Predicting Purchase Trends using Black Friday Sales Dataset

Vibhor Sharma- 17BCE2152; Yadavendra Sakharkar- 17BCE2170; Rishabh Singh-17BCE0094

Every year, Black Friday happens on the fourth Thursday of November. Most products are marked down with discounts, this event is therefore a huge attraction for people and people will definitely start purchasing more. We are trying to build a model to predict Black Friday Sales through a given public dataset on Kaggle to help the shop owners stop the problem of controlling the crowd with by targeting prospective customers. To predict the sales we are going to use some machine learning algorithms like Random Forest, Decision Trees, Linear Regression, XGBoost and so on, and decide which one is the most successful.

***Keywords* — Black Friday Sales, Prediction model, Regression, Neural network, Machine Learning, Rule Based Learning**

I. INTRODUCTION

For huge franchises and large stores, it becomes impossible for them to know about the preferences of individual customers. Some examples of such franchises are Costco, Walmart, and Wholefoods. These stores without any proper knowledge of their customer base are struggling to satisfy the customer needs. Thus, prediction models are needed to better understand customer preferences. Let us give some more context on Black Friday, it is the largest shopping day of the year in United States of America. Black Friday is the day after Thanksgiving Day which marks the beginning of the shopping season for Christmas. A prediction model developed for Black Friday can only be used during that day because customer spending differs drastically between a normal day and a Black Friday; this is because discounts and price reductions attract more customers.

Finally, better visualization techniques are required to portray the findings and help the store owners understand their customers. Our dataset is the Black Friday Sales Dataset in Kaggle. In this dataset we have the information about the Age, Occupation, City, Duration stayed, Marital status, the quantity of products bought of various types and the total amount spent. We are using these inputs to find the most necessary attributes, potentially excluding some attributes. Finally, we arrive at the conclusion from applying these models to find which model is best suited to predict Purchase trend of customers.

II. RELATED WORK

The most important component of getting knowledge from an existing dataset, is to select the most appropriate model that fits the dataset.

[1] This paper tells us that implementation and high classification efficiency However, this method is too dependent on the distribution of samples in the sample space, and has the potential of instability.To this end, the decision tree strategy is acquainted with manage the issue of intrigue characterization, and the imaginative utilization of Local storage technology in HTML5 to acquire the necessary exploratory information. This paper tells us that Both theoretical analysis and experimental results show that the decision tree is used to deal with the problem of prediction of users' interests has obvious advantages in the efficiency and stability.

[2] This paper tells us that a sales forecast model was designed based on BP (Back Propagation) neural network Experimental results show that the BP algorithm is applied to predict the substance of accidental guidelines; it has higher exactness and preferred prescient capacity over the customary regression analysis approach.

[3] This paper focuses on how raw data is converted into valuable data. For the exploration of large multidimensional data, analysts want to identify problem areas at a glance, with the possibility to

effectively drill into issue territories to get important data. Therefore, we need to present an overview of the data and at the same time show detailed information for each data item. For the exploration of large volumes of maldistributed data, the current charts and tables are not able to show important information such as data distribution of multiple attributes. patterns, correlations, trends, and exceptions, and. detailed information, for example, each sales transaction with price, location, time, and so forth.

[4] This paper tells us about tree boosting which is a highly effective and widely used machine

learning method. we depict an adaptable start to finish tree boosting framework called XGBoost, which is utilized broadly by information researchers to achieve state-of-the-art results on many AI challenges. We propose a novel sparsity-aware algorithm for inadequate information and a weighted quantile sketch for inexact tree learning.

[5] Random forests were proposed for building a predictor ensemble with a set of decision trees that grow in randomly selected subspaces of data. Despite practical use, there has been little exploration of the statistical properties of random forests. The procedure of random forest is consistent and adapts to sparsity, in the sense that

its rate of convergence depends only on the number of strong features and not on how many noise variables are present. Substantial gains in classification and regression accuracy can be achieved by using ensembles of trees, where each tree in the ensemble is grown in accordance with a random parameter. Last expectations are gotten by conglomerating over the gathering.

