**SUPERVISED ML - Classification**

**Credit card default Prediction**

**­­­**



**-** **Kaiwalya Zankar**

**Data science trainee at**

**Almabetter**

**ABSTRACT**

Credit risk plays a major role in the banking 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. Aiming at the problem that the credit card default data of a financial institution is unbalanced, which leads to unsatisfactory prediction results This paper provides a performance evaluation of credit card default prediction. Thus, logistic regression, decision tree, and random forest are used to test the variable in predicting credit default and random forest proved to have the higher accuracy.

**PROBLEM STATEMENT**

This project is aimed t 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 to evaluate which customers will default on their credit card payments

**INTRODUCTION**

Credit risk has traditionally been the greatest risk among all the risks that the banking and credit card industry are facing, and it is usually the one requiring the most capita.

Despite machine learning and big data have been adopted by the banking industry, the current applications used by the banks are mainly focused on credit score predicting. The disadvantage of heavily relying on credit score is banks would miss valuable customers who come from countries that are traditionally underbanked with no credit history or new immigrants who have repaying power but lack credit history

Due to the scope of the project and lack of computational resources, this analysis is not intended to be exhaustive, we only applied 3 classification machine learning models

**DATASET PREPROCESSING**

First I remove the column header as it was creating problem in analysis and after I have done some minor chances with converting it into lowercase

As dataset didn’t consist null value no required to update the dataset.

**DATA DESCRIPTION**

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.

**CHALLENGES FACED**

* As data is imbalanced , to balance the data in an appropriate way was a bit tricky.
* Using the appropriate metrics for comparison of the implemented machine learning algorithms.

**APPROACH**

We checked the Outliers and correlation matrix to overcome the noise in the dataset.Also,data was balanced using the SMOTE method and scaled by Standard Scaler transformation. As the Credit card default dataset defines the classification problem. We decided to train the models such as Logistic regression, Random forest Classifier, Gradient Boosting Classifier.Also, we used Hyperparameter Tuning for improvement in the model fitting to understand the better results of the model as well as the metrics.

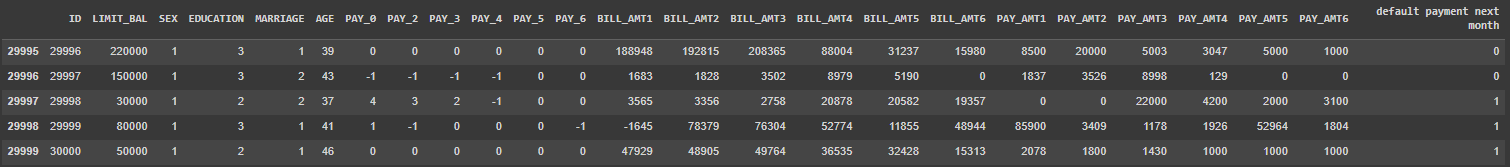
**TOOLS USED**

Following libraries were used for analyzing the data and visualizing:

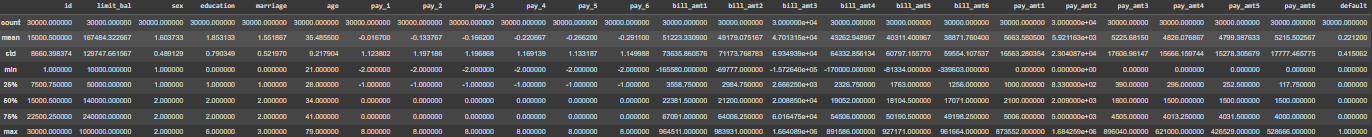
* Numpy: For some math operations in predictions.
* Pandas: Extensively used to load and wrangle with the dataset.
* Seaborn: Used for visualization.
* Matplotlib: Used for visualization
* Scipy – used to slove Scientific and mathematical problems..
* Sklearn: For analysis and prediction.
* Imblearn: Imbalanced-learn (imported as imblearn ) is an open source, MIT-licensed library relying on scikit-learn (imported as sklearn ) and provides tools when dealing with classification with imbalanced classes.
* Warnings: For filtering and ignoring warnings.
* Stats models: For outliers influence.
* Xgboost: XGBoost provides a wrapper class to allow models to be treated like classifiers or regressors in the scikit-learn framework.
* Pprint: The pprint module in Python is a utility module that you can use to print data structures in a readable, pretty way
* train\_test\_split: To split the dataset into training and testing data
* StandardScaler: To scale all the features, so that the Machine Learning model better adapts to the dataset

