**INFO 5082 PROJECT REPORT**

**SEMINAR IN RESEARCH AND RESEARCH METHODOLOGY**

**CREDIT CARD APPROVAL PREDICTIVE ANALYSIS**

**Submitted by**

**NAVYA SOLIPURAM**

**11374807**

**Introduction and Background:**

This is a second hand dataset available at Kaggle website. This dataset consists of 438558 rows and 18 columns with some empty spaces. This project includes the details of the customer represented with an ID. Dataset includes attributes such as ID, gender, owns any property, does he owns a car, marital status, number of children, Annual income and many more. This dataset helps us in predicting the customer for the approval of credit card. Each row represents the unique customer with their details. This project is based on few machine learning techniques. I have learnt machine learning in one of my course so I would like to implement those concepts which I have learnt and get more knowledge on these techniques. The reason why I had chosen this is that I am trying to get into field related to these so this project may help in facing the interview. As I have no experience this is the project which I can show up in the interview.

**Statement of the problem:**

This project stays at the aim to help the bank employees in choosing the customer to be approved for credit card or not. Based on the details of the customers and past history they decide whether to approve or not.

**Review of literature:**

Credit card fraud detection methods On doing the literature survey of various methods for fraud detection we come to the conclusion that to detect credit card fraud there are multiple approaches like.

Various sampling approaches have been proposed in the literature, with random oversampling of minority class cases and random under sampling of majority class cases being the simplest and most common in use; others include directed sampling The second problem in developing supervised models for fraud can arise from potentially undetected fraud transactions, leading to mislabeled cases in the data to be used for building the model.

If one of these or combination of algorithm is applied into bank credit card fraud detection system, the probability of fraud transactions can be predicted soon after credit card transactions by the banks.

**Objectives of the study:**

The objective of this study is to get various visualizations, prediction based on different parameters and choose which parameter would best give the accurate result. Different models have been implemented and based on few factors one model is finalized to be used for prediction of approval.

**Data Collection :**

The credit card approval prediction analysis Dataset is the second-hand data which is available at Kaggle. The dataset consists of 18 columns and more than 4,38,558 rows with a lot of outliers, empty spaces or columns and null values. I removed the null values using EDA process.

Dataset:

[Dataset1](https://www.kaggle.com/rikdifos/credit-card-approval-prediction)

[Dataset2](https://www.kaggle.com/rikdifos/credit-card-approval-prediction?select=credit_record.csv)

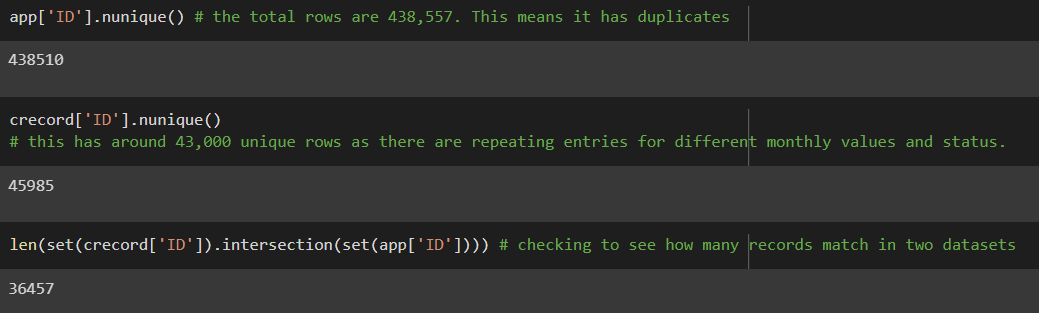
Data dictionary:

All the columns that I used from this dataset are shown below

|  |  |  |
| --- | --- | --- |
| Name | Input | Description |
| ID | Int | Unique ID of the customer |
| CODE\_GENDER | String | Gender of the customer (male/female) |
| FLAG\_OWN\_CAR | String | Customer having car or no |
| FLAG\_OWN\_REALTY | String | Do the customer own the property or no |
| CNT\_CHILDREN | Int | Number of children for the individual |
| AMT\_INCOME\_TOTAL | Int | Total income of the individual |
| NAME\_INCOME\_TYPE | String | Source of income |
| NAME\_EDUCATION\_TYPE | String | Educational qualification of the individual |
| NAME\_FAMILY\_STATUS | String | Marital status |
| NAME\_HOUSING\_TYPE | String | Type of house they live in (rented or own) |
| DAYS\_BIRTH | Int | Number of days the customer lived |
| DAYS\_EMPLOYED | Int | Number of days since they are working on it |
| FLAG\_MOBIL | Boolean | Individual having mobile or not |
| FLAG\_WORK\_PHONE | Boolean | Individual having work mobile or not |
| FLAG\_PHONE | Boolean | Individual having working phone number or not |
| FLAG\_EMAIL | Boolean | Individual having email or not |
| OCCUPATION\_TYPE | String | Occupation type of an individual |
| CNT\_FAM\_MEMBERS | Int | Number of family members of that individual |

**Exploratory data analysis (EDA) and Hypotheses for the Study**

Exploratory data analysis is an approach of analyzing the data sets as a quick summary of what is the dataset having in it. In my dataset firstly I have checked for the duplicates for both the datasets and found that I have many duplicates and also compared the two datasets for common records.

****

Checked for null values using heat map and found that occupation\_type has many null values. So I dropped this column as I don’t want any values to be approximate and wanted to be appropriate.

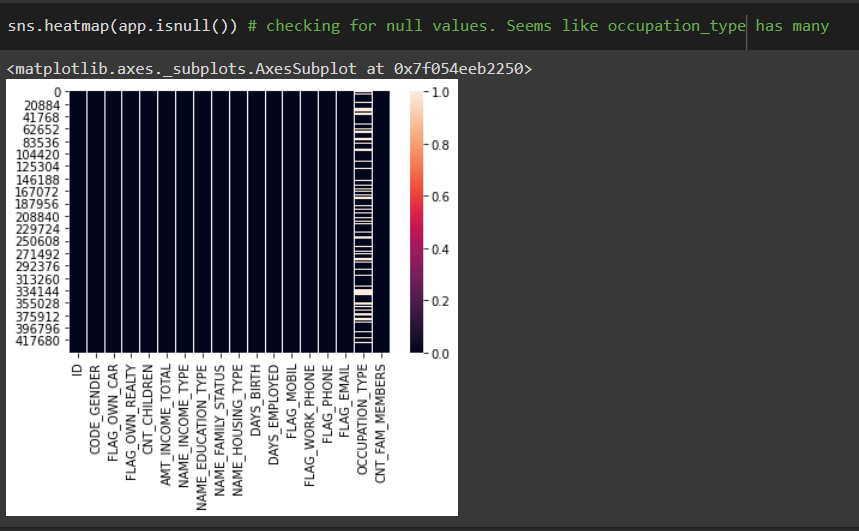
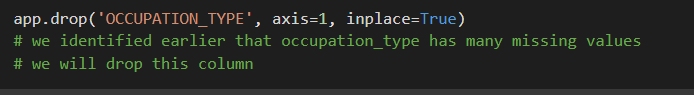
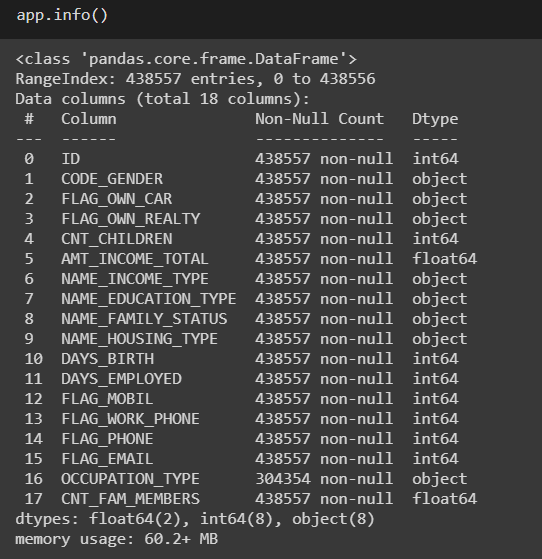