[6] Black Friday is the largest shopping day of the year in the USA following the Thanksgiving holiday. This paper investigates price differences between Black Friday, Cyber Monday, and December 10. Utilizing information extricated from the sites of five retailers, the examination uncovers that cost contrasts alone don't represent the expanded spending on Black Friday and Cyber Monday. The results of the study confirm that consumers buy discounted products for more than economic reasons, because of experiential and hedonic potential that shopping presents. Shipping charges and availability of products are also significant elements of influencing promotional shopping.

[7] The Web-based Multiple Regression Analysis Data Matrices Package empowers contending member groups in the promoting reenactment to contend to apply their insight into multiple regression analyses in deals determining. They use this package to create nine data matrices (one data matrix for each strategic business unit) consisting of relevant predictor and response variables for each of the prior decision periods. These data matrices are screened for potential multicollinearity among the predictor variables using correlation analysis. This package facilitates the integration of computers, the Internet and the World Wide Web into the marketing curriculum.

[8] This research paper tries to show us the possible approaches which can be taken to build a model based on black Friday sales using different datasets.

[9] In this paper, they have made a survey based on different decision tree algorithm so as to give out clear observation on which one is best in which conditions.

[10] In this paper, they abuse a heuristic bootstrap sampling approach joined with the ensemble learning algorithm for the enormous scope protection business information mining, and they proposed an ensemble random forest algorithm that utilized the equal registering capacity and memory-cache mechanism improved by Spark.

III. OBJECTIVES

1. First of all, developing an accurate prediction model for Purchase trend.
2. Comparing the algorithms of various types and its effectiveness on the dataset, the parameter for comparison is RMSE (Root Mean Square Error).
3. Checking the importance of pre-processing and visualization techniques in increase the accuracy
4. Applying hyperparameter tuning to the models, trying to improve them

IV. METHODOLOGY

First, let us start with the description of the dataset that we are going to utilize in his project.
The columns are as follows:
User_ID - unique
Product_ID – of the form (‘P001345’)
Gender – M/F (Bool)
Age - Range
Occupation – Ranging from 1-20

V. City_Category – A/B/C Stay_In_Current_City_Years – Range: 1-4 Marital_Status – Yes/No Product_Category_1 – Range: 1-17 Product_Category_2 – Range: 1-17 Product_Category_3 – Range: 1-17

VI. The defined target value is **Purchase** which contains numerical values ranging form 7000-15000.

The dataset is already divided into two components.

1. Train
2. Test (No purchase values present): These purchase values need to be predicted using various models

Dataset Used:

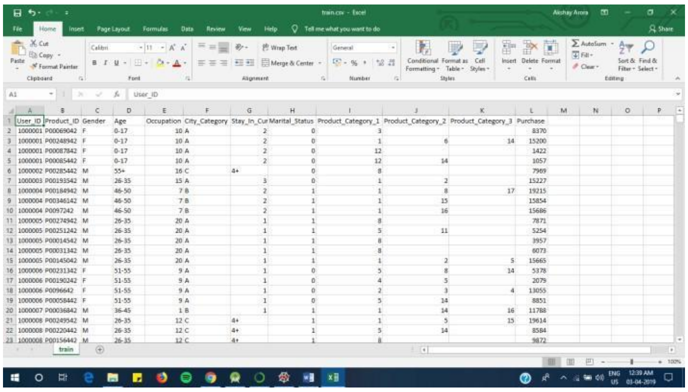


Fig: Train.csv

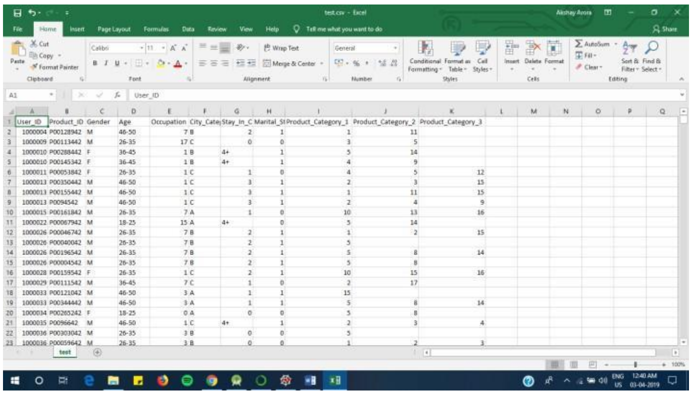


Fig: Test.csv

In the proposed methodology, our system involves the application of machine learning techniques to predict the final target. This is done using the following steps:

1. Data analysis

In this step we just use common descriptive statistics techniques and apply them on our existing data like mean, median, standard deviation, frequency etc. Also, we'll find skew, kurtosis followed by correlation matrix with respect to Purchase values.

2. Data pre-processing

Data pre-processing is an essential step in the process of machine learning. It includes data cleaning and data partitioning

This stage will involve removing all the NA (null) values and replacing them with some integer so that processing can be carried out. Later we can also display the unique value frequencies of all the columns and finally send this data into a modified train and test .csv files.

Because of our dataset being majority numerical in nature, we use the partitioning technique to remove the presence of unique non-numerical values and convert them to numerical.

3. Parameter Selection

Before selecting the models to use for the training process, we need to decide the columns/features that can be used as predictors and drop the others.

This decision needs to be made on the basis on the data analysis done at the first stage of this process.

For eg: when we made the correlation matrix of the dataset, we found that the column "purchase" was highly correlating with the column

I"Occupation". This infers the segment occupation ought to be remembered for the indicators.

4. Application of Machine Learning techniques

The next step, is to use the modified train and test data and apply machine learning algorithms of various types as read about in the survey.

5. Comparison using RMSE

All the various models/algorithms have the same parameter for comparison i.e. the RMSE value (Root mean squared error). It can be defined as the difference between the predicted and the actual values. It is a parameter that when minimum, gives the most accurate model

6. Exploring and trying to find newer approaches for better accuracy.
Now, our aim is to reduce the value of RMSE, to do this we can use various different models, and even explore further to find which model can provide an even better result

VII. IMPLEMENTATION

Requirements:

- Anaconda Navigator
- Jupyter Notebook
- Presence of python machine learning libraries like xgboost and mlxtend

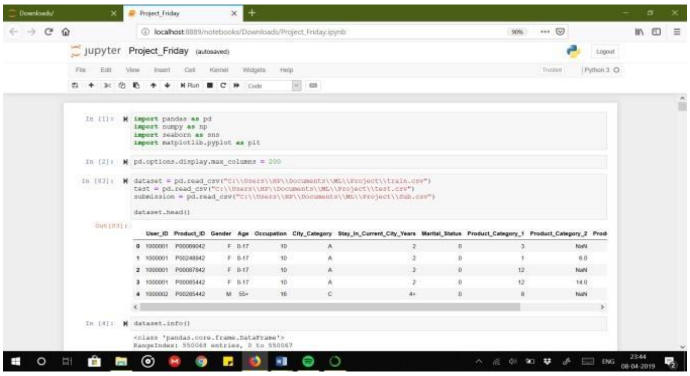
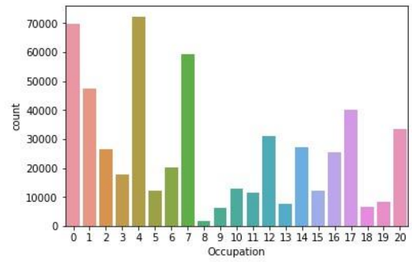


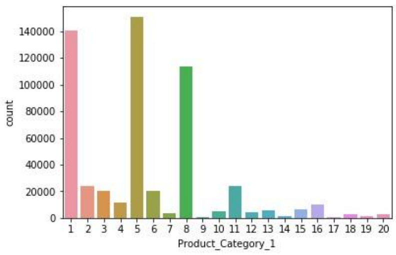
Fig: Interface of Jupyter Notebook

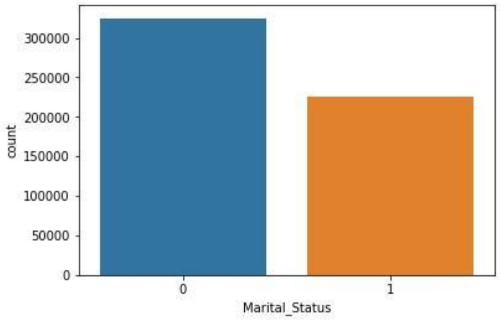
In the above chapter, we defined the steps that we need to follow to get to our final objective. Let us implement the steps one by one.