**DATA INSIGHTS**

**Pandas dataframe**

****

**Describe**

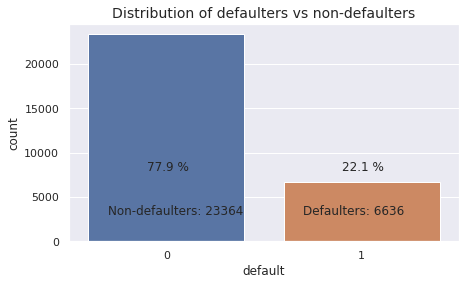
****

**Checking the NAN values-**

|  |  |
| --- | --- |
| id | 0 |
| Limit\_bal | 0 |
| sex | 0 |
| Education | 0 |
| Marriage | 0 |
| Age | 0 |
| Pay\_1 | 0 |
| Pay\_2 | 0 |
| Pay\_3 | 0 |
| Pay\_4 | 0 |
| Pay\_5 | 0 |
| Pay\_6 | 0 |
| Bill\_amt1 | 0 |
| Bill\_amt2 | 0 |
| Bill\_amt3 | 0 |
| Bill\_amt4 | 0 |
| Bill\_amt5 | 0 |
| Bill\_amt6 | 0 |
| Pay\_amt1 | 0 |
| Pay\_amt2 | 0 |
| Pay\_amt3 | 0 |
| Pay\_amt4 | 0 |
| Pay\_amt5 | 0 |
| Pay\_amt6 | 0 |
| Default | 0 |
| Dtype: int64 | |

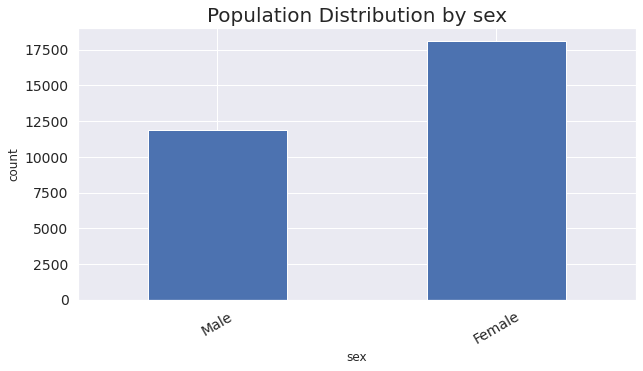
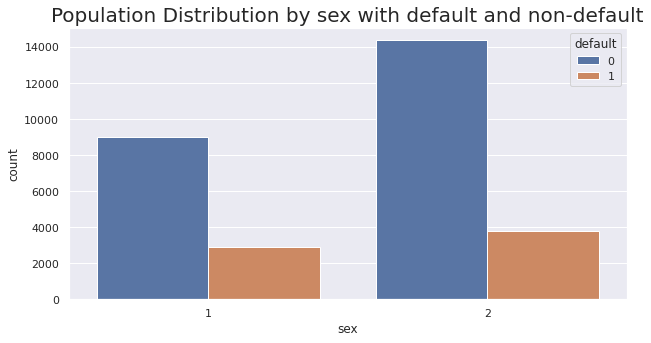
As dataset doesn’t contain any null/nan values

