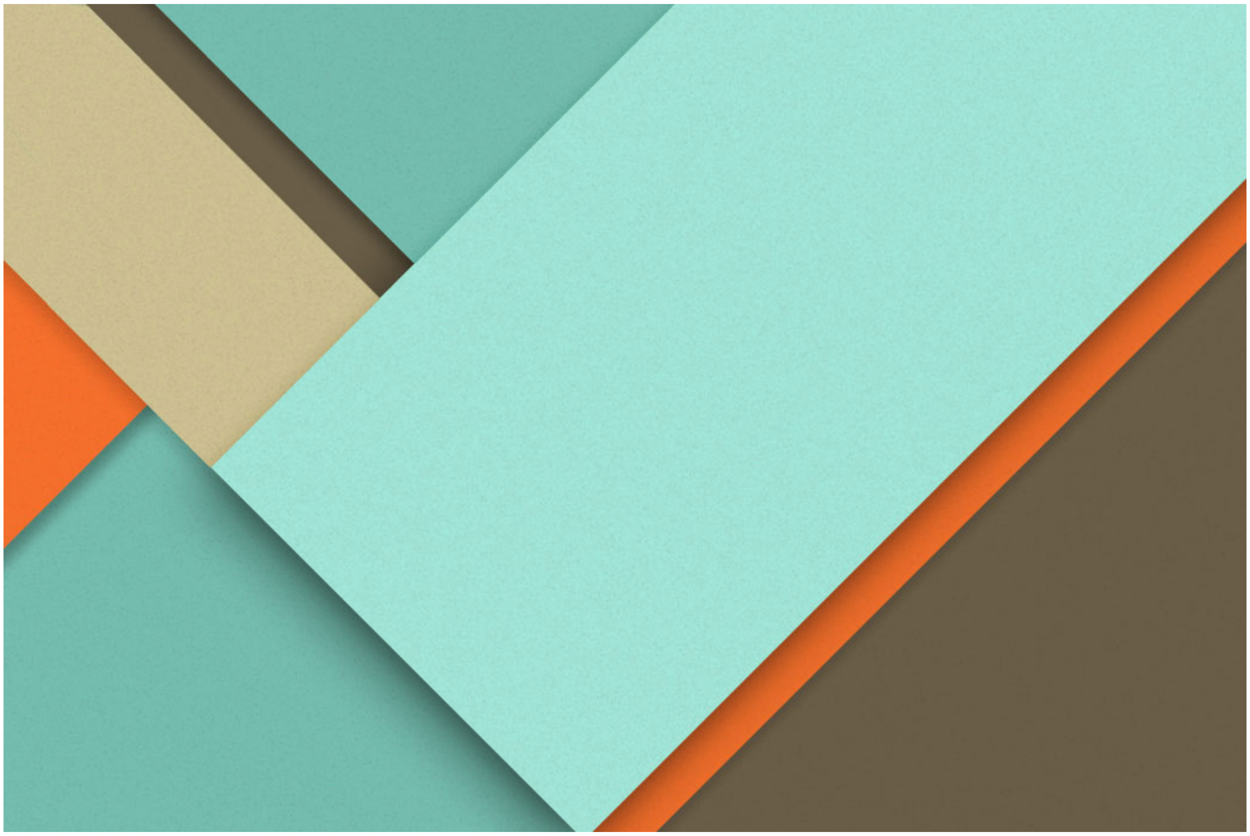


Classification Model

Problem - 2 of Data Science Assignment

07.09.2019

—

Pramod Yadav
IIIT Sri city ,A.P.

Problem Statement

1 . Create a model to predict the active good standing customer on dataset provided.

Goals

1. Model a Binary Classifier which predicts whether the customer is “Active Good Standing Customer ” -Label : 1 or “Active Bad Customer” - Label : 0.

Procedure

Following are the steps involved in solving the assignment -

1. Exploratory Data Analysis
2. Data Preparation
3. Training and fine tuning the model
4. Testing
5. Graphs

Note: The notebook links for all of them are mentioned under Links section at the end.

