UNIVERSITY OF TEXAS AT DALLAS

BUAN 6337.002

ANALYSIS OF BLACK FRIDAY SALES

Group 4

Vishakha Nangia
Mudit Agrawal
Sneha Murugan
Sneha Liza George
Namrata Kamthe
Abhishek Dhingra

Table of Contents

1. Overview

Black Friday is observed on the Friday following Thanksgiving, where people go to stores to shop in bulk. This day is one of the busiest shopping days as 115 Million people rush the stores for shopping and buy commodities worth $655.8 billion[1]. Both customers and retailers alike are highly pressurized to make the right decisions. While customers, based on their needs and requirements, create shopping lists, the retailers are required to understand the needs of the customers and perceive the value proposition of the customers.

This is important in today's context as the market has dynamically evolved from being product centric to become customer-centric and it has become imperative to address the needs of the customers as they dictate the growth of companies based on their evaluation of products and services.

We as analysts want to understand customer's behavior and analyze the factors that can increase the revenue and profit for the retailers. To address this situation, the intuitive question we raised is, "**How does customer demographics influence Black Friday Sales?**"

The data set in our specific study consisted of gender, occupation, city the customer stays in, duration of stay in the current city, marital status, occupation of the customer etc., along with purchase amount. By understanding how demographics of customers affect purchase, a prediction model can be formulated to understand the factors that drive Black Friday sales, specifically revenue and bill amount for each of these purchases. Retailers can concoct the action plan to incorporate the product types that will enforce higher sales, based on these predictions.

2. Black Friday data set

To address the research question, Kaggle's Black Friday data set was used. Before building models for prediction, it is important to have a good comprehension of the variables. Data set contained 12 variables of which some were numerical, and the rest were categorical. The two unique identifiers in the data set are – User_ID and Product_ID: User_ID assigns each customer with an ID while the Product_ID allocates a number to each product.

Ordered categorical variables in Black Friday data set are Age and Stay_In_Current_City_Years. These categorizes customers based on their age and number of years that they are inhabiting in current city of residence, respectively. Nominal data variables are Gender (records customers' gender), Occupation (type of occupation of a customer), City_Category (City in which the customer is currently residing), Marital_Status (married = 1, unmarried = 0). Other nominal variables, Product_Category_1, Product_Category_2 and Product_Category_3 consists of different product types within the 3 mentioned categories. Finally, the target variable – Purchase indicates the total bill amount per transaction.

3. Data Cleaning:

There were 538,000 observations. However, for the variables – Product_Category_2 and Product_Category_3, 70% of observations were missing, therefore smart Imputation was not the right choice to handle the missing data. Hence, we went ahead and removed the observations from the data and used the 30% of data, i.e., 164,278 observations to analyze and build models.

4. Exploratory Data Analysis:

We used ogive plot to see the cumulative distribution of total purchase for different products.

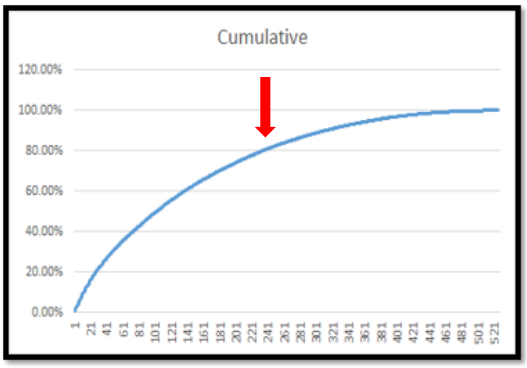


Fig. 4.1 – Ogive Plot

The Fig. 4.1 shows that 50 % of the product types drives 80% of the revenue.

The significance of the above plot can be understood from a recent example of a famous women's clothing company "Charlotte Russe" which filed for bankruptcy in 2019.

The main reason was mismanagement of inventory, as it tried to provide all the types of products in all its stores without taking customers need into consideration. This caused unnecessary higher operating costs

The above graph highlights that the retailer can optimize their operating cost by focusing on limited number of products which drive most of the sales.

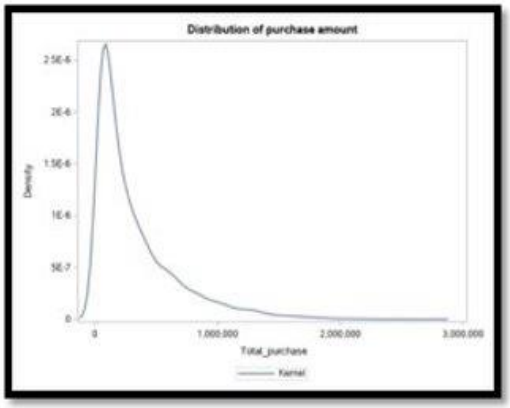


Fig. 4.2 – Distribution of Purchase amount

On performing statistical analysis, we observed the following:

- Purchase ranges from $185 to $23959
- Average purchase - $11661
- Purchase Amount is positively skewed

4.1 Insights on demographics

In the following graphs the one in solids represent trend for all of the product types whereas the shaded ones represent trend for 50% of the product types which drive 80% of the sales.

- Purchase trend as per Gender - Fig.4.1.1 shows that males buy, on an average, greater number of products as compared to females

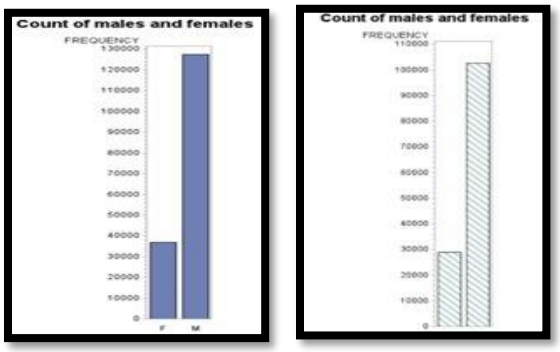


Fig. 4.1.1

- Purchase trend as per Age Groups – Fig.4.1.2 shows that the customers in age group 26-35 buy greater number of products as compared to the customers in other age groups.