**Histogram Plot for defaulters & non defaulters**

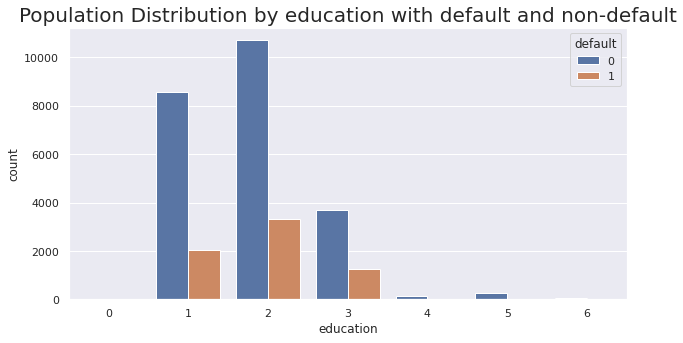


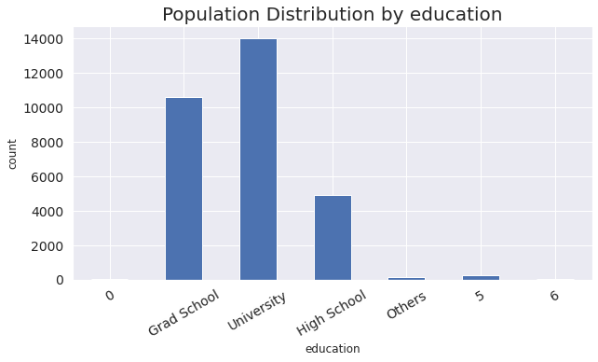
**Relationship between variables and defaullt.**

1. Is default proportion affected by gender

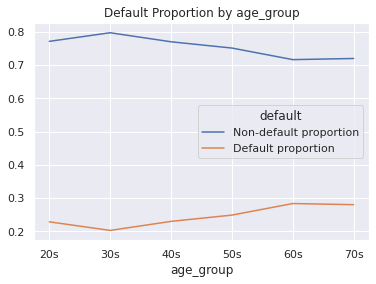


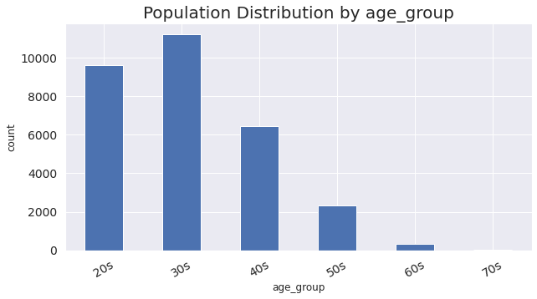
Default proportion among men is higher

1. Is default proportion affected by education?



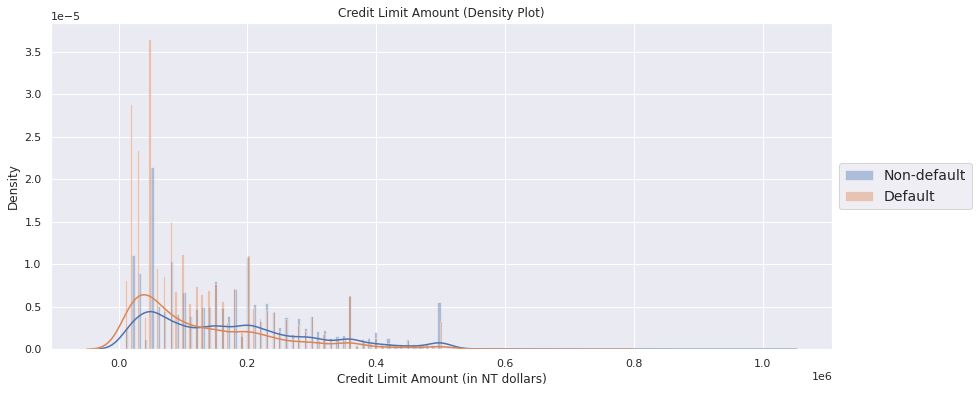
default proportion decreases with higher education level.

