**R for Data Science**

**Group 1 : Black Friday Sales Prediction**

**Team**

Aruna Maurya (113), Srinivas Machiraju (506), Venu Vardhan (508), Mohit Hooda (241)

**Executive Summary**

Everything in the world today is data-driven, and being able to harness this data to draw meaningful insights to learn, analyze and improve upon business models is what we strive to achieve. Having said that, this report will touch upon the problem statement description, a brief description of the data, its source and its key characteristics followed by the findings from the application of different data mining models which include key approaches and comparative results of various techniques applied. We have also tried different types of models to compare and contrast and see the difference in results as an accuracy measure. The model that performed the best, however, was a Decision Tree. We have also to explore the different reasons as to how and why these models have differences in their accuracies.

**Problem description**

A retail company “ABC Private Limited” wants to understand the customer purchase behavior (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high volume products from last month. The data set also contains customer demographics (age, gender, marital status, city\_type, stay\_in\_current\_city), product details (product\_id and product category) and Total purchase\_amount from last month. Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create a personalized offer for customers against different products.

When solving a problem, particularly in the field of Machine Learning, we start from the basic models which bring the usage the Linear Regression for our problem. To improve accuracy or reduce the error of the model, we use several data preprocessing techniques like normalization, handling categorical values, etc. Following this step, we try to use some complex models like Decision Trees, Support Vector Machines, XGBoost algorithm, etc.

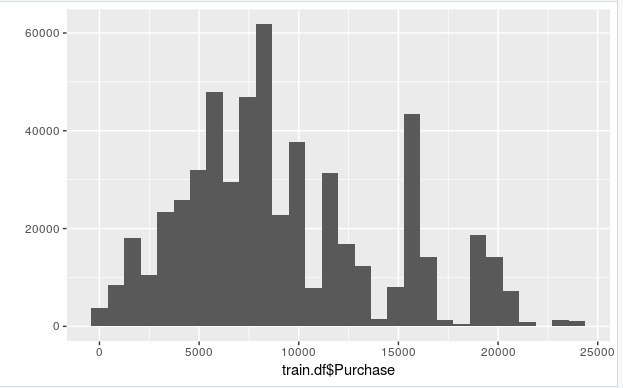
**Dataset Description**

Source: <https://datahack.analyticsvidhya.com/contest/black-friday/>

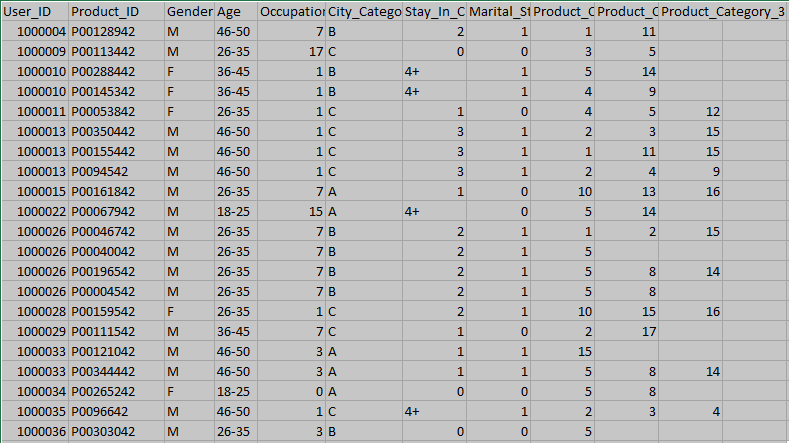
|  |  |
| --- | --- |
| *Variable* | *Definition* |
| User\_ID | User ID |
| Product\_ID | Product ID |
| Gender | Sex of User |
| Age | Age in bins |
| Occupation | Occupation |
| City\_Category | Category of the city(A, B, C) |
| Stay\_In\_Current\_City\_Yearscity | Number of years stay in current |
| Marital\_Status | Marital Status |
| Product\_Category\_1 | Product Category (Masked) |
| Product\_Category\_2 | Product may belong to another category also (Masked) |
| Product\_Category\_3 | Product may belong to another category also (Masked) |
| Purchase | Purchase Amount (Target Variable) |

***Note*:** Model performance will be evaluated on the basis of prediction of the purchase amount for the test data.

Histogram plotting Purchase



**Data Snippet**

****

**Data Preprocessing**

* Data cleansing was achieved after obtaining the summary of the dataset followed by a thorough check of any missing values. Missing values were later replaced by integer 20 as missing values can cause problems in the training phase if not taken care of before.



* Data were partitioned and all the string data types were changed to integer. One hot encoding features of the training and testing dataset as some columns are of string type.

**Results Obtained from Various Models**

1. Linear Regression Model
2. Decision Tree Modelling
3. SVM
4. XGBoost

Out of which, Decision Tree and XGBoost yielded good results. So, we took an weighted average of both the predictions and generated our result.

submit\_final$Purchase <- (submit\_dc$Purchase + 2\*submit\_xg$Purchase)/3

|  |  |  |
| --- | --- | --- |
| *Name of the classifier* | *Initial Accuracy* | *Accuracy after Data Preprocessing* |
| Linear Regression | 6168.147 | 5599.195 |
| Decision Tree | 4653.5352 | 3404.233 |
| SVM | 6431.296 | --- |
| XGBoost | 2786.564 | 2654.343 |
| WA(XGBoost, DT) | --- | 2620.962 |

Source Code: <https://github.com/vchrombie/black-friday-sales>

R Notebook link: <https://vchrombie.github.io/black-friday-sales/black-friday-sales.nb.html>

**Conclusion**

AV LB Username: *srinivasmachiraju*

AV LB Rank: *520*

AV LB Score: *2620.962*

