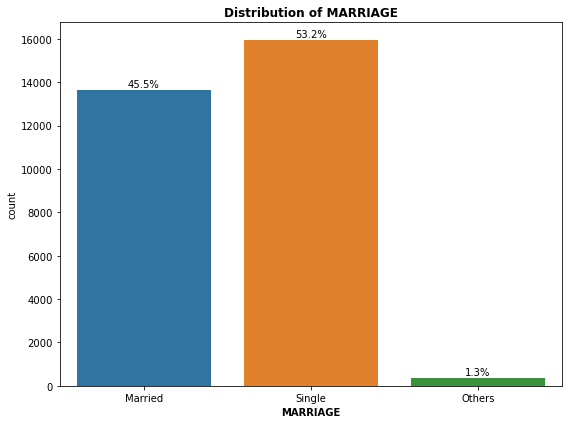
Approximately 40% are male and 60% are Female.

**EDUCATION**



From the below graph, we can see that More credit holders are university students followed by Graduates and then High school students.

**MARRIAGE**

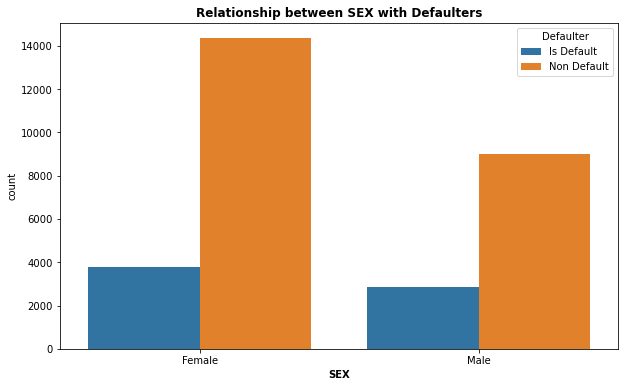


From the above data analysis, we can say that More number of credit cards holder are Single as compared to Married and others.

**Visualize the distribution of Close stock price**

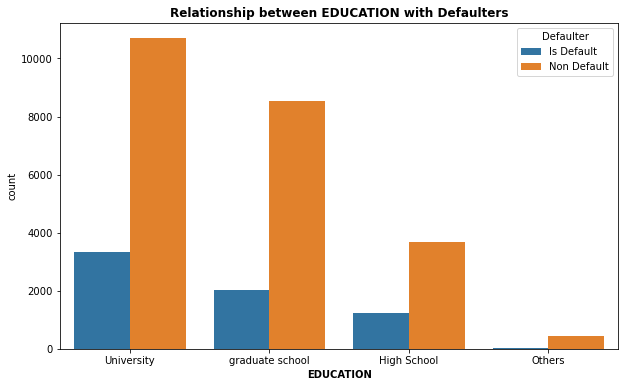
**4.3) Univariate Analysis**

**SEX vs DEFAULTER**



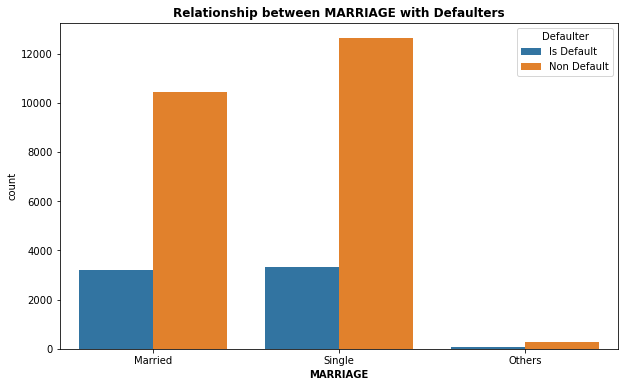
It is evident from the above graph that the number of defaulters has a high proportion of females.

**EDUCATION vs DEFAULTER**



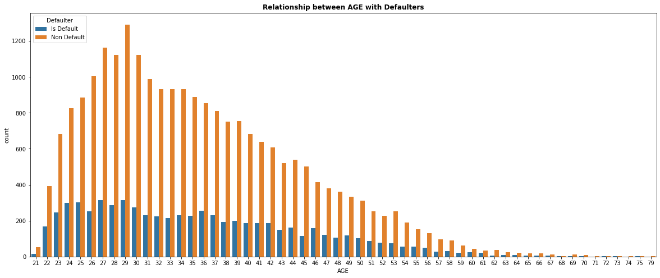
From the above graph, it is clear that those people who are university students have higher default payments w.r.to graduates and high school people.