Fig1 Heatmap showing null values in the dataset

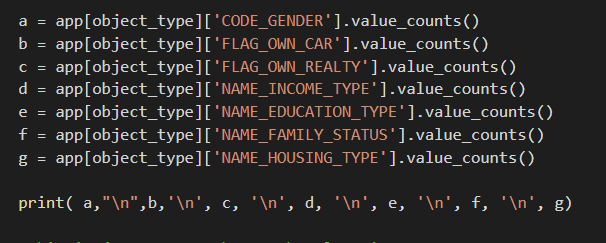


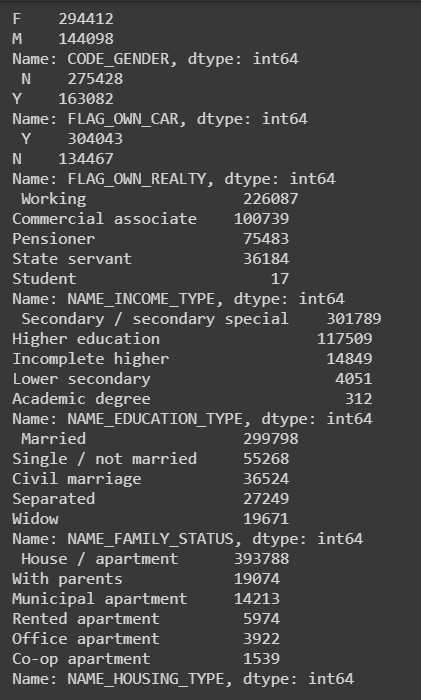
**Data analytics:**

Here is the summary of my dataset



I have displayed below the count for each and every attribute present in the dataset.





**Data visualizations and results:**

1. Type of occupation category most of the customers fall under:

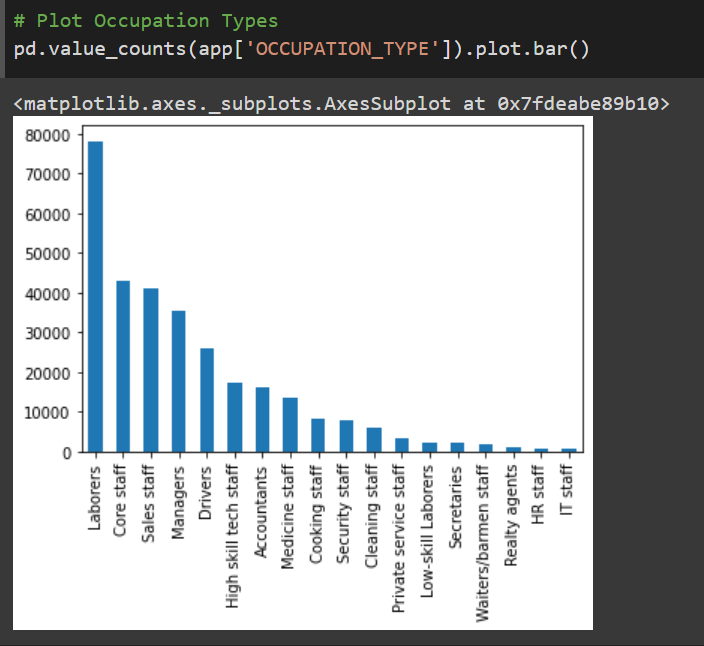
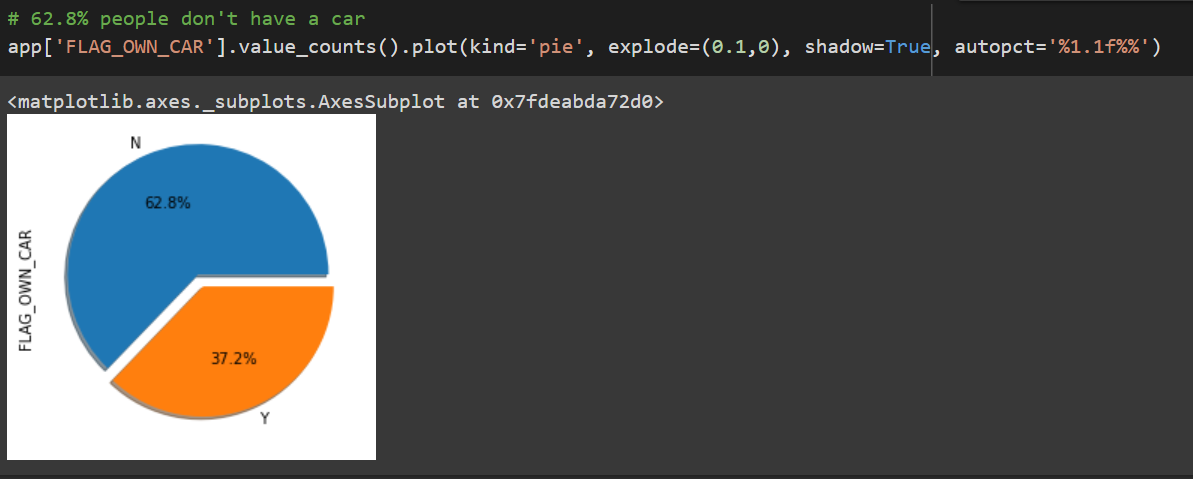


