Kaggle Report

My predictions on "**Find the secret binary outcome**" using different models

Kaggle username	: Aparna Aidith
Private leaderboard score	: 0.89235
Public leaderboard score	: 0.87934

Data Handling

- Importing Data with Pandas
- Cleaning Data
- Exploring Data through Visualizations with Matplotlib

Data Analysis

- Logistic Regression Model
- Linear Regression Model
- XGBRegressor Model
- XGBClassifier Model
- Ridge Model
- Lasso Model
- RandomForest

Valuation of the Analysis

- Cross validation to valuate results locally
- Output the results from the Jupyter Notebook to Kaggle

Common Steps followed

Step 1 : Imported all the necessary tools and standard libraries.

```
In [1]: # This Python 3 environment comes with many helpful analytics libraries installed
        # It is defined by the kaggle/python docker image: https://github.com/kaggle/docker-python
        # For example, here's several helpful packages to load in

        #Lets start by import some of our standard tools
        import numpy as np # linear algebra
        import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
        import matplotlib
        import matplotlib.pyplot as plt
        from scipy.stats import skew
        from scipy.stats.stats import pearsonr

        %matplotlib inline
```

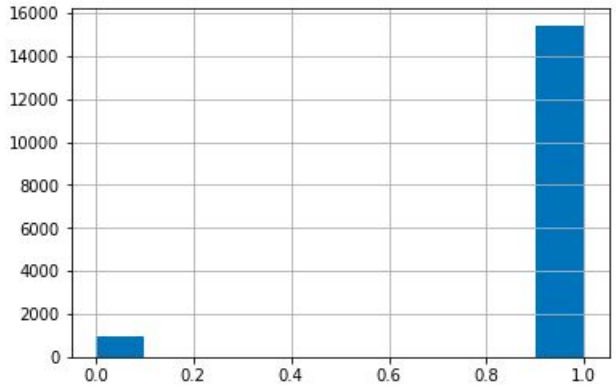
```
In [2]: #Lets read the data
        train = pd.read_csv("/Users/aparnaaidith/Desktop/Kaggle/all (1)/train_final.csv")
        test = pd.read_csv("/Users/aparnaaidith/Desktop/Kaggle/all (1)/test_final.csv")
```

Step 2 : The dataset given was supervised ,so it clearly had Y label in it. I checked whether the target variable is binary by plotting a histogram .The plot clearly showed the target variable is a binary.

Checking that your target variable is binary

```
In [3]: train.Y.hist()
```

```
Out[3]: <matplotlib.axes._subplots.AxesSubplot at 0x1176a1c50>
```

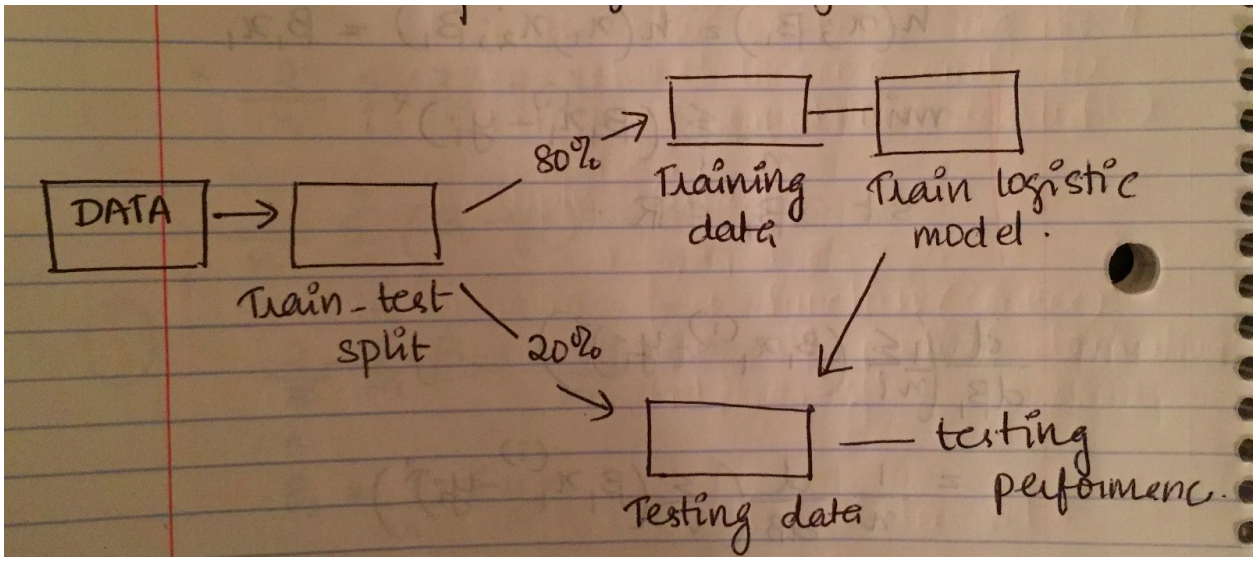


Step 3 : Checked for missing values. It's easy to check for missing values by calling the isnull() method, and the sum() method off of that, to return a tally of all the True values that are returned by the isnull() method.