**MARRIAGE vs DEFAULTER**



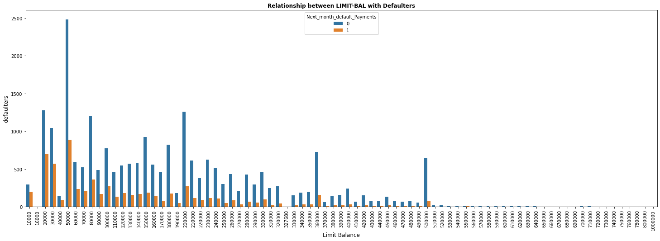
From the above graph, Here it seems that married, or single is most likely to default.

**AGE vs DEFAULTER**



More number of credit card holders aged between 26-32 years and 29 years age is the highest uses of credit cards. Age above 60 years old rarely uses a credit card. Also, more Defaulters are between 27-29 years.

**LIMIT BALANCE vs DEFAULTER**



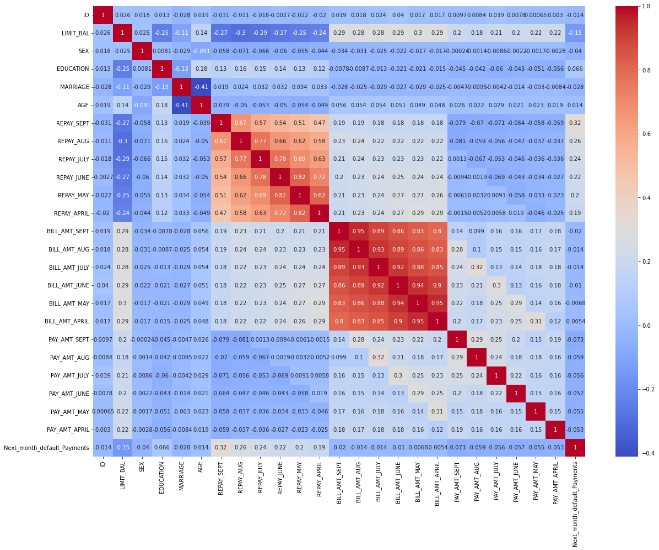
From the below plot, we clear that the majority of the defaulters are those who have a credit limit balance between 20,000 to 3,00,000 After the credit limit of 5,00,000, the number of defaulters is almost negligible.

**4.4) Correlation Analysis**

The correlation analysis has been done to get a better understanding of the dependent and independent variables’ multicollinearity. Multicollinearity may not affect the accuracy of the model as much but we might lose reliability in determining the effects of individual independent features on the dependent feature in your model and that can be a problem when we want to interpret your model.

**Heatmap**

Let’s check the heatmap plotted concerning independent variables.



* The heatmap shows some high correlations between variables and also visualizes how each parameter's correlation will respect every other parameter.
* Above heatmap It seems that there is some negatively correlated feature like age and marriage. and ID is unimportant and has no role in prediction so we will remove it.

**SMOTE (SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE)**

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.

**5) 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 perform one hot encoding on 'EDUCATION','MARRIAGE','PAY\_SEPT', 'PAY\_AUG', 'PAY\_JUL', 'PAY\_JUN', 'PAY\_MAY', 'PAY\_APR'.

**6) Building Machine Learning Algorithm**

The provided data is first cleaned and visualized. We then split the data into the Train set (for Hyperparameter tuning) and Test set (for Model Evaluation). Using accuracy as our evaluation metric, we compare various models and select the classification algorithm based on the lowest accuracy on the Test data.

**6.1) Train/Test Split**

The train/test split was done as 0.33 % on data with a random state of 0. The final dataset which was split into (31307, 84) as Train data and (15421, 84) as Test data.

**6.2) Logistic Regression**

Logistic Regression is a Machine Learning algorithm and is basically used for binary classifications like yes-no, true-false, male-Female, etc.

It takes the linear combination and applies a sigmoid function (logit). The Sigmoid curve gives a value between 0&1.

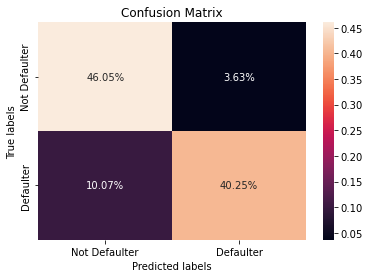
To represent binary/categorical values, dummy variables are used. For the purpose of a special case in logistic regression is a linear regression, when the resulting variable is categorical then the log of odds is used for the dependent variable, and also it predicts the probability of occurrence of an event by fitting data to a logistic function. Such as

O = e^ (I0 + I1\*x) / (1 + e^ (I0 + I1\*x)) (3.1) Where,

O is the predicted output

I0 is the bias or intercept term

I1 is the coefficient for the single input value (x).



