**PROJECT SUMMARY**

|  |  |
| --- | --- |
| Batch Details | **DSE – AUGUST 2022 (GURGAON)** |
| Team Members | Aekansh Singh, Deepak Pareek, Harsh Vardhan Mudgal, Karan Saxena, Sourav Rawat, Yashveen Sangwan |
| Domain of Project | Banking (Finance) |
| Proposed project title | Credit Card Approval |
| Group number | Team 5 |
| Team Leader | Sourav Rawat |
| Mentor Name | Pratik Sonar |

**Date: 23/12/2022**

**Sourav Rawat**

**Signature of the Mentor Signature of the Team Leader**

TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **Sl NO** | **Topic** | **Page No.** |
| 1 | Overview | 1-3 |
| 2 | Business Problem Statement | 4 |
| 3 | Data Description | 5 |
| 4 | Exploratory Data Analysis | 5-6 |
| 5 | Feature Selection/Engineering/Extraction | 7-8 |
| 6 | Basic Modelling | 9 |
| 7 | Model Comparison | 10-13 |
| 8 | Conclusion and recommendations | 14-15 |
| 9 | References | 15 |

**Project Details**

**OVERVIEW**

Credit risk as the board in banks basically centers around deciding the probability of a customer's default or credit decay and how expensive it will end up being assuming it happens. It is important to consider major factors and predict beforehand the probability of consumers defaulting given their conditions.

Which is where a machine learning model comes in handy and allows the banks and major financial institutions to predict whether the customer, they are giving the loan to, will default or not. This project builds a machine learning model with the best accuracy possible using python.

Banks receive a lot of credit card applications. Many of the applications do not get approved for a variety of reasons, like increased loan balances or poor-income levels.

Manually analyzing these applications can be very time-consuming and full of human errors. Thankfully, we can automate this task with the help of machine learning. Below are the concepts and theories that helped understand the project solution and are an integral part of this process. A thorough understanding of them facilitated the development process.

**Business problem statement (GOALS)**

1.What would you achieve by this project ?

Studying past data provided by bank of the applicants, understanding the dependency of these features onto prediction of approval of an application by deploying classification machine learning algorithms striving for a higher accuracy in prediction.

2.How would this help the business or clients ?

By implementing these models (Logistic Regression, Random Forest, Decision tree, XGBoost) also, we will try to increase the accuracy using ensemble techniques. Bank would be able achieve better accuracy in credit card approval and maximizing the revenue for themselves.

The solution here is a Classification based Machine Learning model. It can be implemented by different classification algorithms (like Logistic Regression, Random Forest, Decision tree, XGBoost and so on. Here first we will be performing Data pre-processing step, in which Data Profiling, feature engineering, feature selection, feature scaling, VIF steps are performed and then we are going to build model.

3.What is the further scope of the project ?

Automated web based products on credit card processing for the finance sector.

4.Limitation of the project

Our model is not a generalized model but a business oriented, realistic and specific model -meaning needs to be calibrated and updated with time and business

**DATA DESCRIPTION**

| Feature Name | Description | Remarks |
| --- | --- | --- |
| ID | Client Number |  |
| CODE\_GENDER | Gender |  |
| FLAG*OWN*CAR | Is there a car |  |
| FLAG*OWN*REALTY | Is there a property |  |
| CNT\_CHILDREN | Number of Children |  |
| AMT*INCOME*TOTAL | Annual Income |  |
| NAME*EDUCATION*TYPE | Education Level |  |
| NAME*FAMILY*STATUS | Marital Status |  |
| NAME*HOUSING*TYPE | Way of Living |  |
| DAYS\_BIRTH | Age in days |  |
| DAYS\_EMPLOYED | Duration of work in days |  |
| FLAG\_MOBIL | Is there a mobile phone |  |
| FLAG*WORK*PHONE | Is there a work phone |  |
| FLAG\_PHONE | Is there a phone |  |
| FLAG\_EMAIL | Is there an email |  |
| JOB | Job |  |
| BEGIN\_MONTHS | Record month | The month of the extracted data is the starting point, backwards, 0 is the current month, -1 is the previous month, and so on |
| STATUS | Status | 0: 1-29 days past due 1: 30-59 days past due 2: 60-89 days overdue 3: 90-119 days overdue 4: 120-149 days overdue 5: Overdue or bad debts, write-offs for more than 150 days C: paid off that month X: No loan for the month |
| TARGET | Target | Risk user are marked as '1', else are '0' |

Dataset consists of the historical data around the credit card approval. It has 19 features and 5,37,667 observations.

