IDS 575 Project

Black Friday Sales - A study of sales through consumer behaviors

Ashwin Narayanan – anaray22 - 654325491

Vigneshwaran Giri Velumani – vgiriv2 - 670664553

Contents

Abstract

Analyzed Black Friday Dataset and solved the regression problem by predicting the different factors that most influence the purchase amount. Implemented four different models on top of the baseline model and compared the effectiveness of each model based on two parameters - Root Mean Squared Error (RMSE) and MAE (Mean Absolute Error) and found that the Stepwise Regression Model performs the best among the four models used. Our findings reveal that Men who are aged between 40-50, and occupation type 2, 12 and 17 and city category C predominantly influence the Purchase Amount which is the target variable we are interested in.

Introduction

Problem Statement: 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 behavior of customer against various products which will help them to create personalized offer for customers against different products.

- Customers are purchasing multiple product categories in different combinations like city, marital status, year & quantity.
- As the target variable – Purchase Amount is numeric, we are trying to solve a Regression Problem.
- The data can provide us interesting demographic patterns influencing purchase behavior of customers

Related work

The dataset was taken from Kaggle and analytics Vidhya. So, the works was inspired from many blogs and articles. Among several articles and blogs, we referred, most of them removed or left the product category Features unused since it does not play any vital role in prediction. Some articles imputed the missing values, some just filled with mean values. We have just dropped those features for our analysis. Some of them are listed below:

https://rpubs.com/mohitrajput901/262248

https://rpubs.com/mohitrajput901/262248

https://github.com/wanshun123/black-friday/blob/master/blackfriday.r

The following resource from the Customer Analytics Chapter has been used for Exploratory Data Analysis in Python

https://github.com/jalajthanaki/Customer_segmentation/blob/master/Cust_segmentation_online_retail.ipynb

Models and methods:

Problem Settings: The dataset has 12 Features and 537557 Observations.

There are few features which can be better represented with other data types. Converting characters into factor levels will give more clarity and easier for analysis.

User ID- We could see same user ID's recurring multiple times	**Product ID** – they are unique and bought by multiple customer
Age – They were in 7 categories	**Occupation** – The occupation was categorized into 20 categorie
City categories – They are classified into 3 categories	**Stay in current city** – They are in number of years

Marital status – Two categories Male and Female	**Categories** – There were few null values in this category. Sa could be in multiple categories.
Purchase – It is in numbers and it is the target variables	

Data Cleaning Explanation:

- Removing plus symbol
- Removing missing values and adding zeros
- Change the format of the entries

Now we have the dataset cleaned. The dataset is with following changes ready for analysis.

- No Null values
- With proper data types
- All symbols removed
- All the categorical feature properly grouped into a bucket
- Aggregate the purchases by User ID and take purchases per use and put it in a separate data frame and use it for analysis

Splitting total purchases data frame into train and Test (80/20)

Models:

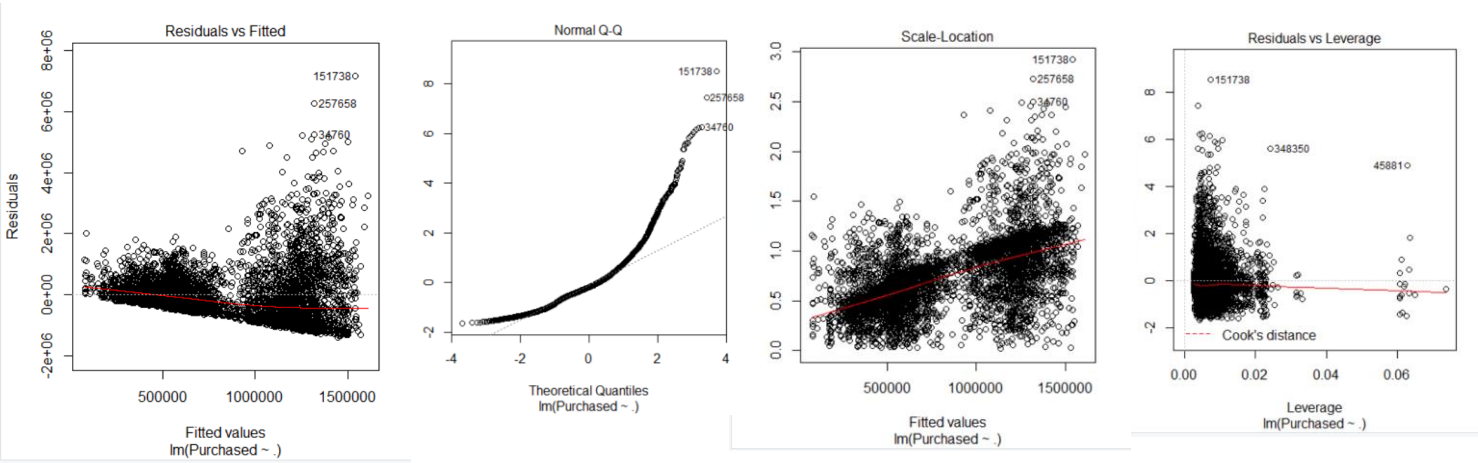
RMSE and MAE are our Model Parameters upon which we have compared and evaluated the performance of our Models.

1. Baseline Model:

```
> # Evaluate RMSE and MAE on the testing data
> RMSE_base <- sqrt(mean((base_mean-test$Purchased)^2))
> RMSE_base
[1] 977077.9
> MAE_base <- mean(abs(base_mean-test$Purchased))
> MAE_base
[1] 664973.2
```

2. Linear Model:

Plots:



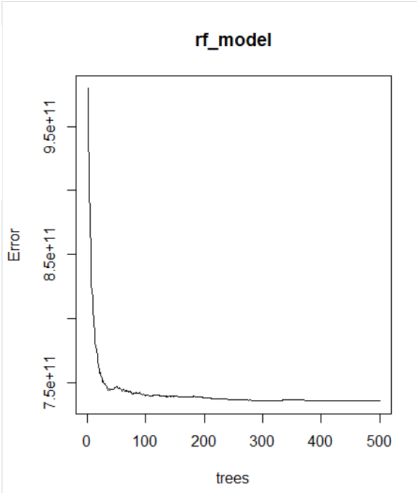
Root Mean Squared Error and Mean Absolute Error

```
> lm_model_rmse
[1] 896198.9
> lm_model_mae
[1] 614793.6
```

Random Forest:

Root Mean Squared Error and Mean Absolute Error

```
> rf_model_rmse
[1] 907894.1
> rf_model_mae
[1] 615031.6
```



Step-wise Regression: Root Mean Squared Error and Mean Absolute Error

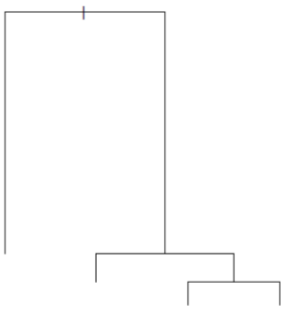
```
> swr_rmse
[1] 840209.6
> swr_mae
[1] 588200
```

Rpart Model:

Root Mean Squared Error and Mean Absolute Error

```
> rpart_rmse
[1] 896305.6
> rpart_mae
[1] 615146.7
```

Plot:



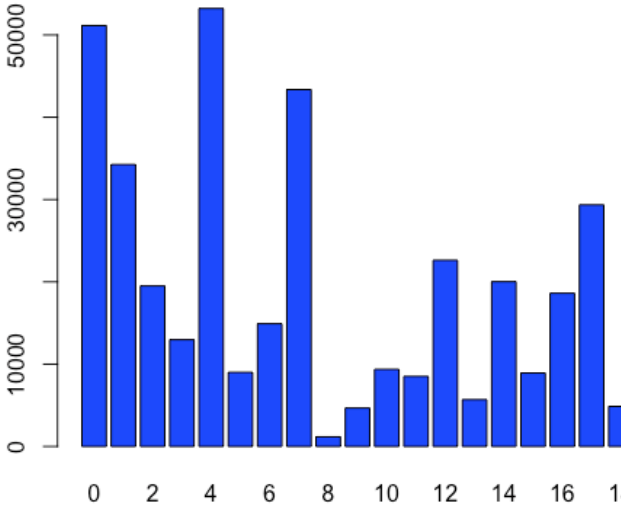
Experimental results

UNIVARIATE ANALYSIS

Various visualizations and plots are used to study the distribution of features in dataset.

OCCUPATION.

- There are 20 categories in the occupation feature.

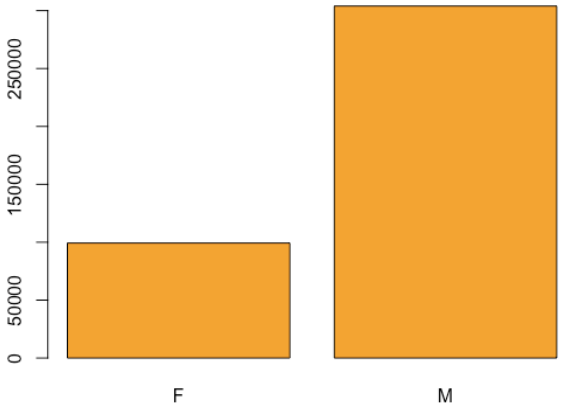


Inference.

Category 4 has the highest number and the other features the distribution is varied. So, there is no interesting pattern.

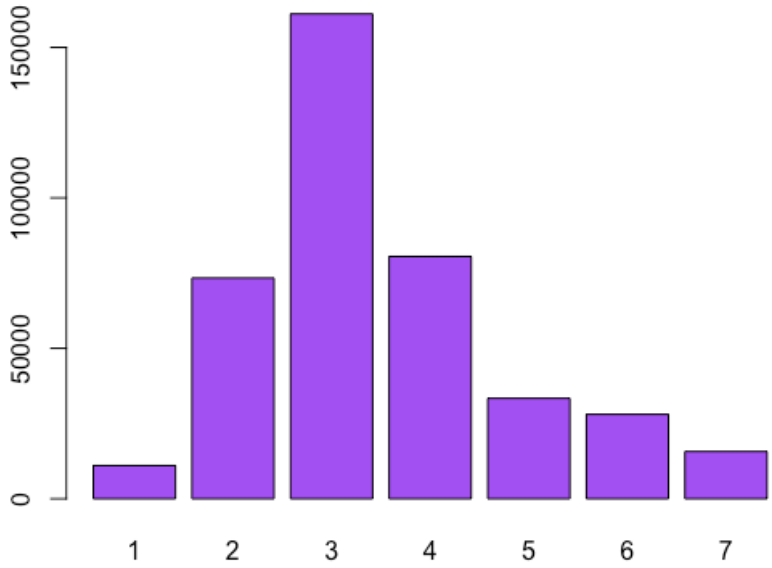
Gender

Clearly more males have bought items during black Friday.



Age

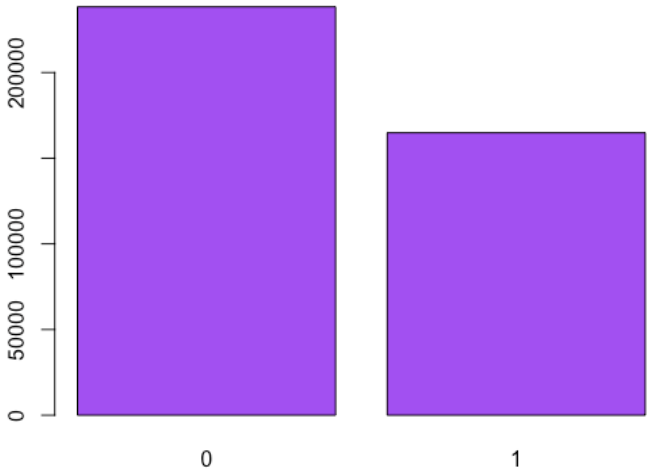
Age Group 3 has the highest purchase.