1. Is proportion of defaults correlated with age?



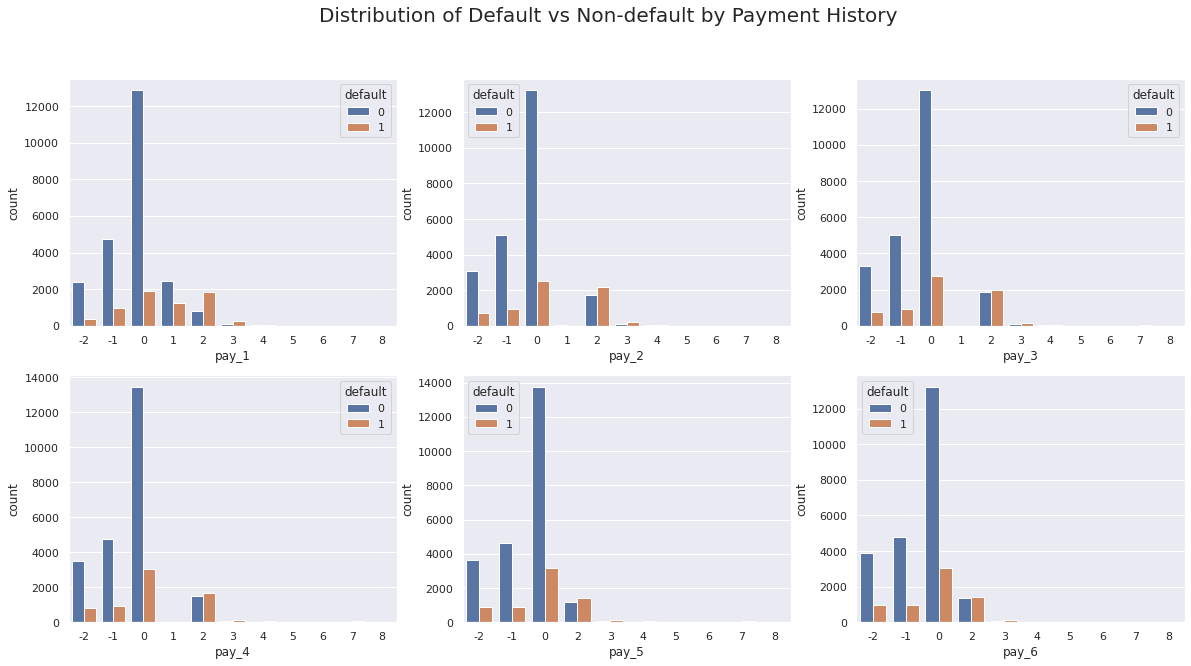
The default proportion is lowest for people in their 30s and steadily rises with age.

1. Is default proportion affected by credit limit?



It seems that people with higher credit limit have significantly lower default proportion.