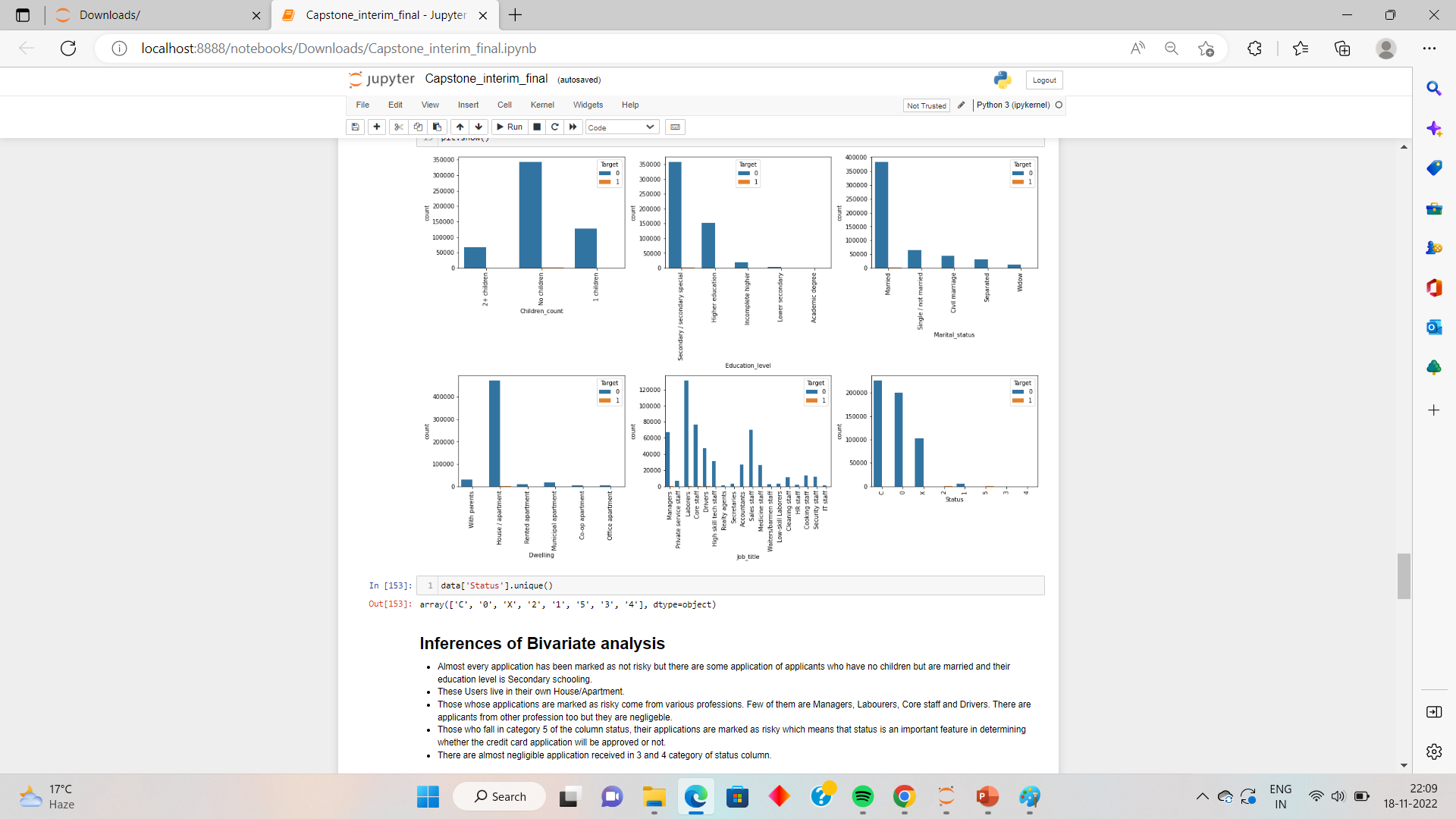
**EXPLORATORY DATA ANALYSIS**

CORRELATION MATRIX-



* The columns which has No and Yes will be converted to 0's and 1's respectively.
* **Univariate Analysis:** Income is Positively skewed which means there are outliers in the data. Age, Years\_employed and Months\_applied all are positively skewed.
* **Bivariate Analysis:** Those who fall in category 5 of the column status, their applications are marked as risky which means that status is an important feature in determining whether the credit card application will be approved or not.
* **Outliers Treatment:** There are outliers in Income and Years\_employed. We treated outliers using IQR.
* **One Hot Encoding:** Encoded the education\_type into higher studies and lower studies, Marital\_status into married, single and others and Dwelling into independent and dependent.
* **INERENCES FOR UNIVARIATE ANALYSIS-**
* 1. Majority of the applicants have no children, followed by 1 child and 2 children.
* 2. Most of the applicants have 'Secondary/secondary special' as education level, followed by Higher education and incomplete higher education which means many of them are high school dropouts.
* 3. A large number of applicants are Living in their own House/Apartment. There are a few who are residing with their parents and in Municipal apartments followed by rented apartments.
* 4. Most of the applicants are laborer, followed by Drivers and sales staff. There are other professions too that have a significant number of applicants.
* 5. Most of the applicants are Married followed by single or unmarried.
* 6. Most of the applicants fall in the C Category of column status which means that they have paid there debt in the same month they took it. There are significant number of applicants too who have no debts on them so they fall in X category.
* 7. almost half the applicants have debts pending which are due for 1-29 days.
* 8. There are certain applicants whose debts have been considered as bad debts by the banks.

**BIVARIATE ANALYSIS-**



* Almost every application has been marked as not risky but there are some application of applicants who have no children but are married and their education level is Secondary schooling.