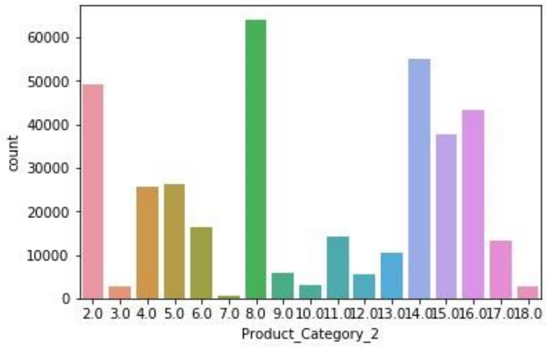
A. 1. Data Analysis

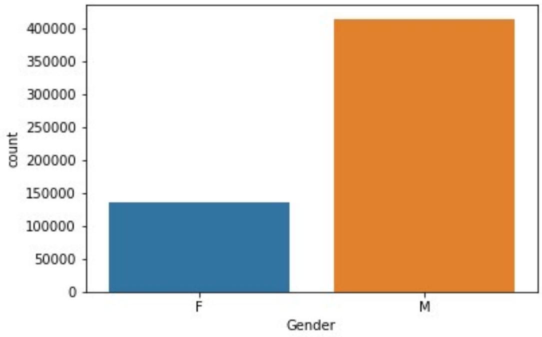
This steps uses basic statistics, building correlation matrices, histograms and finding skewness and kurtosis.











Figures: Bar Graphs of various columns

The correlation matrix is essential for the process of parameter selection.

Code:
corr = numeric_features.corr()

print
(corr['Purchase'].sort_values(ascending=False)[:10],"\ n") f, ax = plt.subplots(figsize=(14, 7))

sns.heatmap(corr, vmax=.8,annot_kws={'size': 14}, annot=True);

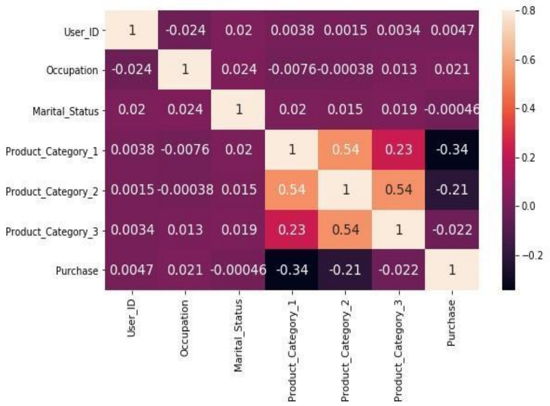


Fig: Correlation Matrix

B. 2. Data Pre-processing

The Data Pre-Processing component in this project consists of the following steps:

- Data Encoding
- Data Scaling
- Removing Outliers
- Removing null values
- Factorizing data
- Removing excess data

C. Data Encoding

```
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder
dataset['User_ID'] = dataset['User_ID'] -
1000000 test['User_ID'] = test['User_ID'] - 1000000 enc = LabelEncoder()

dataset['User_ID'] = enc.fit_transform(dataset['User_ID'])
test['User_ID'] = enc.transform(test['User_ID'])
dataset['Product_ID'] = dataset['Product_ID'].str.replace('P00', '')
test['Product_ID'] = test['Product_ID'].str.replace('P00', '')
```

This step removes the ‘P00’ part from the ProductID so that it can be used in the algorithm fitting process, otherwise a mixture of characters and numbers cannot be utilized. Hence, making it an essential step.

