**Submitted towards partial fulfilment of the criteria**

**For award of PGP-DSE by GLIM**

**Submitted By**

Group No. 8 [Batch: 2018-19]

**Group Members**

Shahrukh Buland Iqbal – DSEFTCJUL18042

Arjun Thumbayil – DSEFTCJUL18006

Sahil Bansal - DSEFTCJUL18014

**Research Supervisor**

P V Subramanian

**Project Report on**

**“Purchase Prediction on Black Friday”**



**ABSTRACT**

This report is prepared solely for academic project purpose. Winter holiday shopping accounts for 25–40% of the total U.S. annual retail sales (NRF, 2011). Because consumers spend so much during this period, retailers anxiously look forward to the holiday shopping season in order to achieve positive financial results for the year. During the holiday season, retailers rely heavily on advertising and consumer word-of-mouth to bring customers into stores and online to shop. In order to maximize their efforts, retailers are eager to understand consumer buying decisions that will aid them in achieving profits during this time. The purpose of the present study is to predict consumer spending on Black Friday so that retailers can make better holiday promotion decisions. Understanding the shopping habits of holiday shoppers will allow retailers to realize why some consumers spend more than others on Black Friday. The contribution of this Analysis is to understand Consumer motivations to spend on Black Friday and the factors behind this motivation. The objective is to understand theoretically these motivations in order to provide needed answers to retailers in terms of how to base their holiday marketing budgets.

After discussions and evaluations along with our mentor PVS, we have planned to apply Multiple Linear Regression model using Python tool for analysis, on a dataset which will suffice our problem statement. Dataset for this project, is pulled out from KAGGLE website.

We have applied 6 different models in which the data is processes using several feature engineering approaches based on analysis and understanding of the data. Categorical Regression, Multiple Linear Regression, Ridge Linear Regression and Lasso Linear Regression are applied to predict the Purchase amount by a Customer in US Dollars. Root Mean Squared Error(RMSE) has been assessed after model parameter optimization and feature selection.

**Acknowledgements**

We are very much glad to place on record our gratitude and appreciation for the guidance and help provided to us by our Mentor **Mr. PV Subramanian** for completing this project. His experience and support guided us to sail through the research process and complete it on time. We would like to extend our sincere thanks to all the faculties for their teaching and guidance and our dear class mates for the support and advice given.

We also wish to thank Project Office, Chennai for the support.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

**Date: January 6, 2019**

**Place: Chennai**

**CERTIFICATE OF COMPLETION**

I hereby certify that the project titled **“Purchase Prediction on Black Friday”** for case resolution was undertaken and completed under my supervision by **Shahrukh Buland Iqbal**, **Arjun Thumbayil** and **Sahil Bansal** of Post Graduate Program in Data Science Engineering (PGP-DSE).

**P V Subramanian**

**Date: Jan 6, 2019**

**Place: Chennai**

Table of Contents

[1. Introduction 7](#_Toc534578585)

[1.1 Objective 7](#_Toc534578586)

[1.1.1 Predicting Purchase 7](#_Toc534578587)

[1.1.2 Pattern Recognition 7](#_Toc534578588)

[1.2 Motivation 8](#_Toc534578589)

[2. Dataset 8](#_Toc534578590)

[2.1 Data Collection 8](#_Toc534578591)

[2.2 Data Description 8](#_Toc534578592)

[2.3 Data Pre-processing 9](#_Toc534578593)

[2.4 Handling Missing Values 9](#_Toc534578594)

[2.5 Handling Outliers 9](#_Toc534578595)

[2.6 Exploratory Data Analysis 10](#_Toc534578596)

[2.6.1 5 Point Summary 10](#_Toc534578597)

[2.6.2 Bivariate Analysis 10](#_Toc534578598)

[2.6.3 Multivariate Analysis 15](#_Toc534578599)

[2.6.4 Statistical Analysis 20](#_Toc534578600)

[3. Feature Engineering 22](#_Toc534578601)

[3.1 Data Conversion 22](#_Toc534578602)

[3.2 Data Binning 22](#_Toc534578603)

[3.3 Feature Creation 25](#_Toc534578604)

[3.3.1 Based on Average Feature Purchase 25](#_Toc534578605)

[3.3.2 Based on Feature Frequency 26](#_Toc534578606)

[3.4 Polychotomization 26](#_Toc534578607)

[3.5 Response Transformation 27](#_Toc534578608)

[4. Model Development 28](#_Toc534578609)

[4.1 Regression Model 28](#_Toc534578610)

[4.1.1 Model Development 29](#_Toc534578611)

[4.1.2 Model Evaluation 29](#_Toc534578612)