Fig 2 Bar graph showing count of different types of occupations

From the above bar graph we can say that most of the customers are laborers and we have 18 different occupation types. And we have very few IT staff. We have nearly 80000 laborers among 4 lakh customers which is almost more than 20%.

2. People who don’t own a car



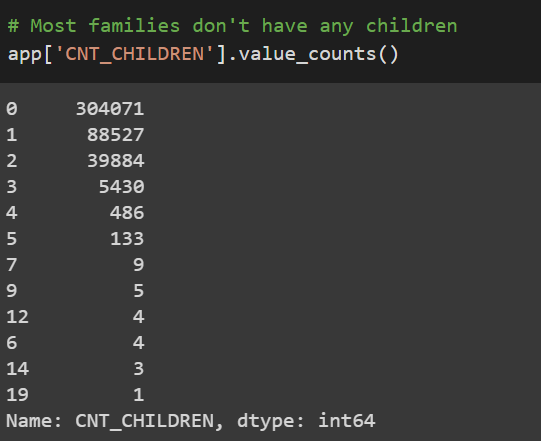
It can be seen that out of all the customers 62.8% people do not own a car. Which can be understood that they might not have enough money to buy a car as we saw that most of the customers are laborers. And only few of the total have a comfort that they have their own car.

3. Gender division



It can be seen that most of the customers are female. Among all 67.1% of them are females and 32.9% are males. By this we can say that most of the customers who applied for credit card are females.

4. Number of children for the customer



Among the customers whom I considered, most of them have no children and few have them have one or more than one children. 304071 customers have no children, 88527 have only one children, 39884 customers have 2 children, 5430 customers have 3 children, 486 customers have 4 children, 133 customers have 5 children, customers having more than 6 children are comparatively negligible.

5. categorizing customers based on their educational qualification.

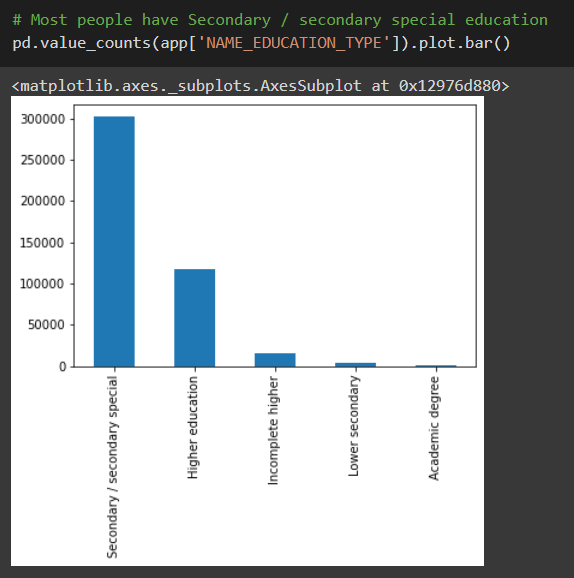


Fig 3 bar graph showing the educational details of the customer

From the above graph most of the customers have completed secondary education. Some of them have completed the higher education, and very few are done with their academic degree.