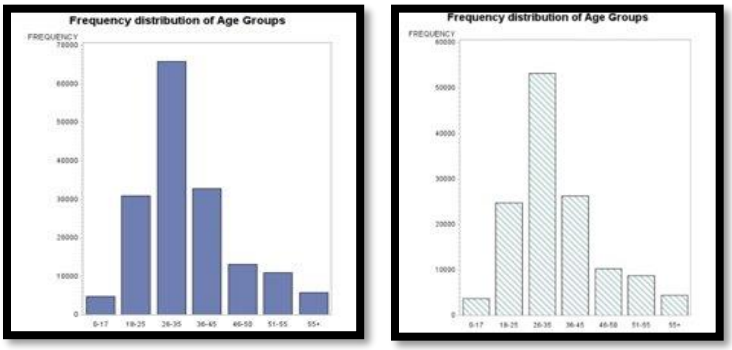


Fig. 4.1.2

- Purchase trend as per Marital Status- Fig.4.1.3 shows that unmarried people tend to buy a greater number of products as compared to the married people.

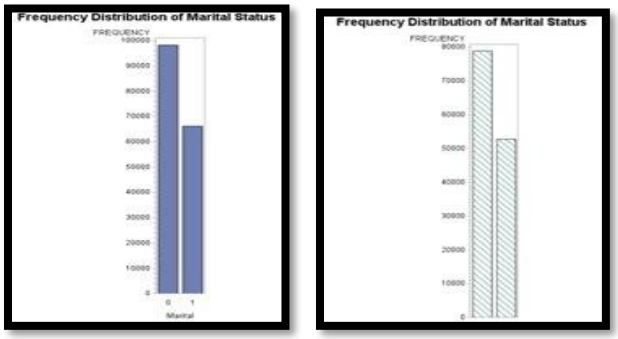


Fig. 4.1.3

- Purchase trend as per years of stay in current city – Fig.4.1.4 shows that the people who live in a current city for 1 year tend to buy more products as compared to the others.

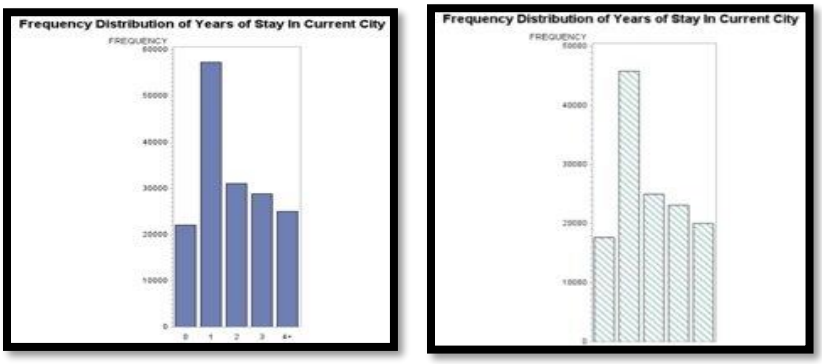


Fig. 4.1.4

From the above analysis we found that the general trend for all the product types and the trend for 50% of the product types is similar, so we will consider all the product types for all the further analysis.

4.2 Purchase behavior for different city categories

To see the product preference for the people living in different cities we used the total and the average purchase amount. We observed the following results –

- City B purchases cheaper but a greater number of products
- City A and C purchase expensive products

The above results were reached on the basis of the sum purchase and mean purchase distributions for different city categories.

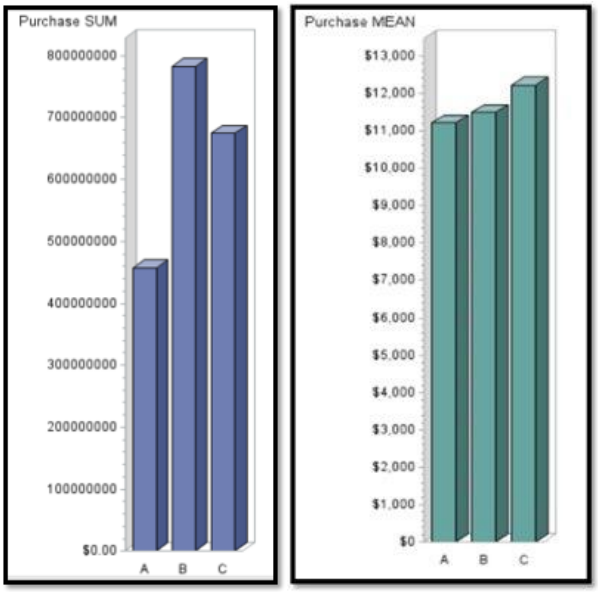
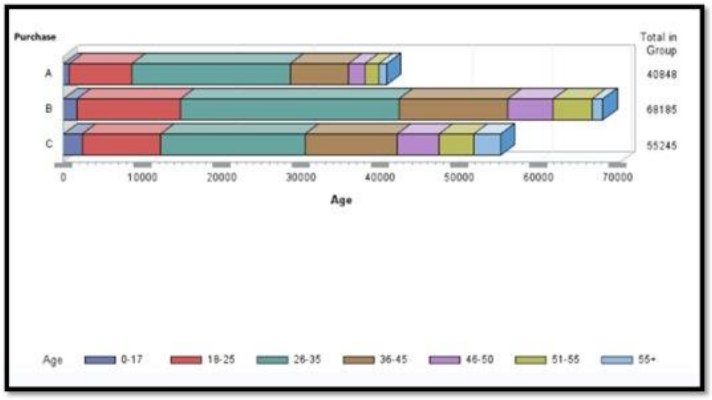


Fig. 4.2.1

Recommendation: For city A and C make expensive products more accessible, while for city B make cheaper products more accessible

Purchase behavior based on age group across different cities – The fig. 4.2.2 shows the Purchase distribution for city A, B, C based on age groups.