* These Users live in their own House/Apartment.
* Those whose applications are marked as risky come from various professions. Few of them are Managers, Laborer, Core staff and Drivers. There are applicants from other profession too but they are negligible.
* Those who fall in category 5 of the column status, their applications are marked as risky which means that status is an important feature in determining whether the credit card application will be approved or not.
* There are almost negligible application received in 3 and 4 category of status column.

**FEATURE EXTRACTION/SELECTION/ENGINEERING-**

1. In JOB feature, we have encoded the categories into 2 categories - Blue collar and White collar jobs
2. In Name\_Housing\_type feature, we have encoded the categories into 2 categories-

Dependent and non-dependent.

1. In Name\_Family\_Status, we have encoded the categories into 3 categories as Married, Single and Others. Where others are:- Divorced or Separated applicants.
2. In Name\_Education\_Type, we have encoded the catgeories into 2 categories- Higher education and lower education.
3. In CNT\_CHILDREN, Since there are only 3 unique values no feature extraction has been done.

**SCALING-**

**ROBUST SCALER-**

Scale features using statistics that are robust to outliers.

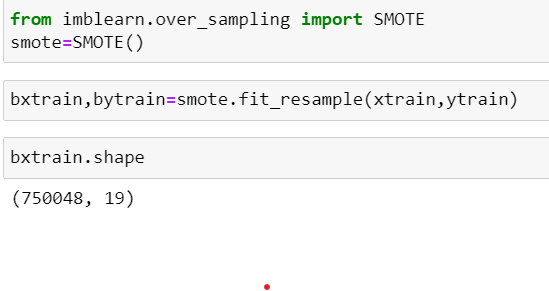
This Scaler removes the median and scales the data according to the quantile range (defaults to IQR: Interquartile Range). The IQR is the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile).

Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Median and interquartile range are then stored to be used on later data using the [**transform**](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html#sklearn.preprocessing.RobustScaler.transform) method.

Standardization of a dataset is a common requirement for many machine learning estimators. Typically this is done by removing the mean and scaling to unit variance. However, outliers can often influence the sample mean / variance in a negative way. In such cases, the median and the interquartile range often give better results.

**USING SMOTE()-**

* Applying SMOTE target variable is highly imbalanced.