I checked for test data and train data and observed there are no missing values

Therefore I concluded the data is clean and I preferred to use the raw data as such without any cleaning and normalising.

Inference from Logistic Regression Model



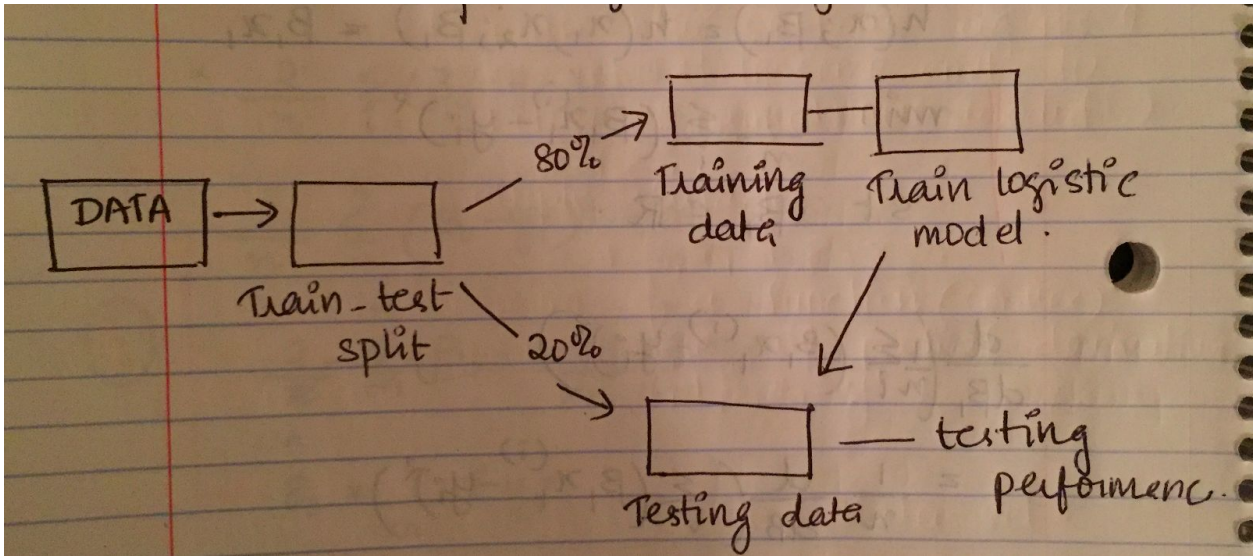
Trained and fitted the train data with logistic regression model and predicted it with the test data.The probability of the result was found using predict_proba function. The result was wrote to a csv file and to validate the analysis I did cross validation and also uploaded the results from csv to Kaggle.

Observed rmse : 0.24055093654697873

Private score : 0.51650

Public score : 0.53466

Inference From Linear Regression Model



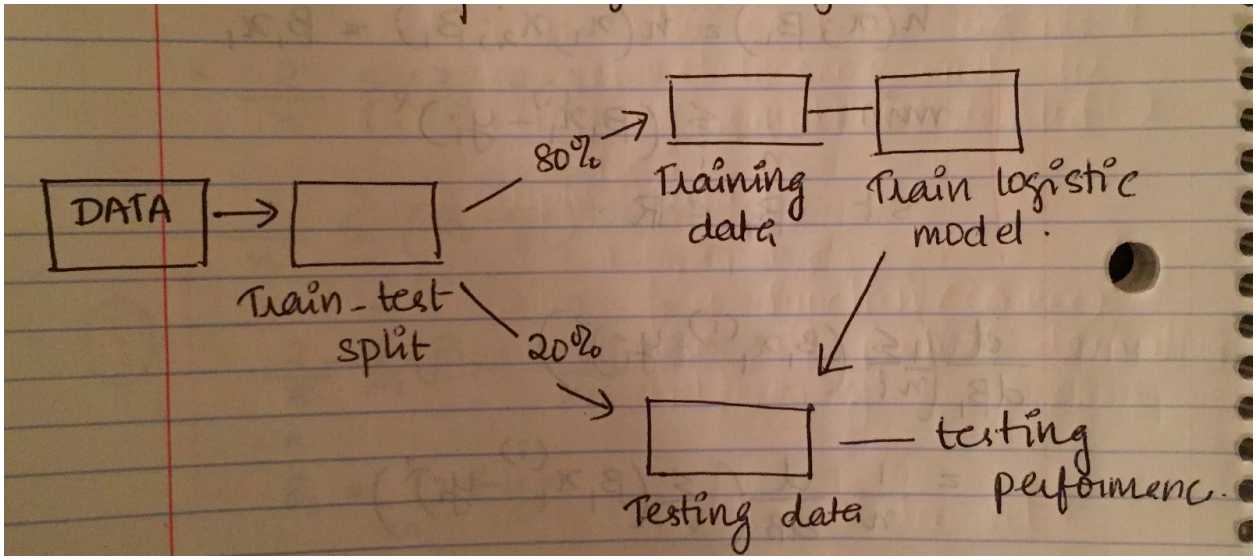
Trained and fitted the train data with linear regression model and predicted it with the test data. The result was wrote to a csv file and to validate the analysis I did cross validation and also uploaded the results from csv to Kaggle.

