# CREDIT CARD RISK ANALYSIS

BACKGROUND:

Credit risk refers to the potential for loss due to the failure of a borrower to make a payment when it is due. The risk is mainly for the lender and it can include complete or partial loss of principal, loss of interest, and disruption of cash flow.

The application of scoring models in today’s business environment covers a wide range of objectives. The original task of estimating the risk of default has been augmented by credit scoring models to include other aspects of credit risk management: at the pre-application stage (identification of potential applicants), at the application stage (identification of acceptable applicants), and at the performance stage (identification of possible behavior of current customers).

ABSTRACT:

Classify Credit Card customers based on their history to good or bad customers

DATA SUMMARY:

The data is provided in the csv format, need to be imported using Pandas library. The data contains both the Customer and Card History.

METHODS SUMMARY:

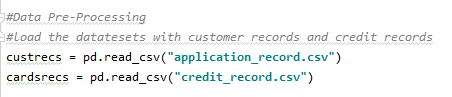
The table shows the list of data pre-processing, analysis, visualization,and model building techniques applied to complete the project

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Task Details | Analytical Techniques | Visualization  Techniques |
| Data Manipulation & Preparation | 1. Perform required data manipulation and cleaning. 2. Perform Uni variate and Bi variate analysis | 1. Descriptive statistics and outlier analysis 2. Independent t-test and VIF check to get important features | 1. Bar Plot, Seaborn (Heatmap), boxplots 2. Boxplot segmentation |
| Model Building & Performance Check | Create a Model and Asses the performance of the models | Build Logistic, Decision tree models, Random forest, and xgboost used hyperparameter  to fine-tune the model. | Visualize the decision tree using graphviz package |

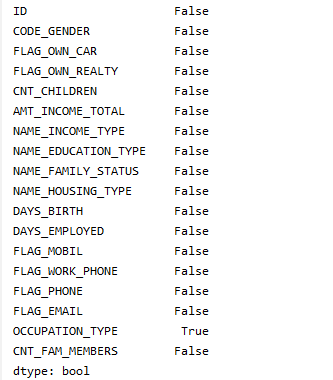
MODEL TO PREDICT DEFAULT CUSTOMERS:

* PART A: DATA MANIPULATION AND CLEANING

Read the input dataset into pandas data frame –

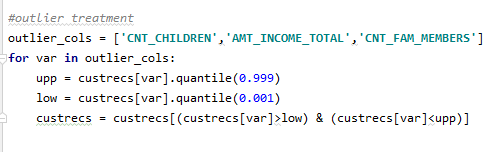


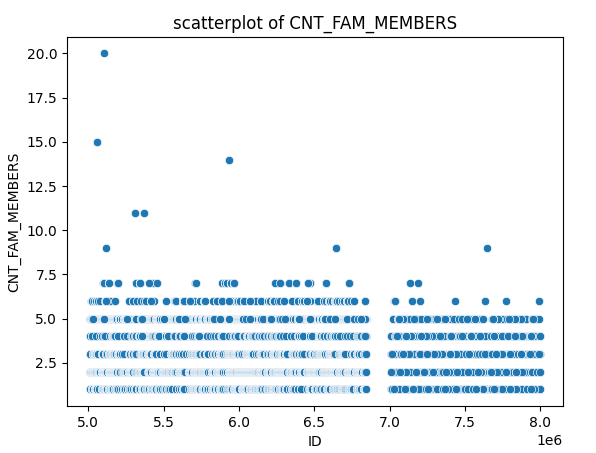
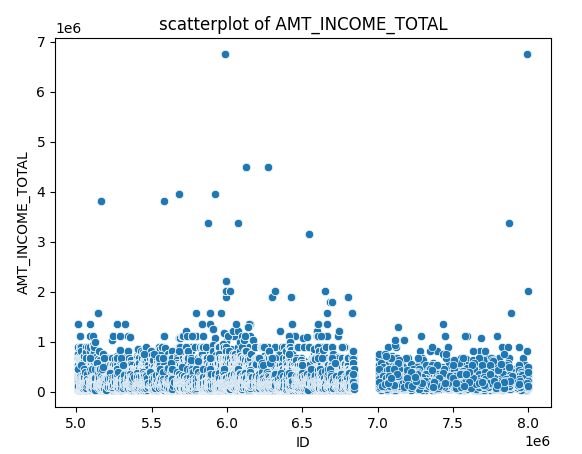
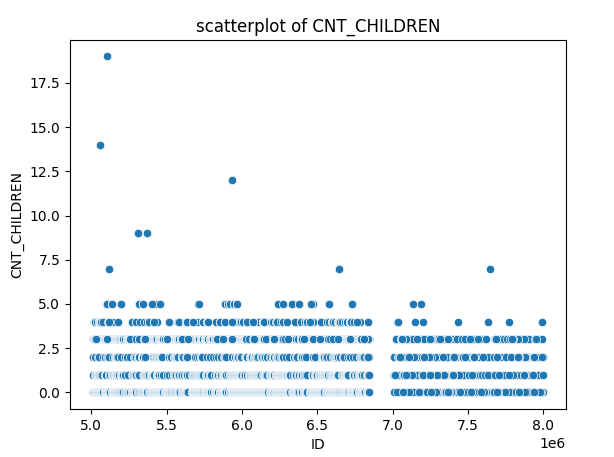
Checked for missing values in the dataset