**WHAT IS KAPPA SCORE, ACCURACY SCORE,F1 SCORE, RECALL AND PRECISION?**

Cohen Kappa Score is a statistic used to measure the agreement between two raters. It can be used to calculate how much agreement there is between raters on a scale from 0-1, with 1 being perfect agreement and 0 being no agreement at all. The higher the score, the more agreement there is between the raters.

.Accuracy classification score. In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels in y\_true.

F1-Score (F-measure) is an evaluation metric, that is used to express the performance of the machine learning model (or classifier). It gives the combined information about the precision and recall of a model. This means a high F1-score indicates a high value for both recall and precision.

Precision and recall are two numbers which together are used to evaluate the performance of classification or information retrieval systems. Precision is defined as the fraction of relevant instances among all retrieved instances. Recall, sometimes referred to as ‘sensitivity, is the fraction of retrieved instances among all relevant instances.

**WHAT IS ROC AND AUV CURVE AND WHY DO WE USE THEM?**

ROC or Area Under Curve/AUC helps us address the problems we face during classification. When checking or visualizing how different classifications of a model are performing, we use these metrics or curves to evaluate the outcome. ROC is short for Receiver Operating Characteristics, and AUC is the Area Under the Curve.

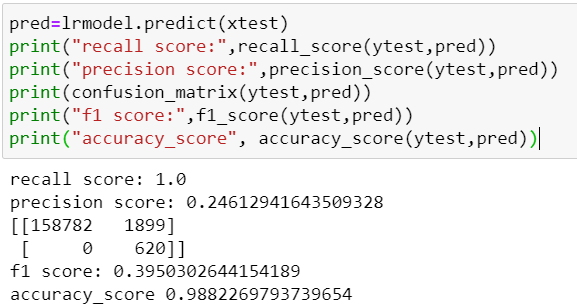
**WHY XGBOOST IS THE BEST MODEL?**

Gradient Boosting is improvised to make it Extreme. XGBoost is a tree based ensemble machine learning algorithm which is a scalable machine learning system for tree boosting. XGBoost stands for Extreme Gradient Boosting. It uses more accurate approximations to find the best tree model.

**BASE MODEL-**

**LOGISTIC REGRESSION-**

* Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable.

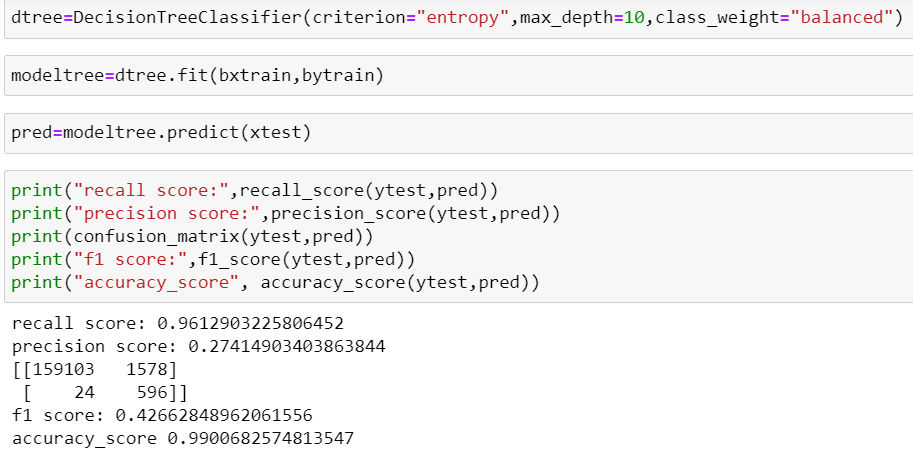


The accuracy of Logistic Regression is 98%.