Description of the Steps involved :

I. Exploratory Data Analysis

In this section , the focus was on getting to know the story conveyed by the data through its attributes and what do they represent .

Some of the steps involved were - plotting distribution plots , treating missing values .

Following were the analysis in brief -

a) There were some columns like ‘Linkedin Verified’ , ‘facebook_verified’ , ‘EPFO Verification Status’ which had high percentage of missing data and were dropped.

b) Dropped descriptive features like ‘application reason ’ etc.

c) Datatype of columns were checked and it comprised of Datetime , object , Float.

d) Separation of the columns on their data type for numeric analysis and encoding .

e) Plot distribution of numeric features - most of the distributions were skewed.

f) Apply appropriate transformation on the column data.

g) Analyse non-numeric columns to derive meaning insights and prepare them for feeding to the model . fior eg .Dob data can be converted into age of the customer .

h) Conversion to float values like 'Amount' which were stored as an object .

II. Data Preparation

Following the insights obtained from the above analysis , the data was prepared and stored into a new file called 'new_processed.csv'

Following were the major steps :

a) For numeric columns - > Power Transformation followed by filling missing values and then z-score normalization

b) Conversion of datatypes of columns like - Amount , Term in Float from object type.

c) Treatment of Datetime columns - convert them in months using today as reference point.

d) Encoding certain columns and making them a categorical variable .

III. Training and Fine tuning the model

1 . The following models were used and implementation was done with scikit-learn -

a) K Nearest Neighbor Classifier

b) Support Vector Classifier (SVC)

c) Logistic Regression

d) Random Forest

2 . The dataset was split into Training and testing (20 %) .

3 . The hyper parameters were tuned by using Grid Search .

4 . After obtaining the best parameters the model was fitted using them.

IV. Testing and Results

1 . After training each model was fed with test data to know the generalization ability of the model .

2 . The dataset was imbalanced in ratio of nearly 5:1 in favor of “Active Good Customers” which was label - 1 as compared to “Active Bad Customers” as Label 0.

3 . The following are considered as evaluation metric as the dataset was imbalanced-

a) Confusion Matrix and other terms derivable from it
b) ROC - AUC score

4 . Every model achieved nearly 1.0 as roc-auc score on the test set .

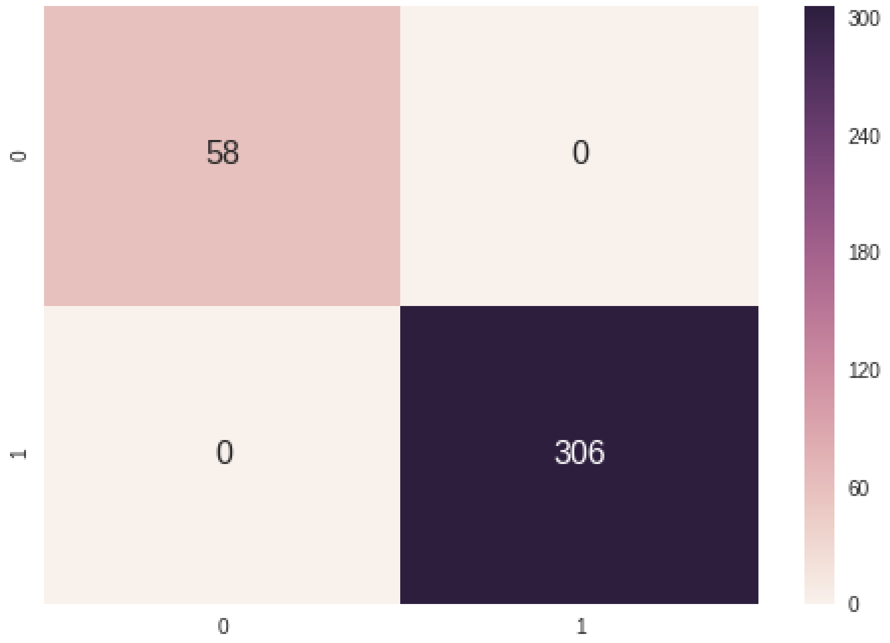
5 . Following tabulates the performance for each model.