- On average Millennials (18-25 & 26-35) buy greater number of products

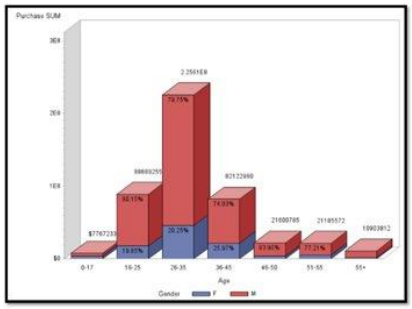


Recommendation: To have higher total number of purchases (or higher total revenue), cater to the needs of millennials

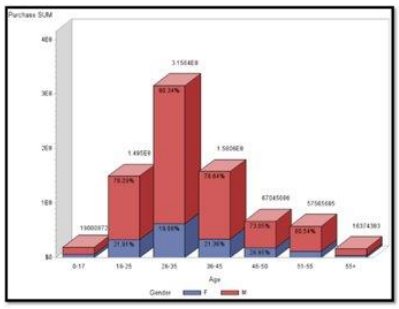
Fig.4.2.2

Purchase behavior based on gender across different cities –

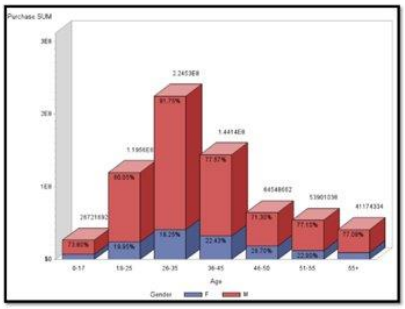
- In general males purchase more than females across all cities and for all age groups



Purchase distribution for City A



Purchase distribution for City B



Purchase distribution for City C

Fig. 4.2.3

Purchase behavior based on marital status across different cities-

- City A has higher proportion of unmarried shoppers (63.25%)
- City B and C have similar proportion of purchases by married and unmarried customers
- City B and C has higher proportion of purchases by married people than City A

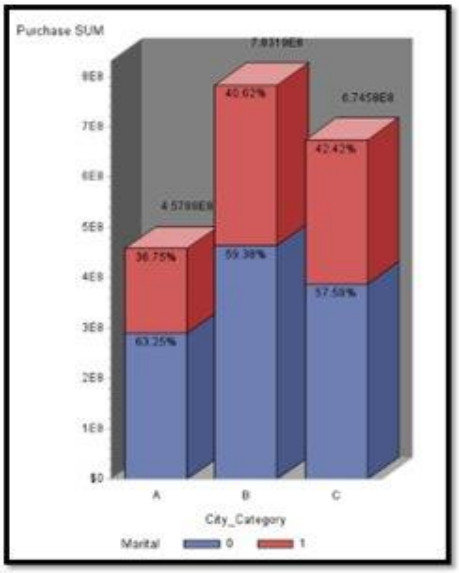
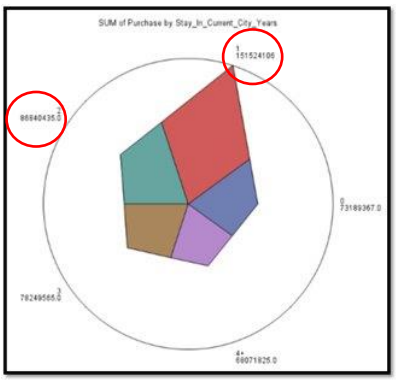


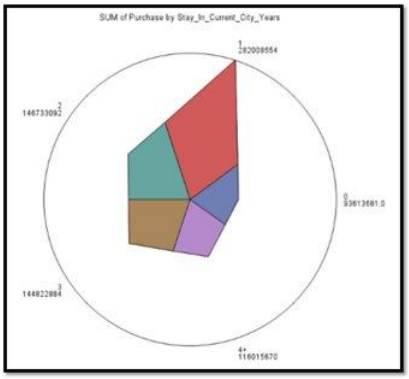
Fig.4.2.4

Purchase behavior based on years of stay in a current city across different cities-

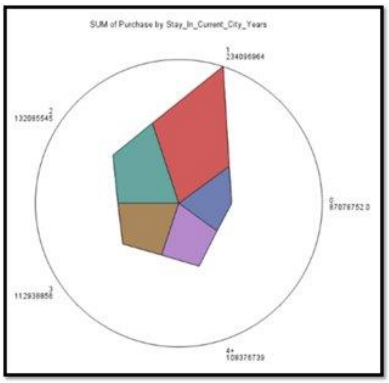
- For city A, B and C the maximum number of products are purchased by the people who have lived in the current city for around 1or 2 years.



Purchase distribution for City A



Purchase distribution for City B

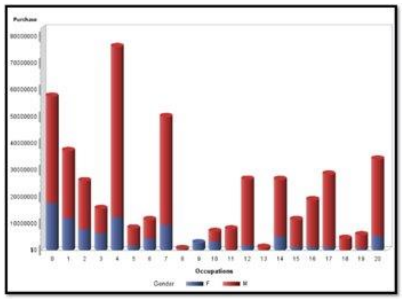


Purchase distribution for City C

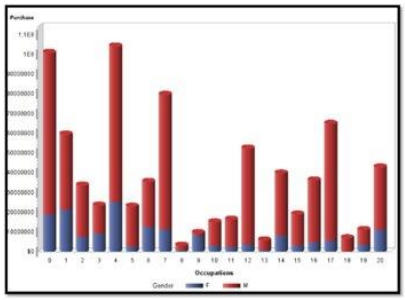
Fig. 4.2.5

Purchase behavior based on occupations across different cities-

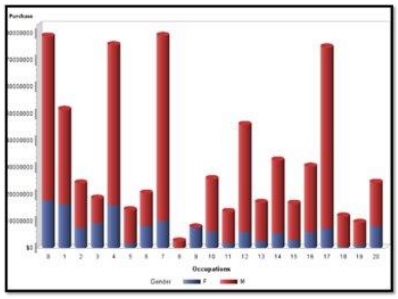
- Occupations have almost similar purchase trends across different city categories
- People from occupation 0,4,7 and 17 tend to buy greater number of products as compared to people from other occupations. May be these occupations have higher pay as compared to other occupations



Purchase distribution for City A



Purchase distribution for City B



Purchase distribution for City C

Fig. 4.2.6

4.3 Market Basket Analysis

We tried to find any relationship among the products that the customers may buy together more frequently, so we performed association data mining for our analysis.