D. Data Scaling

```
scaler = StandardScaler()
dataset['Product_ID'] = scaler.fit_transform(dataset['Product_ID'].va lues.reshape(-1, 1))
test['Product_ID'] = scaler.transform(test['Product_ID'].values.re shape(-1, 1))
```

This process is used to reshape the data, to reduce the time that any ensemble learning algorithm might take.

E. Removing Outliers

```
extra = data.index

[data.Product_Category_1.isin([19,20]))

& (data.source = “dataset”)] data = data.drop(extra)
```

This code, allows us to drop the excess values, that exist in the columns Product_Category_1 and Product_Category_2

F. Removing Null values

```
data.isnull().sum()/data.shape[0]*100 data["Product_Category_2"]=\

data["Product_Category_2"].fillna(1.0).astype("float")
```

data.Product_Category_2.value_counts().sort_i ndex()

```
In [30]:  data.isnull().sum()/data.shape[0]*100

Out[30]:  User_ID                        0.000000
          Product_ID                     0.000000
          Gender                         0.000000
          Age                            0.000000
          Occupation                     0.000000
          City_Category                  0.000000
          Stay_In_Current_City_Years     0.000000
          Marital_Status                 0.000000
          Product_Category_1             0.000000
          Product_Category_2            31.388587
          Product_Category_3            69.648078
          Purchase                      29.808452
          source                         0.000000
          dtype: float64
```

```
In [31]:  data["Product_Category_2"]=\
          data["Product_Category_2"].fillna(-1.0).astype("float")
          data.Product_Category_2.value_counts().sort_index()

Out[31]:  -1.0     245982
           2.0      70498
           3.0       4123
           4.0      36705
           5.0      37165
           6.0      23575
           7.0        854
           8.0      91317
           9.0       8177
           10.0      4420
           11.0     20230
           12.0      7801
           13.0     15054
           14.0     78834
           15.0     54114
           16.0     61687
           17.0     19104
           18.0      4027
          Name: Product_Category_2, dtype: int64
```

G. Factorizing data

data['Gender'],ages = pd.factorize(data['Gender']) print(ages) print(data['Gender'].unique())

data["Gender"].value_counts()

```
In [37]:  data['Gender'],ages = pd.factorize(data['Gender'])
          print(ages)
          print(data['Gender'].unique())
          data["Gender"].value_counts()

          Index(['F', 'M'], dtype='object')
          [0 1]

Out[37]:  1    590031
          0    193636
          Name: Gender, dtype: int64
```

H. Removing Excess data

extra = data.index[(data.Product_Category_1.isin([19, 20])) & (data.source == "dataset")] data = data.drop(extra)

1) 3. Parameter Selection

As stated above this process is done with the help of earlier analysis and statistics. We selected various values from the dataset predictors: All Columns except Purchase target: Purchase

IDcol: User_ID, Product_ID

```
In [58]:  # #Define target and ID columns:
          # target = 'Item_Outlet_Sales'
          # IDcol = ['Item_Identifier','Outlet_Identifier']

          #Define target and ID columns:
          target = 'Purchase'
          IDcol = ['User_ID','Product_ID']

          from sklearn.model_selection import cross_val_score, cross_val_predict
          from sklearn import metrics
```

2) 4. Machine Learning Algorithms
To predict the purchase amount, we implemented various machine learning algorithms and compared them on accuracy and performance metric. Since it is a regression problem, the loss function used is the Root Mean Squared error (RMSE).