**MODEL COMPARISON-**

**DECISION TREE CLASSIFIER-**

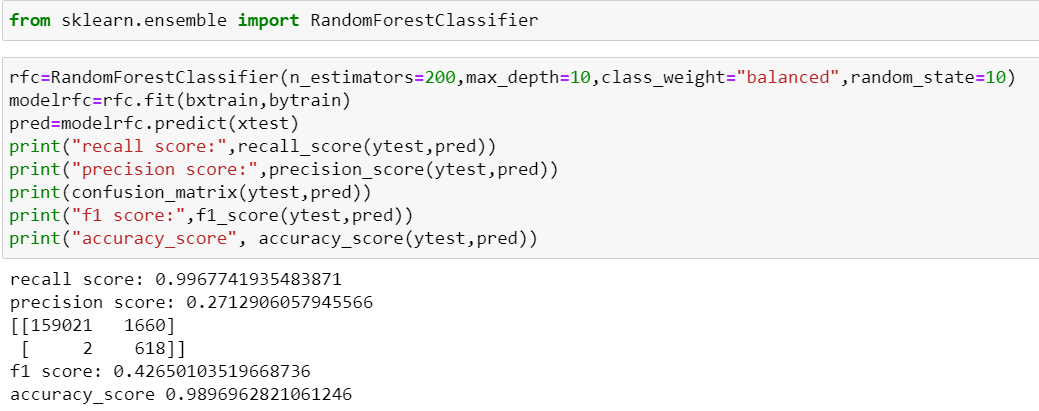
* Decision Tree is a **Supervised learning technique**that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where**internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
* In a Decision tree, there are two nodes, which are the **Decision Node** and**Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.



The accuracy is 99% in Decision Tree Modelling.

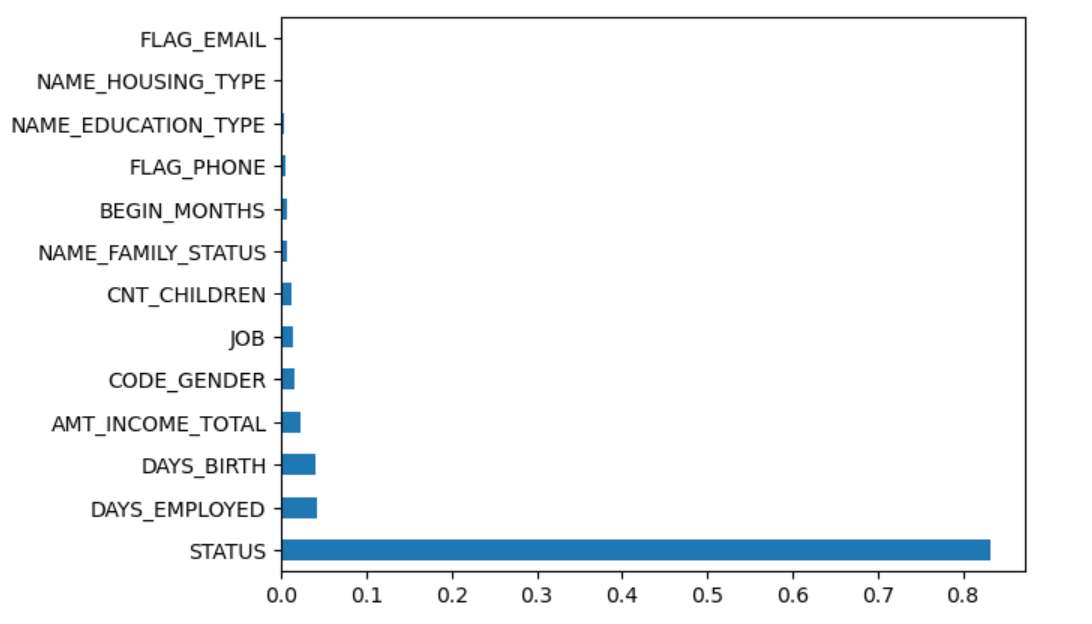
**RANDOM FOREST CLASSIFIER-**

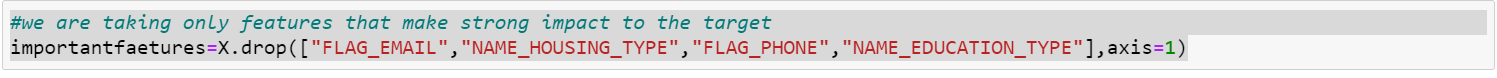
* Random Forest Classifier. Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. In Laymen’s term, So finally, it predicts based on the majority of votes from each of the decision trees made.



The accuracy in this model is 98%.