There are few missing values in the dataset, on further analysis it is found that missing values in the default column belongs to a new set of customers.

Used the concept of winsorization to handle the outliers in the dataset.



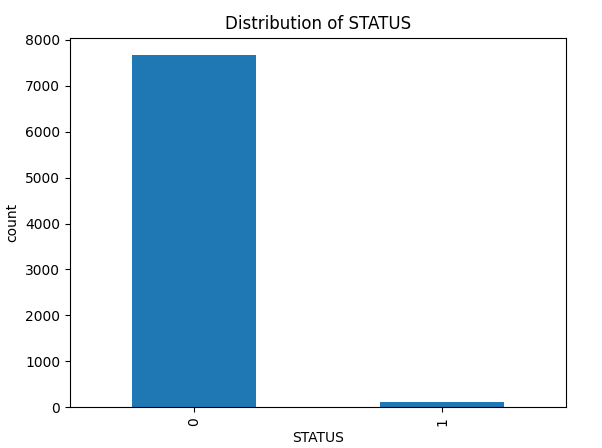


Correlation Plots:

* We can see a positive linear correlation between the family member and the children count. This makes sense, the more the children someone have, the larger the family member count. This is a multicollinearity problem. Meaning that the features are highly correlated. We will need to drop one of them.

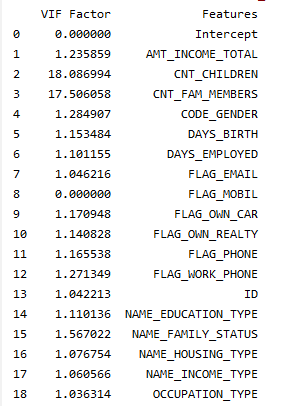
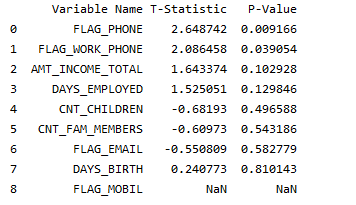


Checked the distribution of default and non-defaulted customers in the dataset, to check whether the dataset is imbalanced or balanced data set.



* PART B: VARIABLE IMPORTANCE – T TEST AND VIF

Performed Independent T-Test on each variable with 95% confidence level.

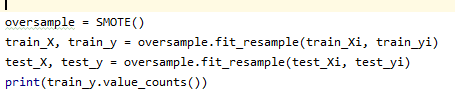


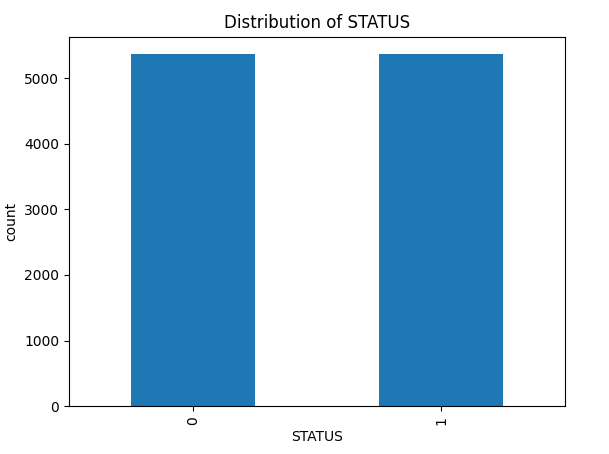
Applied VIF test to find the multicollinearity between the variables.