When I plotted the scatter plots I found few outliers in the columns, CNT\_CHILDREN, AMT\_INOCME\_TOTAL, CNT\_FAM\_MEMBERS. Not to effect the model implementations I have decided to remove the outliers

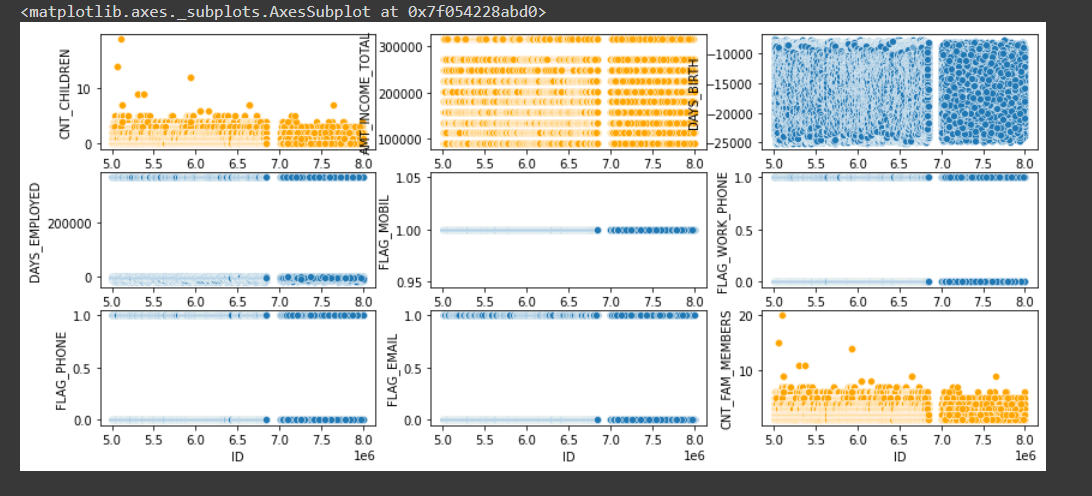


Fig 4 scatterplots showing the outliers