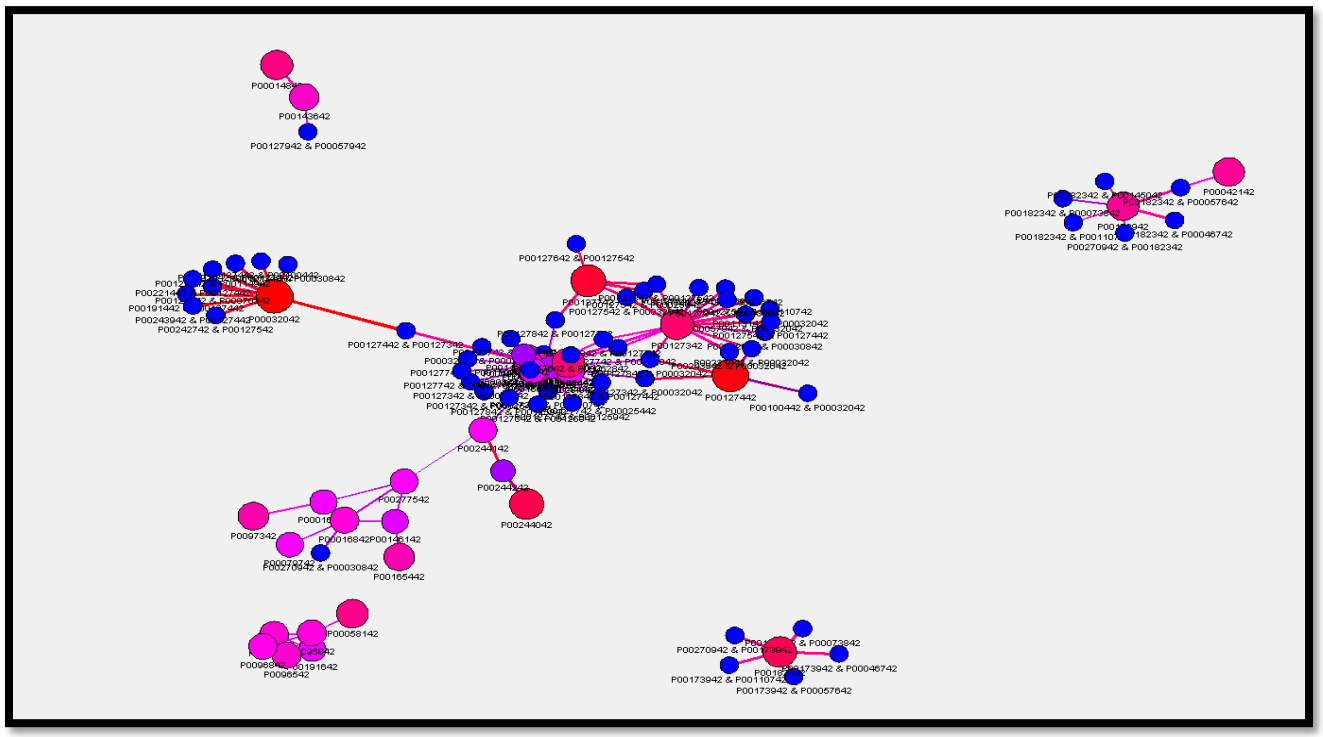


Fig. 4.3.1 Link graph for the various product ids

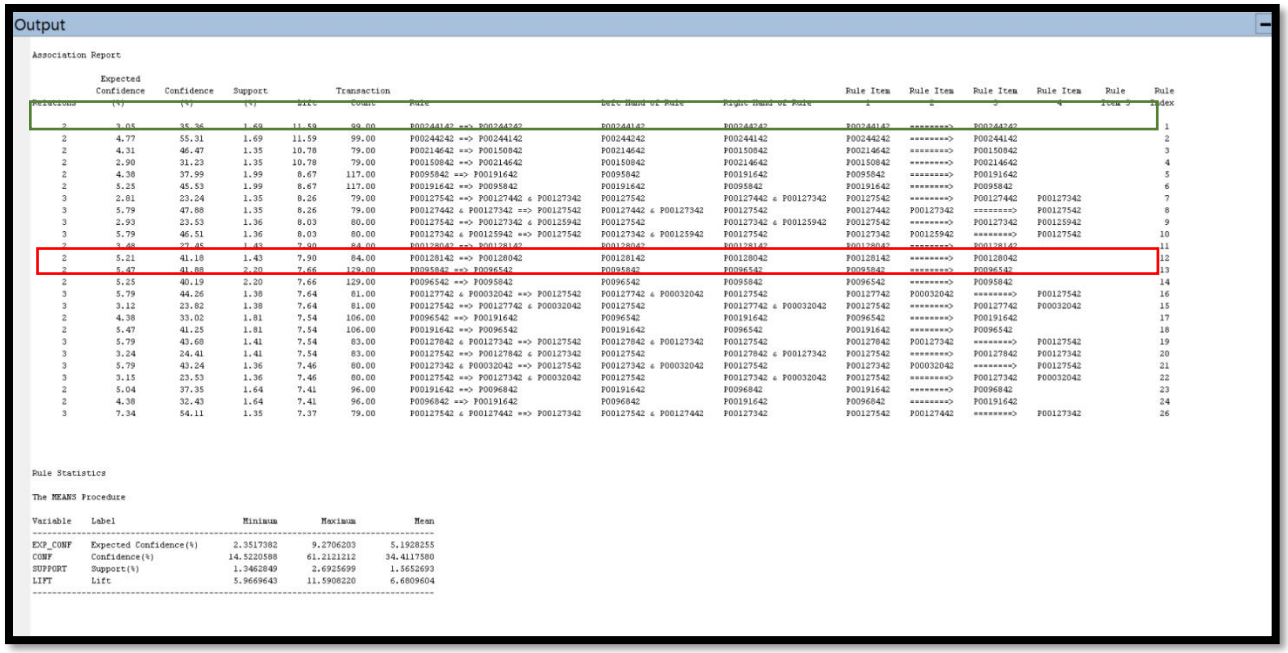


Fig. 4.3.2 Rules with the confidence% and Support%.

Insights

- For the rules highlighted in red, 2.20% of all the customers purchased P0096542 and P0095842 together
- If the customer purchased P0096542 then he/she will also P0095842 with 41.88%
- For rules highlighted in green customers who buy P00244142 are 11.59 times more likely to buy P00244242 than a customer chosen at random.

Market Basket analysis can help the stores in the placement of the products to increase their sales. The products that are bought together frequently should be kept nearby to increase the sales.