* PART C: MODEL BUILDING AND MODEL VALIDATION

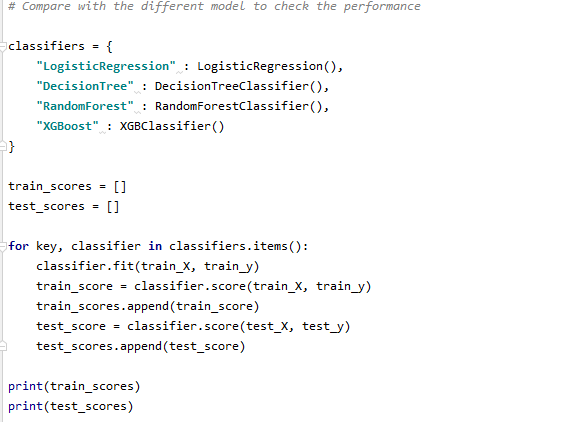
We have used the four model for the evaluation, model are Logistic Regression, DecisionTree, RandomForest, and XGBoost .

Our Dataset is highly unbalanced; hence we used the SMOTE technique to balance the dataset.





After models training and evaluation, found the below interpretations.

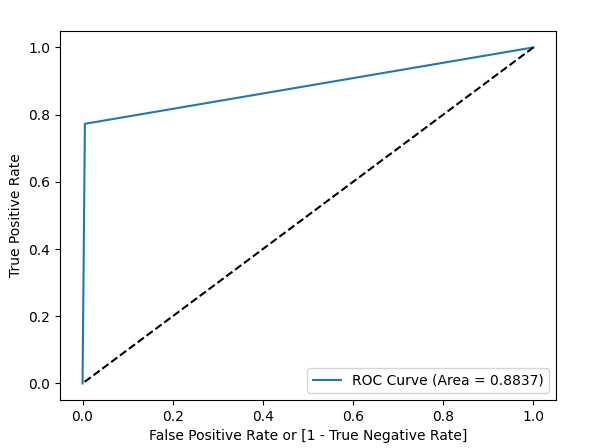
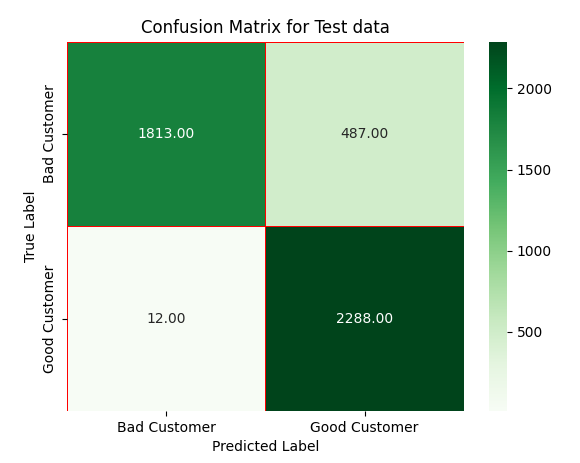




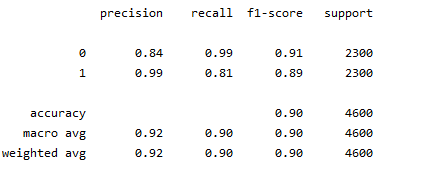
* We found out that XGBoost model is performing best on the train set as well as test set with 90% accuracy
* We will be using XGBoost to predict our values.

Evaluation of the test dataset:

ROC Curves and Confusion Matrix



Confusion Matrix:



Using the ROC curve and recall, we can conclude that our top model is XGBoost