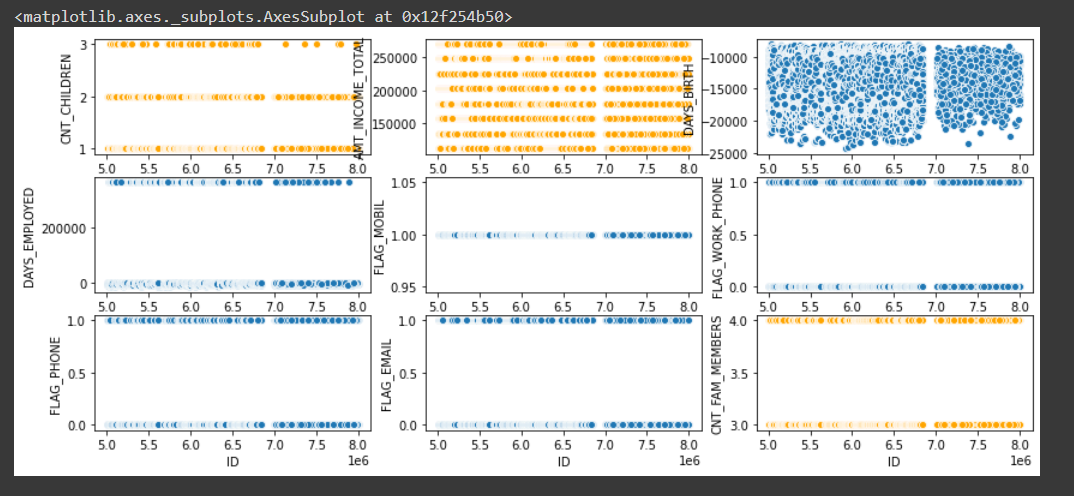
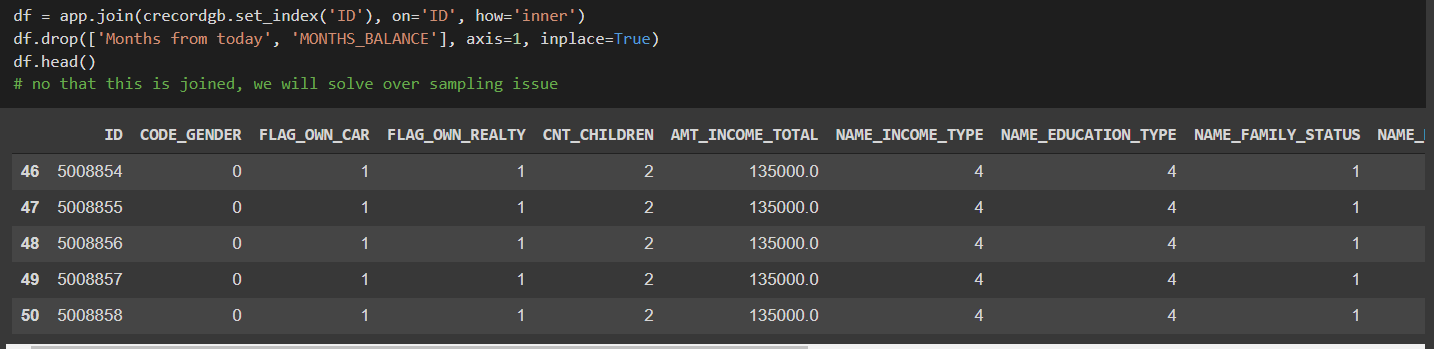
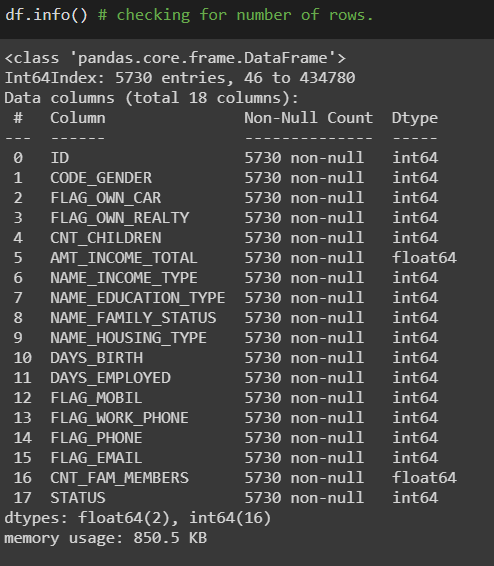


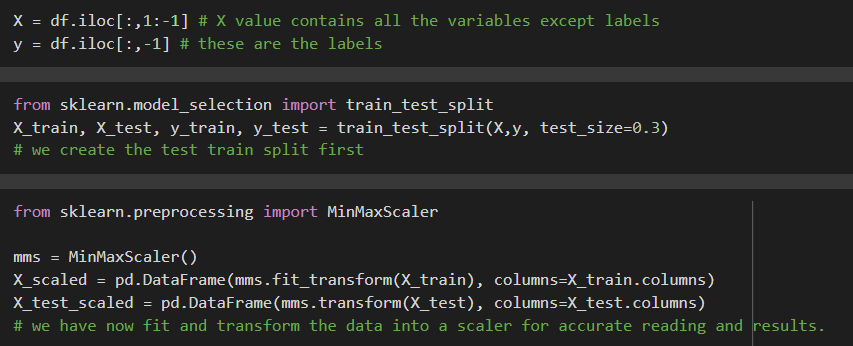
Fig 5 scatterplots after removing the outliers

Joined the two datasets which will solve the sampling issue based on the customer ID.