5. Predictive Modeling

After having gained some initial insights based on exploratory data analysis, we tried to find statistical support through predictive modeling. Various techniques such as Forward Selection, Backward Selection, Stepwise, Bagging and Lasso were used to find the best model for our analysis and their performance was observed for the test data. Although the performance of the models returned by each method was similar, the best model was provided by Forward Selection Technique with Root MSE value of 0.14657 and Adj. R-sq of 0.68. These results were obtained for logarithmic value of the target variable (to counter skewness).

Root MSE	0.14678
Dependent Mean	4.00732
R-Square	0.6820
Adj R-Sq	0.6807
AIC	-372385
AICC	-372380
SBC	-498535
ASE (Train)	0.02146
ASE (Test)	0.02167
CV PRESS	0

Fig 5.1 (Backward Selection)

Root MSE	0.14657
Dependent Mean	4.00732
R-Square	0.6830
Adj R-Sq	0.6816
AIC	-372761
AICC	-372756
SBC	-498765
ASE (Train)	0.02139
ASE (Test)	0.02163
CV PRESS	2843.54650

Fig 5.2 (Forward Selection)

Root MSE	0.14678
Dependent Mean	4.00732
R-Square	0.6820
Adj R-Sq	0.6807
AIC	-372385
AICC	-372380
SBC	-498535
ASE (Train)	0.02146
ASE (Test)	0.02167
CV PRESS	0

Fig 5.3 (Stepwise Selection)

Root MSE	0.14735
Dependent Mean	4.00732
R-Square	0.6793
Adj R-Sq	0.6782
AIC	-371473
AICC	-371470
SBC	-498680
ASE (Train)	0.02164
ASE (Test)	0.02183
CV PRESS	2839.19636

Fig 5.4 (Lasso)

Score Information	
Input Data Set	WORK.BF_TEST
Output Data Set	WORK.TEST_PERFORMANCE
Number of Observations Read	32855
Number of Observations Scored	32854
Number of Residuals Computed	32854
Residual Sum of Squares	710.73911

Fig 5.5 (Bagging) (Root MSE = 0.1471)

Once the best model for our analysis was obtained, we derived coefficients for each parameter to better understand the correlation among the dependent and independent variables. Below graphs shows the result of the regression.

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.075368	0.015653	260.36	<.0001
Gender_F	1	0.015182	0.002185	6.95	<.0001
City_Category_B	1	0.013753	0.001843	7.46	<.0001
City_Category_C	1	0.032125	0.001970	16.30	<.0001
Stay_In_Current_City_Years_0	1	0.009917	0.002585	3.84	0.0001
Stay_In_Current_City_Years_2	1	0.006053	0.002419	2.50	0.0124
Stay_In_Current_City_Years_4+	1	0.004515	0.002655	1.70	0.0890
Age_18-25	1	-0.006706	0.001709	-3.92	<.0001
Age_36-45	1	0.004549	0.001503	3.03	0.0025
Age_46-50	1	-0.014100	0.004192	-3.36	0.0008
Age_51-55	1	0.045307	0.004479	10.12	<.0001
Age_55+	1	-0.011473	0.005790	-1.98	0.0475
Occupations 0	1	0.007859	0.002292	3.43	0.0006
Occupations 1	1	0.001101	0.002479	0.44	0.6570
Occupations 2	1	0.009725	0.002841	3.42	0.0006
Occupations 3	1	0.024900	0.003196	7.79	<.0001
Occupations 4	1	0.013088	0.002365	5.53	<.0001
Occupations 5	1	0.017143	0.003500	4.90	<.0001
Occupations 6	1	0.024674	0.003081	8.01	<.0001
Occupations 7	1	0.015199	0.002357	6.45	<.0001
Occupations 8	1	-0.004114	0.007973	-0.52	0.6058
Occupations 9	1	0.024797	0.004660	5.32	<.0001
Occupations 10	1	0.001562	0.003423	0.46	0.6482

Fig. 5.6

Occupations 11	1	0.010104	0.003717	2.72	0.0066
Occupations 12	1	0.024766	0.002623	9.44	<.0001
Occupations 13	1	0.011291	0.004775	2.36	0.0181
Occupations 14	1	0.019046	0.002774	6.87	<.0001
Occupations 15	1	0.029488	0.003493	8.44	<.0001
Occupations 16	1	0.020412	0.002888	7.07	<.0001
Occupations 17	1	0.019738	0.002473	7.98	<.0001
Occupations 18	1	0.008913	0.004536	1.97	0.0494
Occupations 19	1	-0.014837	0.004153	-3.57	0.0004
Occupations 20	0	0	.	.	.
Product1 1	1	-0.045479	0.016660	-2.73	0.0063
Product1 2	1	-0.104938	0.016826	-6.24	<.0001
Product1 3	1	-0.110036	0.017020	-6.47	<.0001
Product1 4	1	-0.789664	0.016932	-46.64	<.0001
Product1 5	1	-0.384028	0.016718	-22.97	<.0001
Product1 6	1	0.010645	0.016914	0.63	0.5291
Product1 8	1	-0.310056	0.016566	-18.72	<.0001
Product1 10	1	0.098212	0.017235	5.70	<.0001
Product1 11	1	-0.566654	0.017177	-32.99	<.0001
Product1 12	1	-1.094930	0.022376	-48.93	<.0001
Product1 13	1	-1.340025	0.017556	-76.33	<.0001
Product1 15	0	0	.	.	.
Product2 2	1	0.033754	0.007289	4.63	<.0001
Product2 3	1	0.100084	0.008990	11.13	<.0001
Product2 4	1	-0.034363	0.007800	-4.41	<.0001
Product2 5	1	0.019094	0.007386	2.59	0.0097