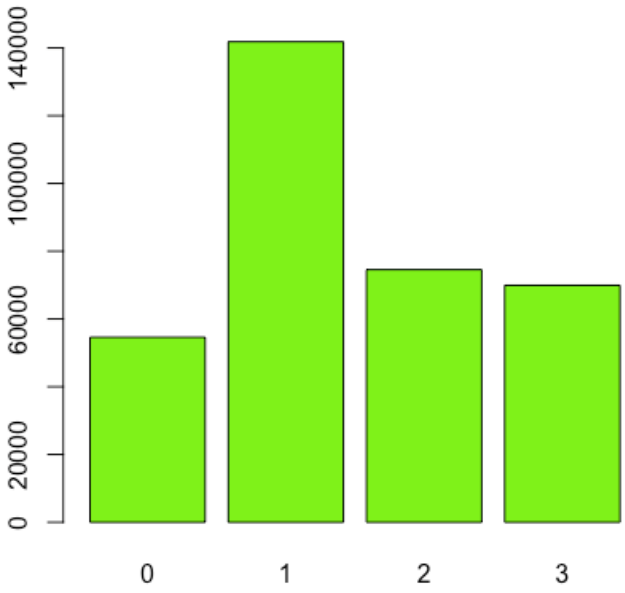
0-17 – Age group 1
18-25 - Age group 2
26-35 – Age Group 2
36-45 – Age Group 3
46-50 - Age Group 4
51-55- Age Group 6
55+ - Age Group 7

Marital Status

0 – Unmarried, 1- Married. The unmarried did more purchases.



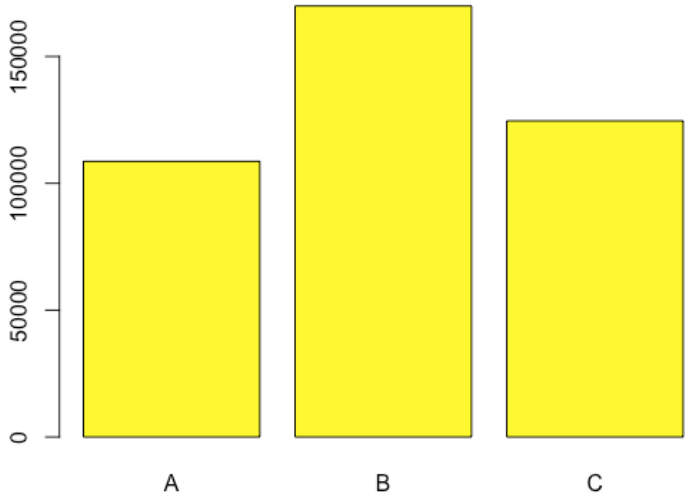
Current Stay in City



These bars represent the number of years a person stayed in the city. People who stayed for a year purchased more.

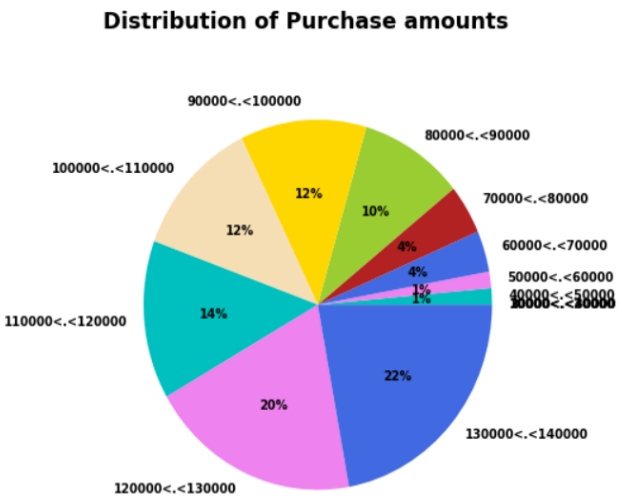
City:

There are three categories of cities A, B, C. No description of these categories.