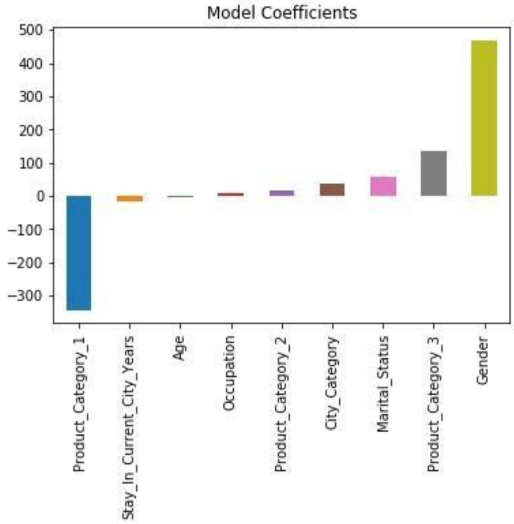
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

3) i. Linear Regression

The linear regression using python's sci-kit library was implemented on the transformed dataset. This was the simplest of the implementations in terms of complexity of the model.
Model Report

RMSE: 4632

CV Score: Mean - 4635 | Std - 35.02 | Min - 4545 |
Max – 4688

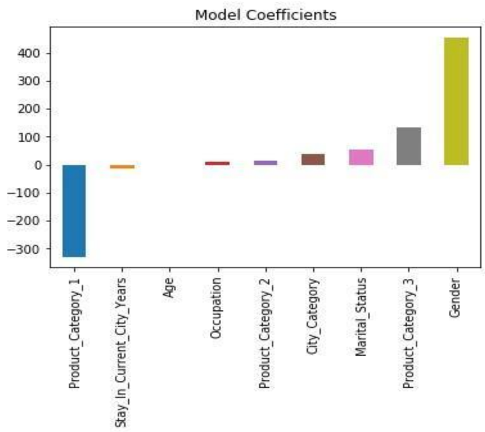


4) ii. Ridge regression

The ridge regression using python's sci-kit library was implemented on the transformed dataset.
Model Report

RMSE: 4817

CV Score: Mean - 4818 | Std - 112.1 | Min - 4741 |
Max – 5293

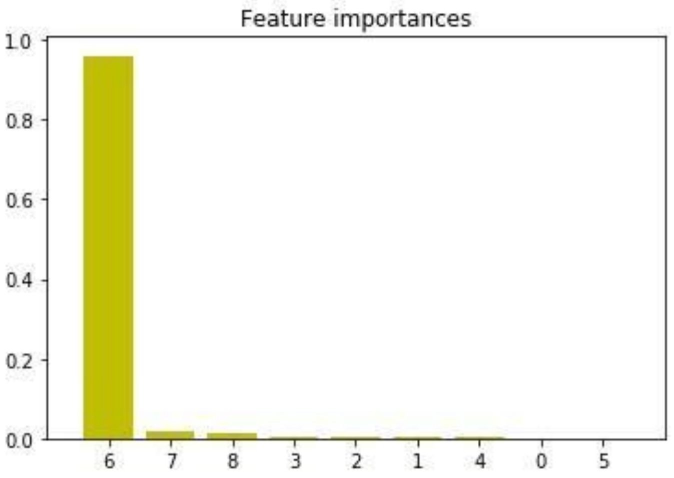


5) iii. Decision Tree Regression

Machine learning algorithms like decision tree and regression are used for developing a simple yet efficient prediction models. We used decision tree on this dataset, and it gave a good RMSE value
Model Report

RMSE: 2996

CV Score: Mean - 3242 | Std - 54.63 | Min - 3031 | Max
– 3289



6) iv. XGBoost Model
The XGBoost is an ensemble learning model, which uses bagging and boosting. This is considered to be a highly robust model.

Model Report
Mean Absolute Error : 240.82192676282142
RMSE : 2926

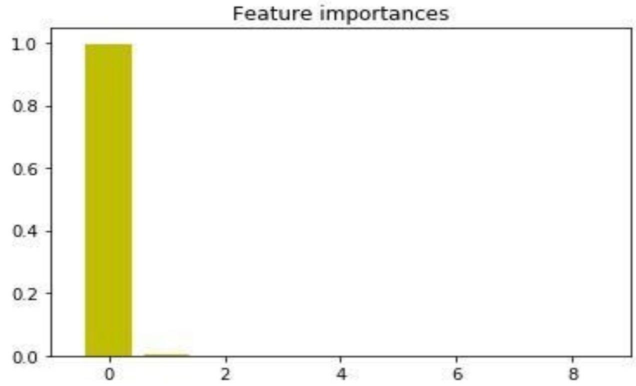
Feature order:
1. feature 6 (0.948004)
2. feature 8 (0.014795)
3. feature 7 (0.011929)
4. feature 3 (0.010370)
5. feature 4 (0.003400)
6. feature 2 (0.003174)
7. feature 1 (0.003119)
8. feature 5 (0.002954)
9. feature 0 (0.002256)

7) v. Random Forest Regression

Random forest Regression uses a multitude of regression Trees, and chooses the result using the best valued tree.