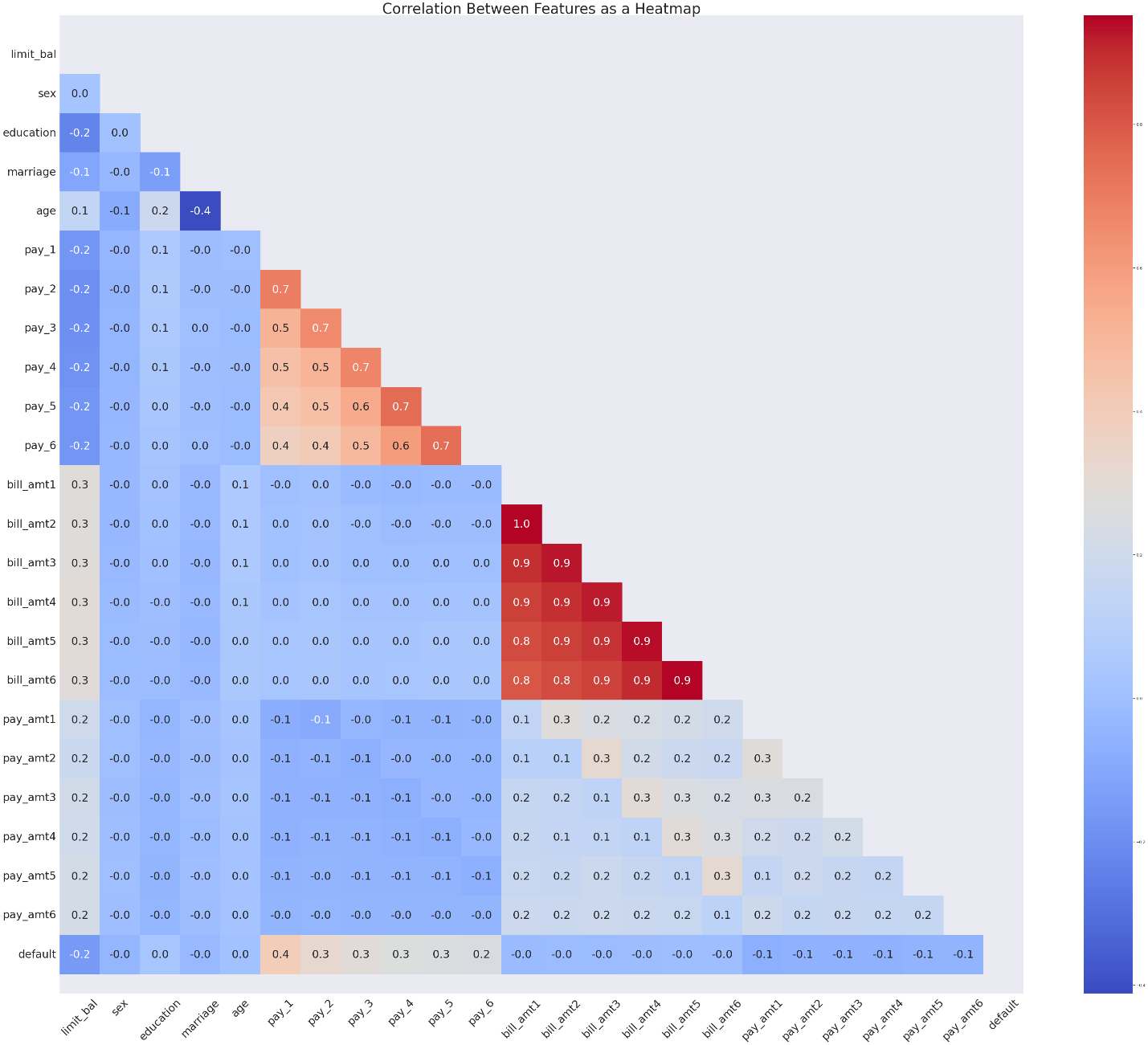
1. Is the default proportion affected by history of past repayment status?



If the person has defaulted for 2 months or more in the past two months, there is a very high chance of them defaulting.

**CORRELATION MATRIX (HEATMAP)**

To refine the data further, a correlation matrix was created amongst all the feature variables to analyze interaction effects.



There is a high correlation among the payment history features and the bill amount features.

**DATA MODELING**

After the data preparation is completed it is ready for the purpose of analysis. Only numerical valued features are taken into consideration. The data were combined and labeled as X and y as independent and dependent variables respectively.

**SPLITTING THE DATASET**

The train\_test\_split was imported from the sklearn.model\_selection. The data is now divided into 70% and 30% as train and test splits respectively. 70% of the data is taken for training the model and 30% is for a test and the random state was taken as 0.

**SCALING THE DATA**

We have used the **Standard Scaler method** to scale the dataset.

Standardize features by removing the mean and scaling to unit variance.

The standard score of a sample x is calculated as:

z = (x - u) / s

where u is the mean of the training samples or zero if with\_mean=False, and s is the standard deviation of the training samples or one if with\_std=False.

**METRICS USED**

* **Classification Report**:A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of your trained classification [model](https://thecleverprogrammer.com/2020/11/27/machine-learning-algorithms-with-python/).

1) **Accuracy**: the proportion of total dataset instances that were correctly predicted out of the total instances

accuracy=(true positives+true negatives)/total

2) **Recall (sensitivity)**: the proportion of the predicted positive dataset instances out of the actual positive instances

sensitivity=true positives/(true positives+false negatives)

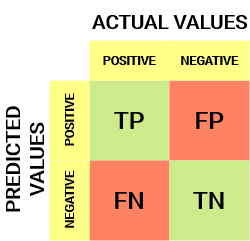
3) **F1 score**: a composite harmonic mean (average of reciprocals) that combines both precision and recall. For this, we first measure the precision, the ability of the model to identify only the relevant dataset instances

precision=true positives/(true positives+false positives)

The F1 score is estimated as

F1=2×(precision×recall)/(precision+recall)

* **Confusion Matrix:**A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.