Model	ROC - AUC Score	No of mis-classifieds
KNN	1.0	0
SVC	0.9999428538773644	2
Logistic Regression	1.0	0
Random Forest	1.0	0

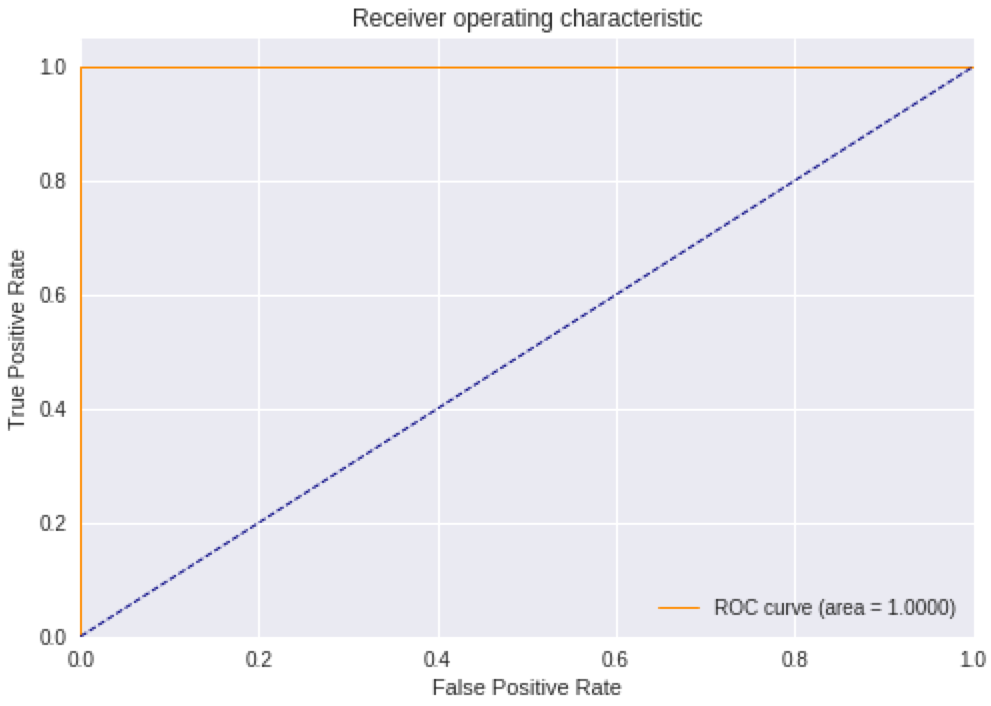
V. Graphs and Confusion Matrix

Following highlights the graphs , plots obtained for each model

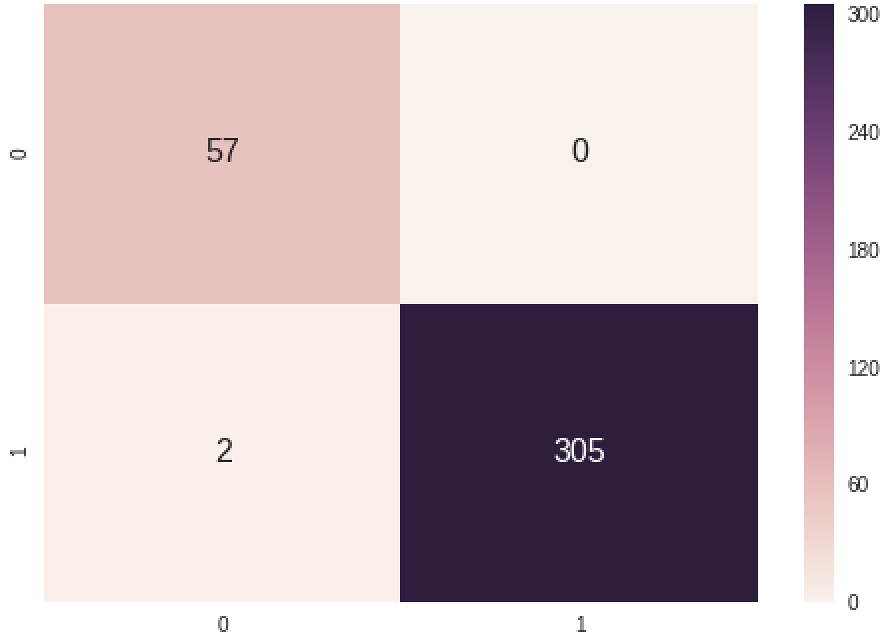
a) K Nearest Neighbor - (n_nieghbors = 2)