**EXTRACTING IMPORT FEATURES USING RFE-**

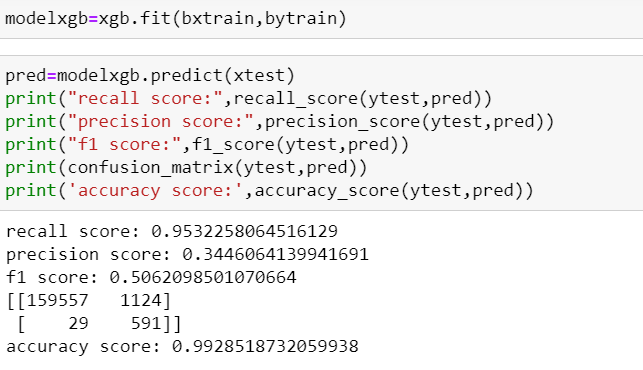




**XGB CLASSIFIER-**

* Classification Example with XGBClassifier in Python. The XGBoost stands for eXtreme Gradient Boosting, which is a boosting algorithm based on gradient boosted decision trees algorithm. XGBoost applies a better regularization technique to reduce overfitting, and it is one of the differences from the gradient boosting.

The accuracy with XGB Classifier is 99%

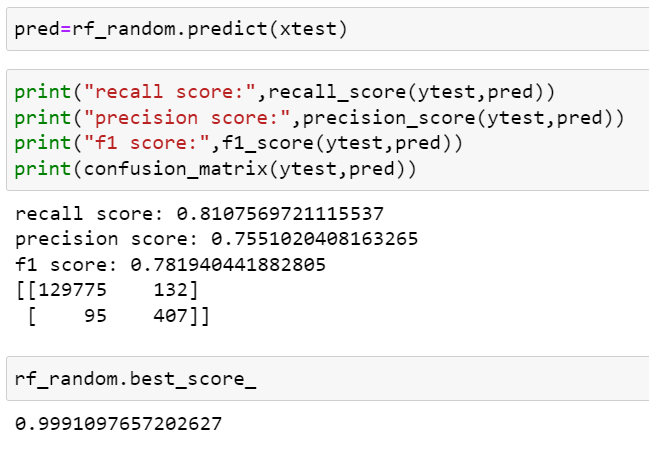


**USING HYPERPARAMETER TUNING-**

**RANDOMIZED SEARCH CV-**

Randomized search on hyper parameters.

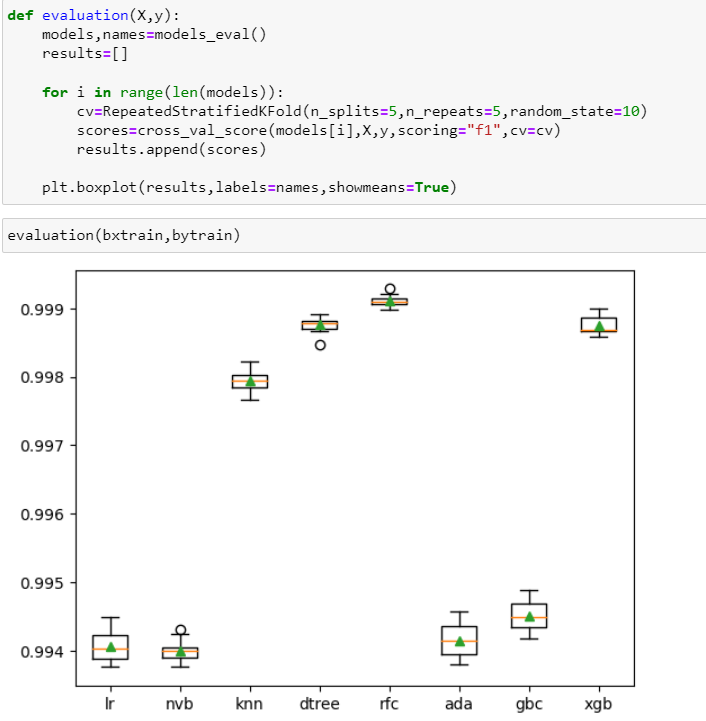
RandomizedSearchCV implements a “fit” and a “score” method. It also implements “score\_samples”, “predict”, “predict\_proba”, “decision\_function”, “transform” and “inverse\_transform” if they are implemented in the estimator used.The parameters of the estimator used to apply these methods are optimized by cross-validated search over parameter settings.