* True Positive (TP) -The predicted value matches the actual value.The actual value was positive and the model predicted a positive value
* True Negative (TN) -The predicted value matches the actual value.

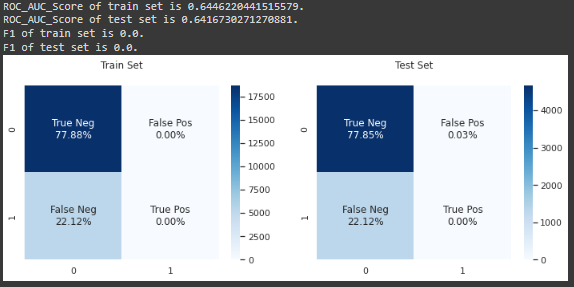
The actual value was negative and the model predicted a negative value

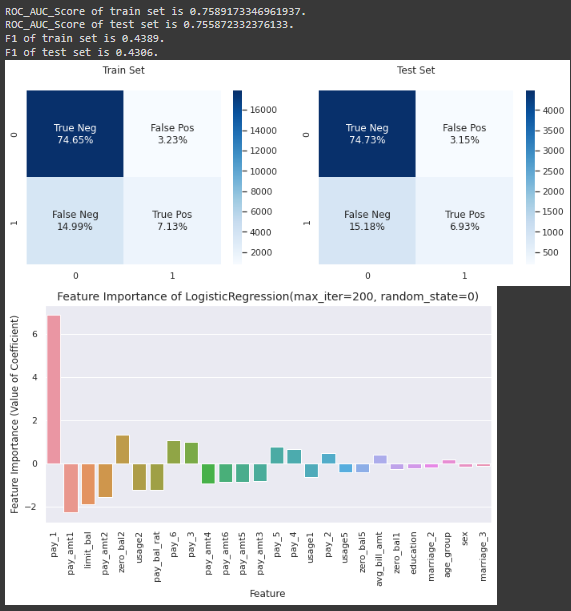
* False Positive (FP) – Type 1 error :-The predicted value was falsely predicted.The actual value was negative but the model predicted a positive value
* False Negative (FN) – Type 2 error:-The predicted value was falsely predicted.The actual value was positive but the model predicted a negative value

**MODEL IMPLEMENTATION**

* **Logistic Regression -**A logistic regression is a type of statistical procedure. It is used to refer specifically to the problem in which the dependent variable is binary, that is the number of available categories is two, while the problem with more than two categories is referred to as multi logistic regression.