Evaluation Metrics:

Training accuracy = 0.86597

Testing Accuracy = 0.8644

Precision Score = 0.8042

Recall Score = 0.9160

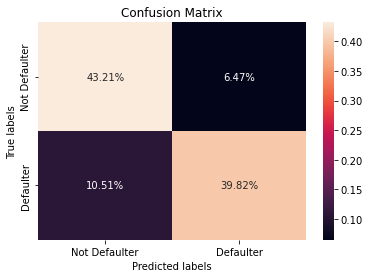
F1\_Score = 0.8565

ROC\_AUC score = 0.8697

**6.3) Decision Tree Classifier**

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

The objective of the Decision tree algorithm is to find the relationship between the target column and the independent variables and Express it as a tree structure



Evaluation Metrics :

Training accuracy = 0.8503

Testing accuracy = 0.8317

Precision score = 0.8600

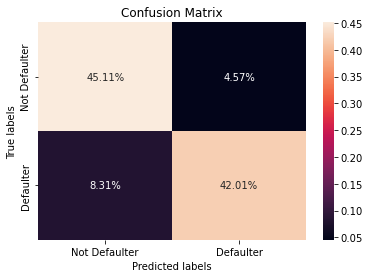
Recall score = 0.7706

F1\_score = 0.8217

ROC-AUC score = 0.83054

**6.4) Random Forest Classifier**

The random forest is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree



Evaluation Metrics:

Training accuracy = 0.9995

Testing Accuracy = 0.8747

Precision score = 0.8395

Recall score = 0.9046

F1\_score = 0.8708

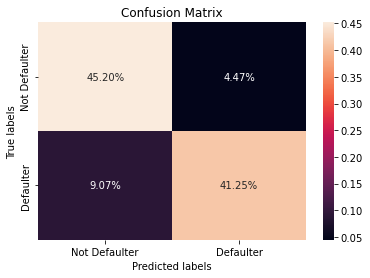
ROC\_AUC score = 0.8765

**6.5) XG-BOOOT CLASSIFEIR**

XG-BOOST is a powerful iterative learning algorithm based on gradient boosting.

Regularizations to avoid overfitting Tree pruning using a depth-first approach.

It is generally used for very large dataset



Evaluation Metrics:

Training accuracy = 0.9156

Testing accuracy = 0.8656

Precision score = 0.8186

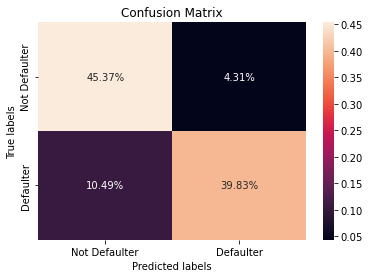
Recall score = 0.9052

F1\_score = 0.8597

ROC\_AUC score = 0.8689

**6.6)** **KNN- CLASSIFIER**

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand but has a major drawback of becoming significantly slows as the size of that data in use grows.



Evaluation Metrics:

Training accuracy = 0.8724

Testing accuracy = 0.8548

Precision score = 0.8041

Recall score = 0.8968

F1\_score = 0.8479

ROC\_AUC score = 0.8586

**8) Conclusion**

1)first We started with data inspection, viewed the data distribution

2)By Visualization we have checked the distribution of defaulters vs non-defaulters and we see around 78% are non-defaulters and 22% are defaulters.

3)the distribution of sex, Education, and Marriage with respect to the defaulter. and we found in Sex more defaulter is Female, in Education, more number of the defaulters is a university student and in Marriage more number of the defaulters by single.

4)After that we built a model (Logistic Regression, Decision Tree, Random Forest, XG-Boost classifier, and KNN), and all of them in, the best accuracy has obtained from the Random Forest Classifier.

5)Using a Logistic Regression classifier, we can predict with 86.38% accuracy, whether a customer is likely to default next month. Using the Decision Tree classifier, we can predict with 82% accuracy whether a customer is likely to default next month OR not. Using Random Forest, we can predict with 87% accuracy whether a customer will be a defaulter in the next month. Using XGBoost Classifier, we can predict with 86.64% accuracy whether a customer will be a defaulter in the next month. And By applying KNN Classifier with 85% accuracy whether a customer will be a defaulter in the next month.

6)From the Above evaluation table Logistic regression model has the highest recall, if the business cares about recall the most, then this model is the best candidate. If the balance of recall and precision is the most important metric, then Random Forest is the ideal model. but Since the Random Forest classifier has also a higher Recall score.so I would recommend Random Forest.

7)From the above evaluation table we can also see that the Random Forest Classifier having Recall, F1-score, and ROC Score values equal 90.46%, 86.92%, and 87.53% resp. and XGBoost Classifier having Recall, F1-score, and ROC Score values equal 90.29%, 86.11%, and 86.92% resp.

8)From the models that are applied to the dataset, we can conclude that these two Random Forest and XG-Boost are giving the best evaluation metric (Recall, F1-score, and ROC-AUC score) and with the help of these two models we are the best to predict whether the credit card is the default or not default according to our analysis.