The f1 score of the model is 78% and the best score comes out to be 99%.

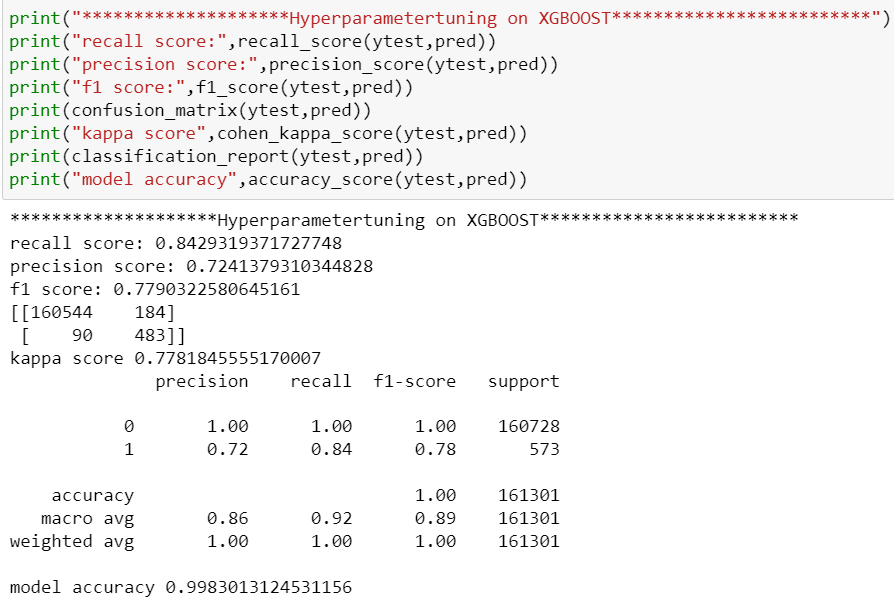
**CONCLUSION-**

**FINAL MODEL –**

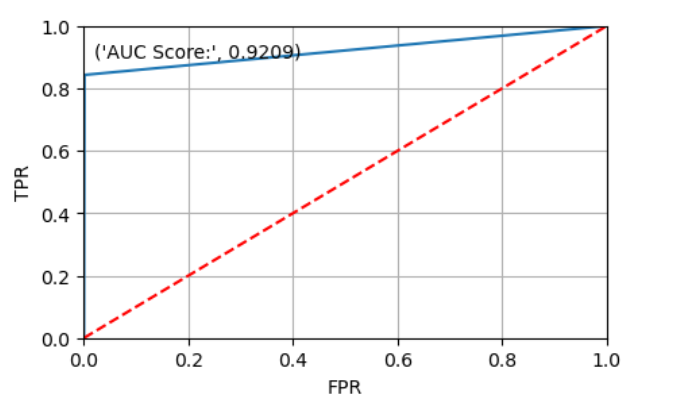
Defining a function to check the best fitted model



XGBoost with HyperParameter Tuning –



**The accuracy is the highest with the XGBoost Classifier HyperParameter Tuning Model with 99.8%.**

****

**REFERENCES –**

1. <https://www.kaggle.com/datasets/laotse/credit-card-approval>

2.S. B. Kotsiantis. Supervised Machine Learning: A Review of Classification Techniques. Informatica 31 (2007) 249-268 *Web*

3. Dataset, UCI http://archive.ics.uci.edu/ml/datasets/credit+approval

4. http://www.cs.stir.ac.uk/courses/ITNP4B/lectures/kms/4-MLP.pdf

5.http://localhost:8888/notebooks/Downloads/MajorProjectBM/Predicting%20Credit%20Card%20Approvals/notebook.ipynb

6. Medium. 2021. *Credit Card Approval Prediction Model in Python*. [online] Available at: <https://medium.com/@ashish.tripathi1207/credit-card-approval-prediction-model-in-python-c0e07677058e> [Accessed 29 June 2021].

7. http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

8. Http://rstudio-pubs-static.s3.amazonaws.com/73039\_9946de135c0a49daa7a0a9eda4a67a72.html