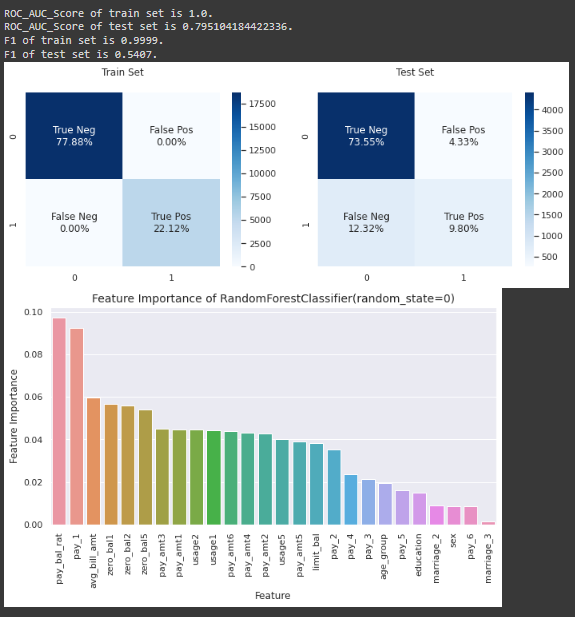
**Orignal data**

****

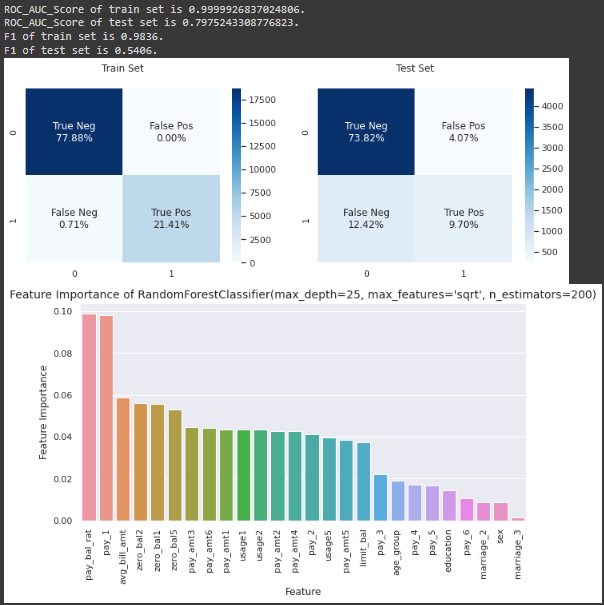
**Engineered data**

* **Random Forest Classifier -** The Random Forest algorithm falls under the Supervised Learning category. Random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting

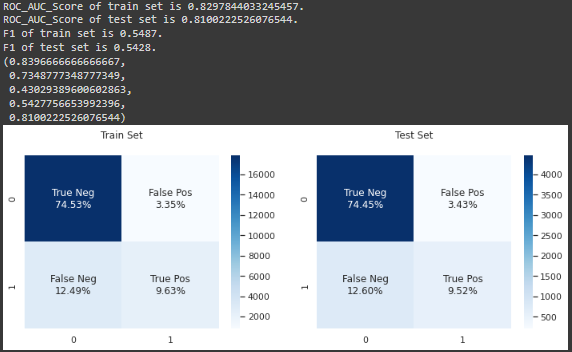
**Un tuned data output**



**Tuned data output**

****

* **Gradient boosting Classifier:** Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive mode.l



**COMPARISON OF MODEL**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Model | Accuracy | Precision | Recall | F1 score | AUR\_ROC |
| **0** | Logistic Regression (original data) | 0.778500 | 0.000000 | 0.000000 | 0.000000 | 0.641673 |
| **1** | Logistic Regression (engineered data) | 0.816667 | 0.687603 | 0.313489 | 0.430642 | 0.755872 |
| **2** | Untuned Random Forest Model | 0.833500 | 0.693396 | 0.443105 | 0.540690 | 0.795104 |
| **3** | Tuned Random Forest Model | 0.835167 | 0.704600 | 0.438583 | 0.540641 | 0.797524 |
| **4** | Gradient Boosting Classifier | 0.829667 | 0.709191 | 0.389601 | 0.502918 | 0.776989 |

* Other than the naive logistic regression on the original dataset, all of the models have pretty good AUC\_ROC scores. For example, the gradient boosting classifier assigns a higher probability of default to a defaulter over a non-defaulter with almost 81% certainty. Also, a lot of the engineered features are some of the most important features and that implies that the data reflects what an analyst would intuitively think to be true.
* The credit limit is a good indicator of financial stability. Whatever mechanism the bank is currently using works well and some of the features that go into choosing the credit line can be used directly in the model for default prediction.
* Although a gradient boosted model or a tuned random forest model might perform better, it might be worth working on a logistic regression model (especially if the model complexity can be reduced by dropping irrelevant features). A deployable model doesn't just need to be reliable, it also needs to be fast.

**CONCLUSION**

* Human characteristics are not the most important predictors of default, the payment status of he most 2 months and credit limit are.
* The best predictor (by far) of delinquency is the behavior in the past couple of months, and in particular the last month.
* From the modeling, we are able to classify default risk with accessible customer data and find a decent model.
* Lastly, I suggest the model output probabilities rather than predictions, so that we can achieve higher accuracy and allow more control for human managers to quantify default risk.