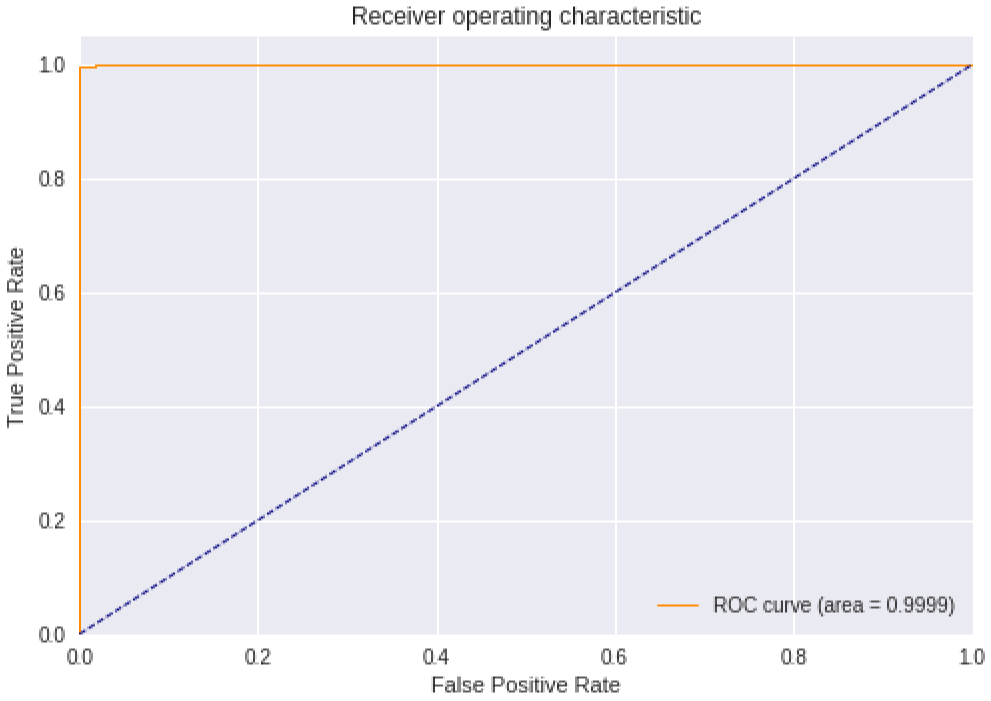
Actual labels (X axis) ->



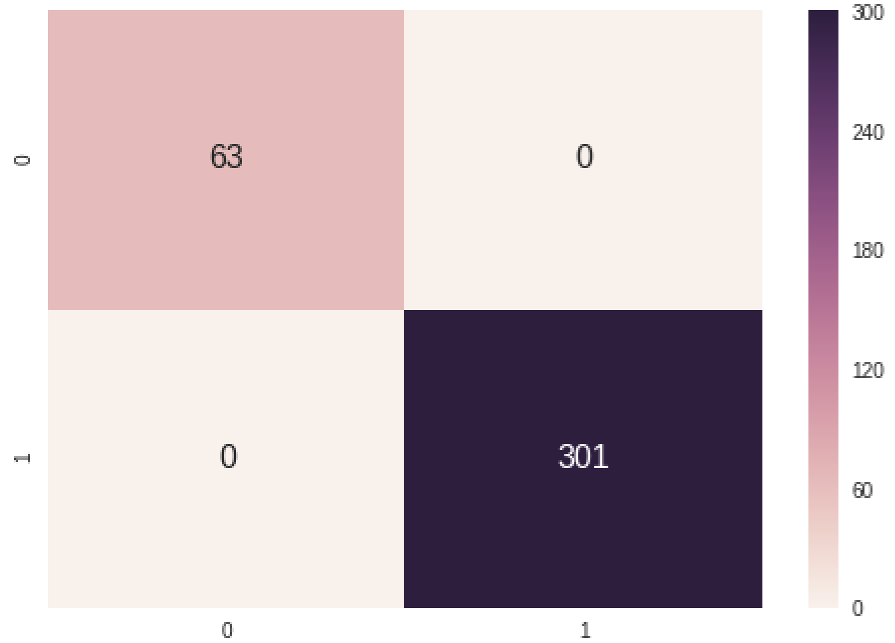
b) SVC



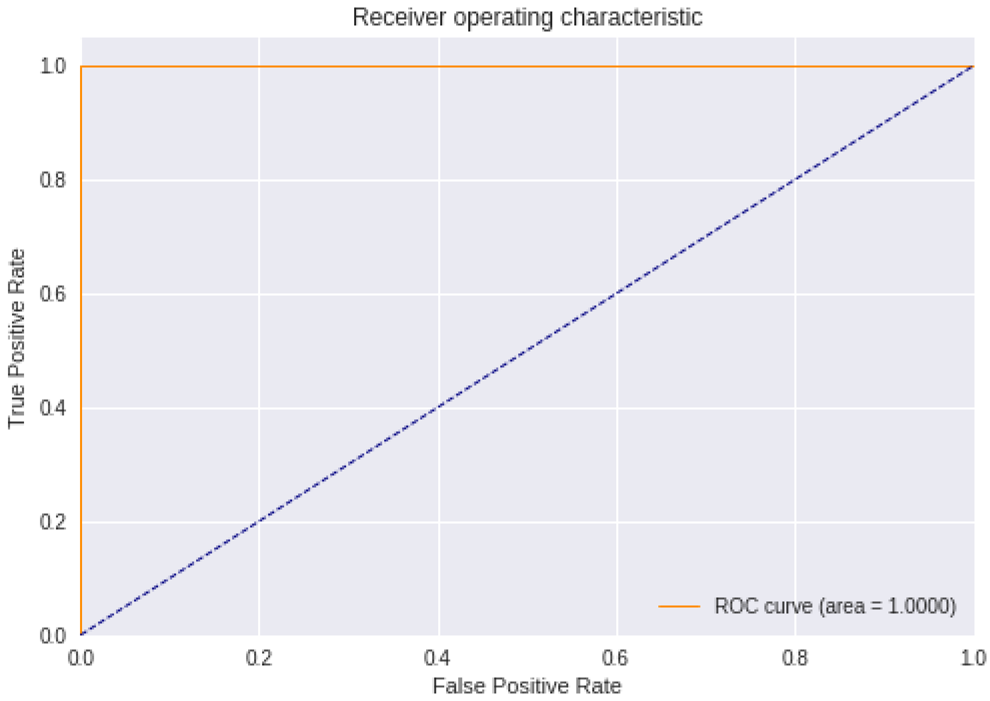
Actual Labels (X axis) ->



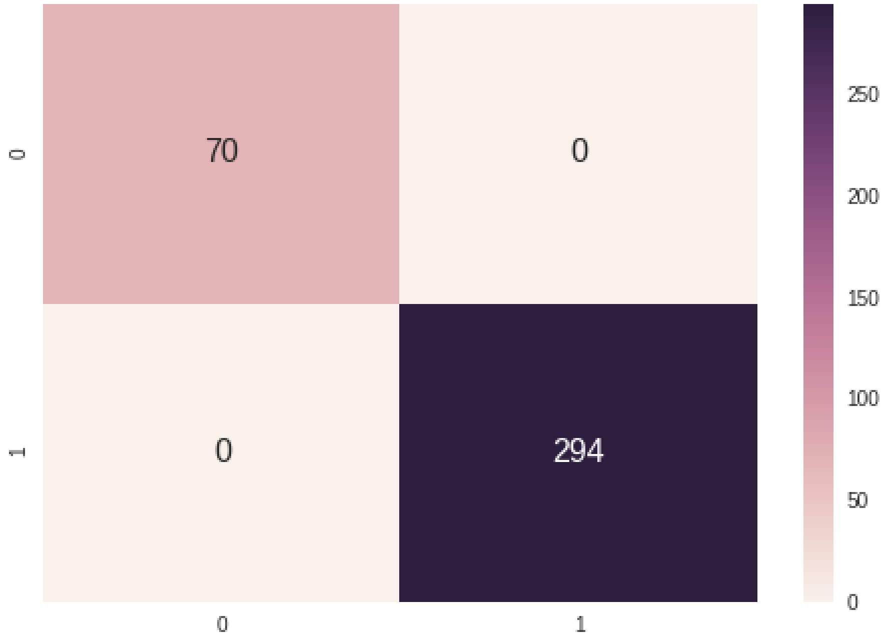
c) Logistic Regression



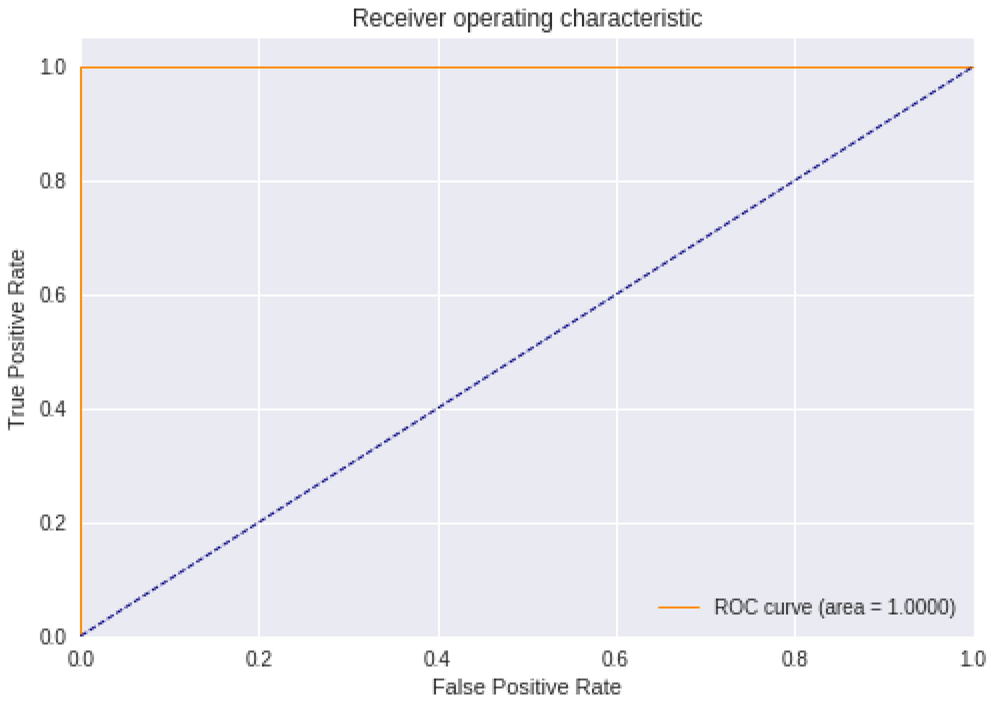
Actual Labels (X axis) ->



d) Random Forest



Actual Labels (X axis)->



Links :

The following links can be used to display jupyter notebook on https://nbviewer.jupyter.org/

(it also contains Colab links).

1 . EDA :
https://github.com/pramod1997/submission_data_science/blob/master/EDA.ipynb

2 . Data Preparation :
https://github.com/pramod1997/submission_data_science/blob/master/Data_prep.ipynb

3 . KNN :

https://github.com/pramod1997/submission_data_science/blob/master/KNN.ipynb

4 . Logistic Regression :
https://github.com/pramod1997/submission_data_science/blob/master/Logistic.ipynb

5 . SVC :
https://github.com/pramod1997/submission_data_science/blob/master/SVM.ipynb

6 . Random Forest :
https://github.com/pramod1997/submission_data_science/blob/master/RandomForest.ipynb