Product3 9	1	0.014343	0.003454	4.15	<.0001
Product3 10	1	-0.025870	0.006130	-4.22	<.0001
Product3 11	1	-0.020908	0.005230	-4.00	<.0001
Product3 12	1	0.013646	0.003607	3.78	0.0002
Product3 13	1	-0.020567	0.004039	-5.09	<.0001
Product3 14	1	0.009808	0.003152	3.11	0.0019
Product3 15	1	-0.012514	0.003081	-4.06	<.0001
Product3 16	1	0.007229	0.003056	2.37	0.0180
Product3 17	1	0.045409	0.003301	13.76	<.0001
Product3 18	0	0	.	.	.
Gender_F*City_Category_B	1	-0.018424	0.002730	-6.75	<.0001
Gender_F*City_Category_C	1	-0.012378	0.002845	-4.35	<.0001
City_Category_B*Age_18-25	1	0.006423	0.002434	2.64	0.0083
City_Category_B*Age_46-50	1	0.036708	0.004906	7.48	<.0001
City_Category_B*Age_51-55	1	-0.028857	0.005223	-5.52	<.0001
City_Category_B*Age_55+	1	0.054366	0.007633	7.12	<.0001
City_Category_C*Age_36-45	1	0.008316	0.002486	3.35	0.0008
City_Category_C*Age_46-50	1	0.024871	0.004987	4.99	<.0001
City_Category_C*Age_51-55	1	-0.028592	0.005333	-5.36	<.0001
City_Category_C*Age_55+	1	0.016010	0.006671	2.40	0.0164
City_Category_B*Stay_In_Current_City_Years_0	1	-0.020461	0.003397	-6.02	<.0001
City_Category_B*Stay_In_Current_City_Years_2	1	-0.005304	0.003062	-1.73	0.0832
City_Category_B*Stay_In_Current_City_Years_4+	1	-0.007422	0.003351	-2.22	0.0268
City_Category_C*Stay_In_Current_City_Years_0	1	-0.012653	0.003521	-3.59	0.0003
City_Category_C*Stay_In_Current_City_Years_2	1	0.000720	0.003175	0.23	0.8205
City_Category_C*Stay_In_Current_City_Years_4+	1	0.005716	0.003454	1.65	0.0980

Fig. 5.7

5.1 Results and Interpretations

We obtained few interesting results from the above regression and interpreted them to provide meaningful business insights.

Purchase behavior for different age groups

AGE(ref. 26-35)	A	B	C
18-25	-0.0067	-0.0003	-0.0067
36-45	0.0045	0.0045	0.0128
46-50	-0.0141	0.0226	0.0108
51-55	0.0453	0.0165	-0.0167
55+	-0.0114	0.0429	0.0045

These are the relative values of coefficients for different age groups for different city types keeping all other factors the same. The coefficient value for the 18-25 age group are negative irrespective of the city whereas it is positive for individuals in the age group of 36-45. This means that the average bill amount will be lower for 18-25 age group while it will be higher for the 36-45 age group.

This may be because the people in 18-25 age group are either unemployed or have a lower income base whereas the 36-45 age group individuals are family persons with relatively higher income base. So, may be the former, in general, buy cheaper products and thus lower bill amount while the latter buy branded/expensive products and thus higher bill amount.

Purchase behavior for gender

Gender(ref. Males)	A	B	C
F	0.0152	-0.0032	0.0028

Above table shows how gender affects the average percent change in bill amount for different city categories keeping all other factors the same. If a person is Female, then the average percentage change will be negative for city category B and positive for city category A and C

Purchase behavior for number of years spent in a city

City_years (ref. 1 year)	A	B	C
0	0.0099	-0.0106	-0.0028
2	0.0060	0.0007	0.0067
4+	0.0045	-0.0029	0.0102

The table gives relative value of coefficients for different city years for the three city categories. Other factors have been kept the same. If someone has spent 2 years in the city, the average percent increase in the bill amount is positive irrespective of the city.

This may be because people are buying bigger commodities around that time resulting in on average, higher bill amount.

6. Recommendations

- For higher average bill amount, retailers should target females in city category A and C and males in B
- Retailers looking to maximize revenue should target age group of 18-25 and the retailers specifically looking for higher average bill amount, should target age group of 36-45
- Retailers should maintain higher proportion of products (such as Air conditioner, etc.) which are preferred by the people who have stayed in the city for almost two years
- For city A and C make expensive products more accessible, while for city B make cheaper products more accessible