Model Report

RMSE : 2956
CV Score : Mean - 2973 | Std - 20.58 | Min - 2936 | Max - 3008
Mean Absolute Error : 3.7333049827565437



Hence, we can see above we implemented all the algorithms, we read about in the Literature Survey and implemented them, we got the RMSE value for each model and the features that affect it the most.

We used the model to predict the purchase values, and stored the predicted Purchase values in separate .csv /excel files.

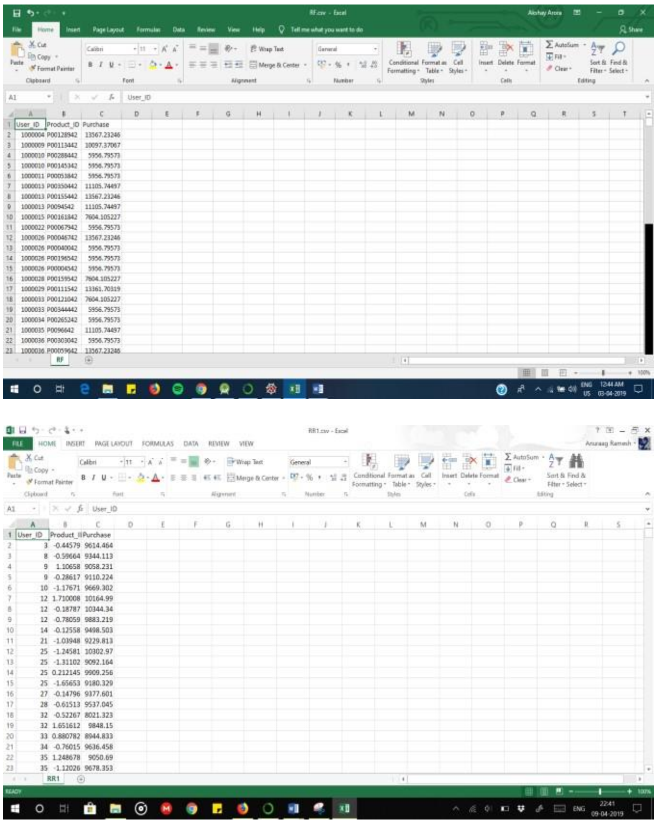


Fig: Excel files that contain the predicted values.

VIII. ALGORITHM

After applying all the basic machine learning algorithms, as stated in the section above. We tried to explore the various parameters of the dataset and landed upon the concept of rule-based learning.
The concept of Rule Based Learning uses the rules that are used to classify a Decision Tree and then utilizes the same, to formulate certain guidelines.
Next step, is to use those guidelines to increase the accuracy of our prediction.

```
In [94]:  def find_node(tree_, current_node, search_node, features):

              child_left = tree_.children_left[current_node]
              child_right = tree_.children_right[current_node]

              split_feature = str(features[tree_.feature[current_node]])
              split_value = str(tree_.threshold[current_node])


              if child_left != -1:
                  if child_left != search_node:
                      left_one = find_node(tree_, child_left, search_node, features)
                  else:
                      return(str(split_feature)+" <= "+str(split_value))
              else:
                  return ""

              if child_right != -1:
                  if child_right != search_node:
                      right_one = find_node(tree_, child_right, search_node, features)
                  else:
                      return(str(split_feature)+" > "+str(split_value))
              else:
                  return ""


              if len(left_one)>0:
                  return(str(split_feature)+" <= "+str(split_value)+", "+left_one)
              elif len(right_one)>0:
                  return(str(split_feature)+" > "+str(split_value)+","+right_one)
              else:
                  return ""
```