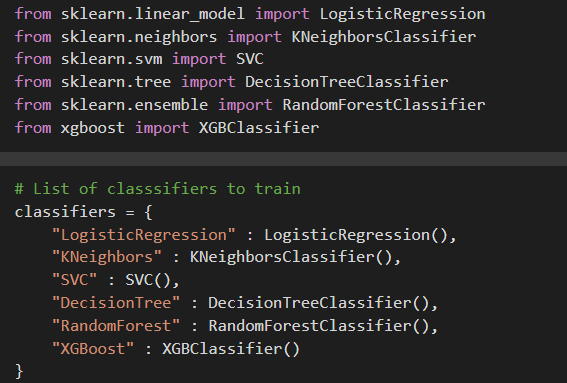
After all the changes implemented on the dataset we have got 5730 entries and 18 columns. Then divided the data into training and testing data. Then balanced the training and testing data sets.



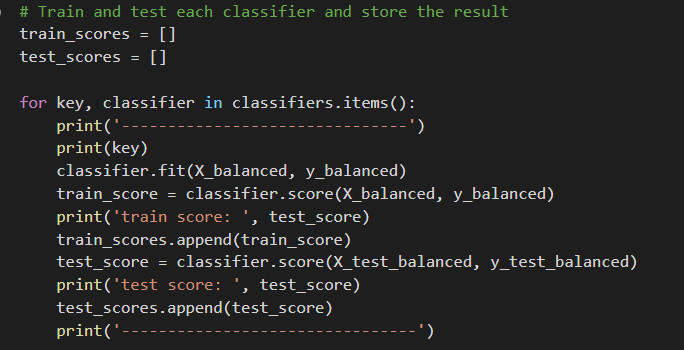


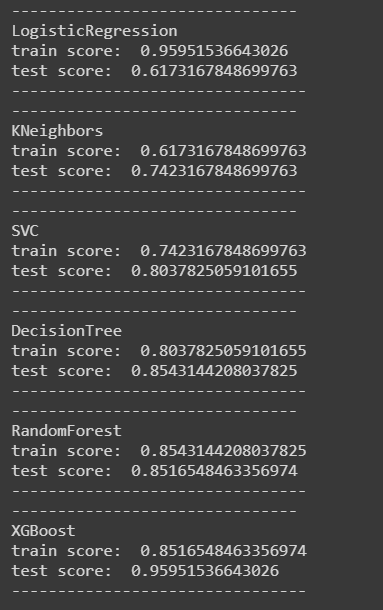
Then I have performed the analysis for those datasets using different models. I have used the following methods:

* Logistic regression
* KNeighbors
* SVC
* Decision tree
* Random forest
* XGBoost



We have got the results for training and testing datasets as below:





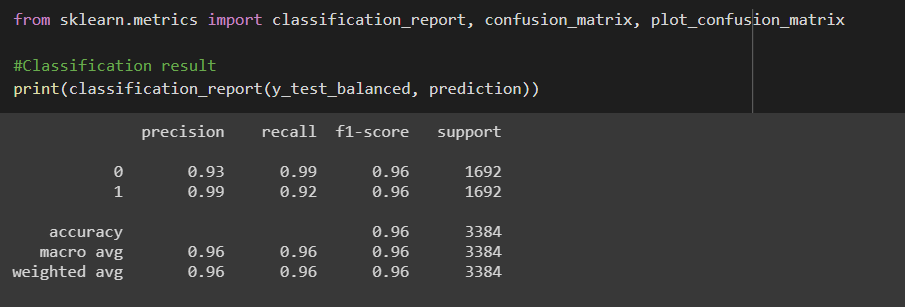
We can observe that XGBoost model is performing the best with 95% accuracy. So I have performed the analysis on XGBoost classifier.

XGBoost is a decision tree based machine learning algorithm which uses the gradient boosting framework. The advantages of implementing this model is that it is scalable and accurate for implementation.

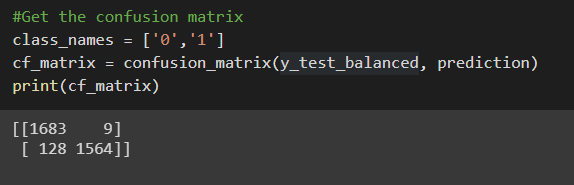
Ensembles are constructed from decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. This is a type of ensemble machine learning model referred to as boosting.

Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. This gives the technique its name, “gradient boosting,” as the loss gradient is minimized as the model is fit, much like a neural network.

Classification result is shown below. We can observe that f1-score is about 0.96



I also represented the confusion matrix for this model.

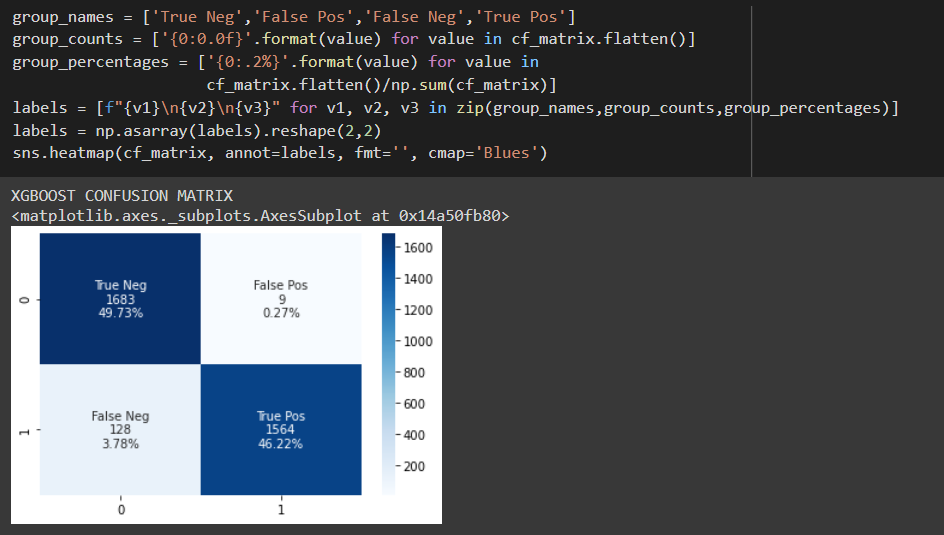


Confusion Matrix: Confusion matrix is a tabular summary of the number of correct and incorrect predictions made by a model.

As our classification outcomes are only 3, we get a confusion matrix of size 2\*2; Each Cell in our confusion matrix signifies one of the following scenarios.

What is the confusion matrix?

|  |  |
| --- | --- |
| Actual-pass,  predicted-fail  True negative | Actual - Fail,  Predicted - Success  False Positive |
| Actual - Success,  Predicted - Fail  False Negative | Actual - Success,  Predicted - Success  True Positive |



**Conclusion:**

This project gives us the useful details such as on what parameter bases the customer is approved for credit card and which customer can be given a credit card based on their details and history analysis. This mainly helps the manager who decides whether the customer can be approved for credit card or not.

In future we can likewise improve the exactness of our models by advancing various hyper parameters for the XGBoost models and furthermore to attempt some different models like KNN and SVM with an focus on accuracy, we can likewise improve our feature selection using regression models to discover which features have great connection with progress.

**Bibliography:**

1. Gui, L. (2019). Application of Machine Learning Algorithms in Predicting Credit Card Default Payment. Retrieved from escholarship.org website: <https://escholarship.org/uc/item/9zg7157q>

*2. Credit Card Approval Prediction*. (n.d.). Kaggle.com. <https://www.kaggle.com/rikdifos/credit-card-approval-prediction>

3. <https://www.researchgate.net/publication/328026972_Application_of_Machine_Learning_Algorithms_in_Credit_Card_Default_Payment_Prediction>

*4. [PDF] Review Paper on Credit Card Fraud Detection - Free Download PDF*. (n.d.). Silo.tips. Retrieved April 26, 2021, from

<https://silo.tips/download/review-paper-on-credit-card-fraud-detection>

*5. Credit Card Fraud Detection: A Systematic Review | Request PDF*. (n.d.). ResearchGate. Retrieved April 26, 2021, from <https://www.researchgate.net/publication/338352966_Credit_Card_Fraud_Detection_A_Systematic_Review>

<https://ijarcce.com/wp-content/uploads/2012/03/IJARCCE2A-s-malvika-Survey-on-Credit-Card-Fraud-final.pdf>

‌