Appendix – A

```
PROC IMPORT OUT= WORK.BlackFriday DATAFILE="C:\Users\vxn180007\Desktop\SAS
Assignment\Project\BlackFriday.csv"
     DBMS=CSV REPLACE;
     GETNAMES=YES; DATAROW=2;
RUN;
data BlackFridaydata;
set BlackFriday;
Marital = put(Marital_Status, 1.);
Product1 = put(Product_Category_1, 2.);
Product2 = put(Product_Category_2, 2.);
Product3 = put(Product_Category_3, 2.);
Occupations = put(Occupation, 2.);
ProductID = put(Product_ID, 9.);
format purchase value dollar10.2;
run;
/*goptions colors=(red green blue);*/
/*gender count*/
proc gchart data=BlackFridaydata;
vbar Gender / type=freq;
title 'Count of males and females';
run;
/*average across gender*/
proc gchart data=BlackFridaydata;
vbar Gender / discrete type=mean
sumvar=Purchase mean;
title 'Average Spending across Gender';
run;
```

```
/*marital count*/
proc gchart data=BlackFridaydata;
vbar Marital / type=freq;
title 'Frequency Distribution of Marital Status';
run;
/*average across marital status*/
proc gchart data=BlackFridaydata;
vbar Marital / discrete type=mean
sumvar=Purchase mean;
title 'Average Spending across Marital Status';
run;
 /*age category count*/
proc gchart data=BlackFridaydata;
vbar Age / type=freq;
title 'Frequency distribution of Age Groups';
run;
/*average purchase across age*/
proc gchart data=BlackFridaydata;
vbar Age / discrete type=mean
sumvar=Purchase mean;
title 'Average Spending across Age Groups';
run;
```

```
/* Generate pie chart for city category count */
proc gchart data=Blackfridaydata;
  pie City_Category / type=freq
            other=0
            midpoints='A' 'B' 'C'
            value=none
            percent=arrow
            slice=arrow
            noheading;
   title 'Frequency Distribution of Cities';
run;

proc gchart data=BlackFridaydata;
hbar City_Category / discrete type=mean
sumvar=Purchase mean;
title 'Average Spending across City_Categories';
run;
 /*Stay_In_Current_City_Years  count*/
proc gchart data=BlackFridaydata;
vbar Stay_In_Current_City_Years / type=freq;
title 'Frequency Distribution of Years of Stay In Current City';
run;
/*average purchase across Stay_In_Current_City_Years */
proc gchart data=BlackFridaydata;
vbar Stay_In_Current_City_Years  / discrete type=mean
sumvar=Purchase mean;
title 'Average Spending across Years of Stay In Current City';
run;
```

```
/*Occupations count*/
proc gchart data=BlackFridaydata;
vbar Occupations  / type=freq;
title 'Frequency Distribution of Occupations';
run;
/*average across occupations*/
proc gchart data=BlackFridaydata;
vbar Occupations / discrete type=mean
sumvar=Purchase mean;
title 'Average Spending across Occupations';
run;
 /*gender and marital count  */
proc gchart data=Blackfridaydata;
   vbar Gender / subgroup=Gender group=Marital  type=freq
                legend=legend1 space=0 gspace=4
                maxis=axis1 raxis=axis2 gaxis=axis3;
title 'Frequency Distribition of Gender across Marital Status';
run;
/* gender and marital average sales*/
proc gchart data=Blackfridaydata;
   vbar gender / subgroup=gender group=Marital  discrete type=mean
sumvar=Purchase mean
                legend=legend1 space=0 gspace=4
                maxis=axis1 raxis=axis2 gaxis=axis3;
title 'Average Sales across Gender and Marital Status';
run;
```

```
/*age and gender count */
proc gchart data=Blackfridaydata;
  vbar Age / subgroup=Age group=Gender  type=freq
            legend=legend1 space=0 gspace=4
            maxis=axis1 raxis=axis2 gaxis=axis3;
title 'Frequency Distribition of Age Groups across Gender';
run;
/* age & gender average sales*/
proc gchart data=Blackfridaydata;
  vbar age / subgroup=age group=gender  discrete type=mean
sumvar=Purchase mean
            legend=legend1 space=0 gspace=4
            maxis=axis1 raxis=axis2 gaxis=axis3;
title 'Average Sales across Gender and Marital Status';
run;
/*city category & stay in years */
proc gchart data=Blackfridaydata;
  vbar City_Category / subgroup=City_Category group=Stay_In_Current_City_Years  type=freq
            legend=legend1 space=0 gspace=4
            maxis=axis1 raxis=axis2 gaxis=axis3;
title 'Frequency Distribition of City Categories and Years of Stay ';
run;
```

```
/* age & gender average sales*/
proc gchart data=Blackfridaydata;
  vbar City_Category / subgroup=City_Category group=Stay_In_Current_City_Years   discrete type=mean
sumvar=Purchase mean
              legend=legend1 space=0 gspace=4
              maxis=axis1 raxis=axis2 gaxis=axis3;
title 'Average Sales across City Categories and Years of Stay ';
run;
/*top selling wip P00110742*/
proc sql;
select Product_id, count(Product_id) as county from BlackFridaydata
group by Product_id
order by county desc;
quit;
goptions colors=(red green blue white yellow brown);
goptions reset=all border;

proc gchart data=Blackfridaydata;
hbar3d age / midpoints=(30 40 50)
freq freqlabel="Total in Group" subgroup=gender autoref
    maxis=axis2
    raxis=axis1
    legend=legend1
    coutline=black;
run;
quit;
```

```
proc gchart data=Blackfridaydata;
hbar3d age / midpoints= (30 40 50)
freq freqlabel="Total in Group" subgroup=gender autoref
sumvar=Purchase mean
    maxis=axis2
    raxis=axis1
    legend=legend1
    coutline=black;
run;
quit;
proc gbarline data=Blackfridaydata;
bar City_Category / discrete sumvar=Purchase space=4;
run;
quit;
data BlackFridaydata_CatA;
set Blackfridaydata;
where City_Category = 'A';
run;
data BlackFridaydata_CatB;
set Blackfridaydata;
where City_Category = 'B';
run;
data BlackFridaydata_CatC;
set Blackfridaydata;
where City_Category = 'C';
run;
```

```
axis1 label = ( f= "Arial/bold" "Product Subcategory") major= (Width = 20);
axis2 label = (f= "Calibri/bold" "Purchase") major = (width = 20);
run;
proc gchart data=BlackFridaydata_CatA;
format quarter roman.;
format Purchase dollar8.;
vbar3d Age / sumvar=Purchase mean subgroup=Gender inside=subpct
    outside=mean
    width=13
    space=10
    maxis=axis1
    raxis=axis2
    cframe=white
    legend=legend1;
run;
quit;
proc gchart data=BlackFridaydata_CatB;
format quarter roman.;
format Purchase dollar8.;
vbar3d Age / sumvar=Purchase mean subgroup=Gender inside=subpct
    outside=mean
    width=9
    space=4
    maxis=axis1
    raxis=axis2
    cframe=white
    legend=legend1;
run; quit;
```

```
proc gchart data=BlackFridaydata_CatC;
format quarter roman.;
format Purchase dollar8.;
vbar3d Age / sumvar=Purchase mean subgroup=Gender inside=subpct
    outside=mean
    width=9
    space=4
    maxis=axis1
    raxis=axis2
    cframe=white
    legend=legend1;
run;
quit;
proc gchart data=BlackFridaydata;
format quarter roman.;
format Purchase dollar8.;
block Gender / sumvar=Purchase
type=mean