The above is the find_node function that is used to extract the rules one by one
The rules extracted using the learning method are:
'Product_Category_1 <= 3.5, Occupation
<= 13.5, Product_Category_1 <= 2.5,
Product_Category_2 > 0.5,Gender <= 0.5,
Product_Category_3 <= 16.5,
Stay_In_Current_City_Years <= 2.5,
Stay_In_Current_City_Years <= 1.5,
Product_Category_3 <= 15.5,
Product_Category_3 <= 6.5'

After extracting the rules, we create a brand new model, by putting the rules found inside a decision tree and then predicting the same.
Model Report
RMSE: 2996
CV Score: Mean - 3242 | Std - 54.63 | Min - 3031 | Max – 3289

Finally, we see that we got the minimum RMSE value, using the XG Boost algorithm we applied. Hence, proving it was a success.

IX. RESULT

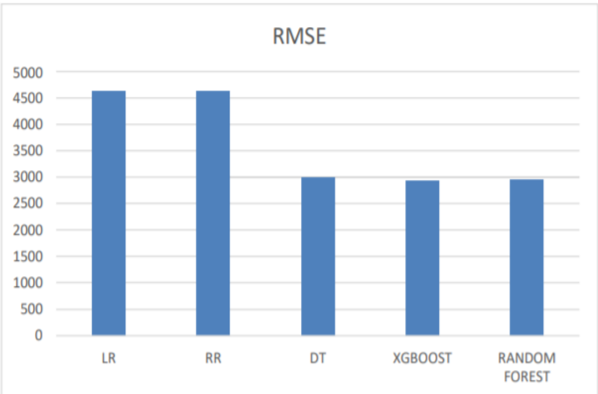


Fig: Tabulating all the RMSE values

Finally, we can compare all the RMSE values that we got form all the models, and we can easily see that XG Boost approach was the best one.
The predicted values form XG Boost closely matched to the already present values. Hence proving our model was a successful one.

X. CONCLUSION

We can conclude by saying that we were able to complete all the objectives listed above successfully.
We applied XG Boost which gave us a minimum RMSE value of 2926.
We were successfully able to predict the Purchase values and trend that was our primary objective, the predicted values are stored in .csv file for later consumption.

XI. REFERENCES

[1] Liu Bing, Shi Yuliang.: Prediction of User's Purchase Intention Based on Machine Learning. International Conference on Soft Computing Machine Intelligence 2(5),(2016)

[2] Yuquan Qin, Haimin Li.: Sales Forecast Based on BP Neural Network. 2(5)

[3] Daniel A. Keim, Ming C. Hao.: Value-Cell Bar Charts for Visualizing Large Transaction Data Sets. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS 2(5), (2007)

[4] Tianqi Chen, Carlos Guestrin.: PXGBoost: A Scalable Tree Boosting System

[5] G´erard Biau.:Analysis of a Random Forests Model. Journal of Machine Learning Research, 2012

[6] Maria Petrescu, Micah Murphy.: black Friday and Cyber Monday: a case study. Int. J marketing and Retailing, (2013)

[7] Aspy Palia.: ONLINE SALES FORECASTING WITH THE MULTIPLE REGRESSION ANALYSIS DATA MATRICES PACKAGE. Developments in Business Simulation and Experiential Learning, (2004)

[8] Esther Swilley, RonaldE.Goldsmith.:Black Friday and Cyber Monday : Understanding consumer intentions on two major shopping days. Journal of Retailing and Consumer Services, (2013)

[9] Himani Sharma, Sunil Kumar.:A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR), (2013)

[10] WEIWEI LIN, ZIMING WU, LONGXIN LIN, ANGZHAN WEN, JIN LI.:An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. SPECIAL SECTION ON RECENT ADVANCES IN COMPUTATIONAL INTELLIGENCE PARADIGMS FOR SECURITY AND PRIVACY FOR FOG AND MOBILE EDGE COMPUTING, (2017)