Observed rmse : 0.2335483191970563 (less compared to logistic)

Private score : 0.52205

Public score : 0.54262

Inference From Ridge Regression Model



Trained and fitted the train data with ridge regression model and predicted it with the test data. The result was wrote to a csv file and to validate the analysis I did cross validation and also uploaded the results from csv to Kaggle.

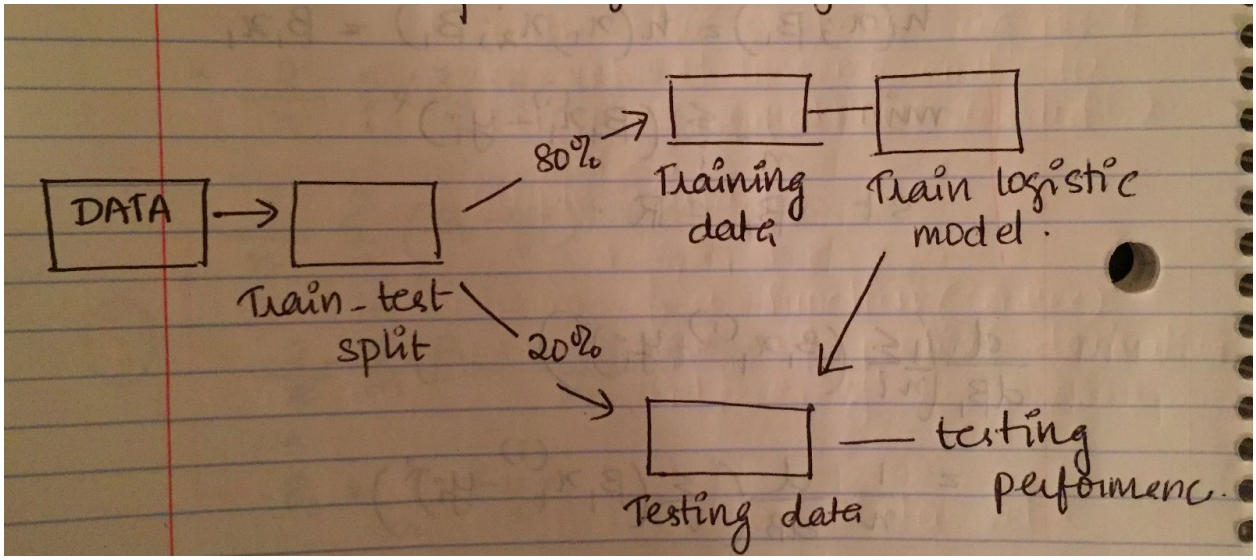
The alpha parameter was given the value 1 .

Observed rmse : 0.23354816726777577 (little bit less compared to linear)

Private score : 0.54627

Public score : 0.57998

Inference From Lasso Regression Model



Trained and fitted the train data with lasso regression model and predicted it with the test data. The result was wrote to a csv file and to validate the analysis I did cross validation and also uploaded the results from csv to Kaggle.

The alpha parameter wag given the value 10e-7 .

Observed rmse : 0.2335480393401645 (little bit less compared to ridge)

Private score : 0.54851

Public score : 0.59131

Inference From XGBRegressor

Trained and fitted the train data with XGBRegressor model and predicted it with the test data. The result was wrote to a csv file and to validate the analysis I did cross validation and also uploaded the results from csv to Kaggle.

Observed rmse : 0.19229360352419217 (less compared to all the models !!!)

Private score : 0.85627

Public score : 0.84942

Inference From XGBClassifier

Trained and fitted the train data with XGBClassifier model and predicted it with the test data. The result was wrote to a csv file and to validate the analysis I did cross validation and also uploaded the results from csv to Kaggle.

I tried tuning the parameters of the XGBClassifier and when max_depth was given a value of 7 I observed the best prediction.

Observed rmse : 0.1997496882143059

Private score : 0.89235

Public score : 0.87768

Inference From RandomForest

Trained and fitted the train data with random forest model and predicted it with the test data. The result was wrote to a csv file and to validate the analysis I did cross validation and also uploaded the results from csv to Kaggle.

Observed rmse : 0.19842226958215162

Private score : 0.79221

Public score : 0.79861