group=City_Category
subgroup=Marital
legend=legend1
noheading;
run;
quit;
```

```
proc gchart data=BlackFridaydata;
format Purchase dollar8.;
vbar3d City_Category / sumvar=Purchase subgroup=Marital inside=subpct
    outside=mean
    width=9
    space=4
    maxis=axis1
    raxis=axis2
    cframe=gray
    legend=legend1;
run;
quit;
proc gchart data=BlackFridaydata_CatA;
star Stay_In_Current_City_Years / sumvar=Purchase type = mean;
run;
quit;
proc gchart data=BlackFridaydata_CatB;
star Stay_In_Current_City_Years / sumvar=Purchase type = mean;
run;
quit;
proc gchart data=BlackFridaydata_CatC;
star Stay_In_Current_City_Years / sumvar=Purchase type = mean;
run;
quit;
```

```
proc gchart data=BlackFridaydata_CatA;
format Purchase dollar8.;
vbar3d Occupations / sumvar=Purchase subgroup=Gender inside=subpct
    shape = cylinder
    outside=mean
    width=9
    space=4
    maxis=axis1
    raxis=axis2
    cframe=White
    legend=legend1;
run;
quit;
proc gchart data=BlackFridaydata_CatB;
format Purchase dollar8.;
vbar3d Occupations / sumvar=Purchase subgroup=Gender inside=subpct
    shape = cylinder
    outside=mean
    width=9
    space=4
    maxis=axis1
    raxis=axis2
    cframe=White
    legend=legend1;
run;
quit;
```

```
proc gchart data=BlackFridaydata_CatC;
format Purchase dollar8.;
vbar3d Occupations / sumvar=Purchase subgroup=Gender inside=subpct
    shape = cylinder
    outside=mean
    width=9
    space=4
    maxis=axis1
    raxis=axis2
    cframe=White
    legend=legend1;
run;
quit;

proc gchart data=Blackfridaydata;
hbar3d City_Category / midpoints=(30 40 50)
freq freqlabel="Total in Group" subgroup=age autoref
    maxis=axis2
    raxis=axis1
    legend=legend1
    coutline=black;
run;
quit;
```

```
goptions reset = global;
proc gchart data=BlackFridaydata;
format Purchase dollar8.;
vbar3d Product1 / sumvar=Purchase subgroup=City_Category inside=subpct
    outside=mean
    width=9
    space=4
    maxis=axis1
    raxis=axis2
    cframe=white
    legend=legend1;
run;
quit;
proc gchart data=BlackFridaydata;
format Purchase dollar8.;
vbar3d Product2 / sumvar=Purchase subgroup=City_Category inside=subpct
    outside=mean
    width=9
    space=4
    maxis=axis1
    raxis=axis2
    cframe=White
    legend=legend9;
run;
quit;
```

```
proc gchart data=BlackFridaydata;
format Purchase dollar8.;
hbar3d Product3 / sumvar=Purchase subgroup=City_Category inside=subpct
    outside=mean
    width=9
    space=4
    maxis=axis1
    raxis=axis2
    cframe=White
    legend=legend1;
run;
quit;
proc gchart data=BlackFridaydata;
  pie Product1  /
          across=2
          clockwise value=none
          slice=outside percent=outside;
run;
quit;
proc gchart data=BlackFridaydata;
  pie Product2  /
          across=2
          clockwise value=none
          slice=outside percent=outside;
run;
quit;
```

```
proc gchart data=BlackFridaydata;
  pie Product3  /
          across=2
          clockwise value=none
          slice=outside percent=outside;
run;
quit;
proc surveyselect data=BlackFridaydata out=bf_sampled outall samprate=0.8 seed=2;
run;
 data bf_training bf_test;
   set bf_sampled;

   if selected then
      output bf_training;
   else
      output bf_test;
run;
Data bf_training;
set bf_training;
log_purchase = log10(purchase);
run;
Data bf_test;
set bf_test;
log_purchase = log10(purchase);
run;
```

```
/*ASE in train vs. test data */

/* forward selection with cross validation as criteria with 10-folds */

proc glmselect data=bf_training testdata=bf_test seed = 2  plots=all;

class Gender (split) City_Category (split) Stay_In_Current_City_Years (split) Age  (split) Marital (split)

Product1 Product2 Product3 ProductID Occupations;

model log_purchase = Gender City_Category Stay_In_Current_City_Years Age Marital Occupations
Product1

Product2 Product3  City_Category*Gender City_Category*Age City_Category*Marital
City_Category*Stay_In_Current_City_Years

  /selection=forward(select=cv) hierarchy=single cvmethod=random(10)showpvalues ;

performance buildsscp=incremental;

run;


/*ASE in train vs. test data */

/* backward selection with cross validation as criteria with 10-folds */


proc glmselect data=bf_training testdata=bf_test seed = 2  plots=all;

  class Gender (split) City_Category (split) Stay_In_Current_City_Years (split) Age  (split) Marital (split)

Product1 Product2 Product3 ProductID Occupations;

model log_purchase = Gender City_Category Stay_In_Current_City_Years Age Marital  Product1
Product2 Product3 ProductID

  /selection=backward(select=cv) hierarchy=single cvmethod=random(10)showpvalues ;

performance buildsscp=incremental;

run;

/*ASE in train vs. test data */

/* stepwise selection with cross validation as criteria with 10-folds */
```

```
proc glmselect data=bf_training testdata=bf_test seed = 2  plots=all;
  class Gender (split) City_Category (split) Stay_In_Current_City_Years (split) Age  (split) Marital (split)
Product1 Product2 Product3 ProductID Occupations;
model log_purchase = Gender City_Category Stay_In_Current_City_Years Age Marital Occupations
Product1 Product2 Product3 ProductID
  /selection=stepwise(select=cv) hierarchy=single cvmethod=random(10)showpvalues ;
performance buildsscp=incremental;
run;


proc HPFOREST data=bf_training;
    target purchase / level= interval;
    input gender Marital/ level = binary;
    input city_category Stay_In_Current_City_Years Product1 Occupations Age Product2 Product3
ProductID/ level= ordinal;
run;
proc glmselect data=bf_training testdata=bf_test seed = 2  plots=all;
  class Gender (split) City_Category (split) Stay_In_Current_City_Years (split) Age  (split) Marital (split)
Product1 Product2 Product3 ProductID Occupations;
model log_purchase = Gender City_Category Stay_In_Current_City_Years Age Marital Occupations
Product1 Product2 Product3 ProductID
  /selection=lasso(choose=cv stop=none) hierarchy=single cvmethod=random(10)showpvalues ;
performance buildsscp=incremental;
run;
proc glmselect data=bf_training testdata=bf_test seed = 2  plots=all;
class Gender (split) City_Category (split) Stay_In_Current_City_Years (split) Age  (split) Marital (split)
Product1 Product2 Product3 ProductID Occupations;
model log_purchase = Gender City_Category Stay_In_Current_City_Years Age Marital Occupations
Product1 Product2 Product3 ProductID
/selection=forward(select=cv) hierarchy=single cvmethod=random(10)showpvalues ;
```

```
modelaverage nsamples=50 tables=(ParmEst(all));
performance buildsscp=incremental;
score data=bf_test out=test_performance residual=res;
run;
```

Appendix – C

Citations: https://www.retailcustomerexperience.com/blogs/black-friday-4-ways-retailers-can-prepare/