People in category B has made highest number of purchases.

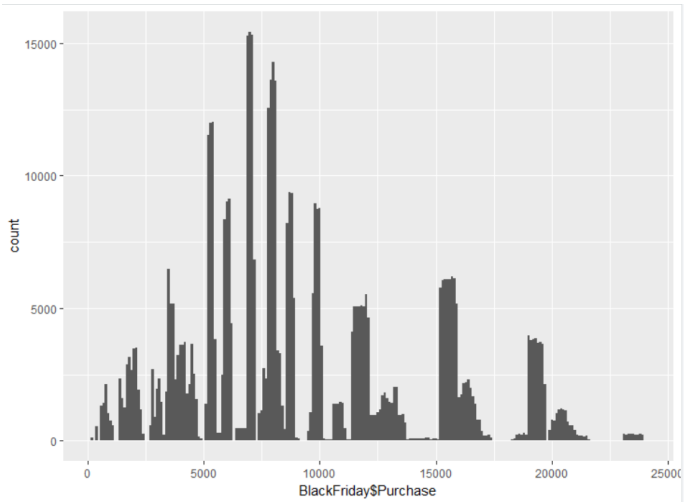
Exploratory Data Analysis:

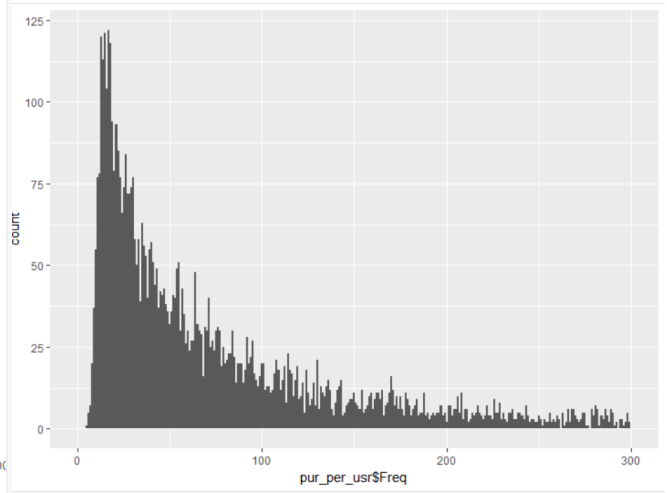


Bivariate Analysis

- Demographic information like city, age, gender can be visualized together. By relating them together we could derive a lot of insights.
- This provides some useful insight, it gives the number of purchases made by people from different demographic information.

ggplot and Purchase per user:





Sample table after aggregating total purchases:

Purchased	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status
333481	F	0-17	10	A	2	0
810353	M	55+	16	C	4+	0
341635	M	26-35	15	A	3	0
205987	M	46-50	7	B	2	1
821001	M	26-35	20	A	1	1
379450	F	51-55	9	A	1	0
234427	M	36-45	1	B	1	1

Results: Comparison of Performance based on Model Parameters:

Model	RMSE	MAE
Baseline Model	977077.9	664973.2
Linear Regression	896198.9	614793.6
RF Model	907894.1	615031.6
Stepwise Regression	840209.6	588200
rpart	896305.6	615146.7

From the Model parameters RMSE and MAE we can observe that Stepwise Model performs better since it has least RMSE and MAE.

Discussion

After running series models, we discovered stepwise regression is the best model as it gave a better r-square value. Running linear regression for entire dataset was time consuming, it took a lot of memory space. So finding aggregated purchase amount (basket amount) and using that value in the model was the best way to analyze and get the models running.

Conclusion

We used various cleaning techniques and models to understand the dataset. The key takeaway is that the male unmarried who are in the age group of 40 - 50 tend to purchase more. So unmarried Gen-X male tend to buy purchase more.