[4.1.3 Model Optimization 29](#_Toc534578613)

[4.2 Statistical Learning 35](#_Toc534578614)

[5. Recommendations: 36](#_Toc534578615)

[5.1 Based on Descriptive Analytics: 36](#_Toc534578616)

[5.2 Based on Predictive Analytics 37](#_Toc534578617)

[5.3 Based on Prescriptive Analytics 37](#_Toc534578618)

[6 Conclusions: 37](#_Toc534578619)

[7 References 37](#_Toc534578620)

[8 APPENDIX 38](#_Toc534578621)

[8.1 Python Code For Statistical Analysis 38](#_Toc534578622)

[8.2 Python Code For Feature Engineering 38](#_Toc534578623)

[8.3 Python Code For The Models 38](#_Toc534578624)

**List of Tables and figures**

[**Table 1** **DATA VARIABLES** 8](#_Toc534578804)

[**Table 2** **PURCHASE Five Point Summary** 10](#_Toc534578805)

[**Table 3: Age – City Category** 16](#_Toc534578806)

[**Table 4: AGE - GENDER** 16](#_Toc534578807)

[**Table 5: Age – Marital Status** 17](#_Toc534578808)

[**Table 6: City – Gender** 17](#_Toc534578809)

[**Table 7: City – Marital Status** 18](#_Toc534578810)

[**Table 8: CITY – OCCUPATION** 18](#_Toc534578811)

[**Table 9: CITY – STAY** 19](#_Toc534578812)

[**Table 10: Chi Square Analysis** 21](#_Toc534578813)

[**Figure 1: AGE V/S PURCHASE** 11](#_Toc534578814)

[**Figure 2: CITY V/S PURCHASE** 12](#_Toc534578815)

[**Figure 3: GENDER V/S PURCHASE** 12](#_Toc534578816)

[**Figure 4:MARITAL STATUS V/S PURCHASE** 13](#_Toc534578817)

[**Figure 5: OCCUPATION V/S PURCHASE** 14](#_Toc534578818)

[**Figure 6: PRODUCT\_CATEGORY\_1 V/S PURCHASE** 15](#_Toc534578819)

[**Figure 7: Pre Binning Occupations** 23](#_Toc534578820)

[**Figure 8: Post Binning Occupation** 24](#_Toc534578821)

[**Figure 9: Pre Binning Product Category 1** 25](#_Toc534578822)

[**Figure 10: Post Binning Product Category 1** 25](#_Toc534578823)

[**Figure 11: Target Distribution** 28](#_Toc534578824)

[**Figure 12: Normal Probability Plot of log(Purchase) indicates normality.** 29](#_Toc534578825)

[**Figure 13: Results** 30](#_Toc534578826)

[**Figure 14: Residual Distribution** 33](#_Toc534578827)

[**Figure 15: Normal Probability Plot of Residuals** 34](#_Toc534578828)

[**Figure 16: Visible Heteroskedasticity** 35](#_Toc534578829)

[**Figure 17: Removed Heteroskedasticity** 35](#_Toc534578830)

# **Introduction**

The day after Thanksgiving in the U.S. is called Black Friday (BF) and serves as the  
traditional start to the holiday shopping season (Shay, 2015). BF represents a unique  
consumption ritual that blends traditional shopping for better deals with holiday rituals for  
social relations (Thomas and Peters, 2011). Known for deep discounts (e.g., doorbusters), BF  
shopping manifests adventure, competition and urgency around getting great deals. With  
doorbuster deals and festive shopping environments on BF, many families in the U.S. have  
come to enjoy BF as one of the popular social activities during the Thanksgiving holidays  
(Thomas and Peters, 2011). Although online shopping on BF outpaced instore shopping in  
2016 (Wahba, 2016), BF is still important (David, 2016) for several reasons: it is one of the  
most important shopping days for malls and stores, it is the biggest sales day of the  
Thanksgiving weekend, and more people shop on BF than any other day during Thanksgiving  
week (Halkias, 2016). Although Cyber Monday is gaining popularity, BF shopping continues  
to be popular because of an abundance of doorbuster deals, instant gratification, and the  
benefit of social shopping (VerHaar, 2015). Some shoppers are loyal to BF and anticipate  
having fun shopping on BF while getting doorbuster deals (Sander, 2013)

## **Objective**

### **1.1.1 Predicting Purchase**

* Based on the given Data, we are attempting to make a model that shall be able to predict a Customer’s Purchase.
* The end goal is to produce a Machine Learning model that will be able to account for the Purchase of a particular Customer based on several input factors such as Age, Sex, Years spend in the City, etc.

### **1.1.2 Pattern Recognition**

* We study various factors that can cause a certain customer to spend more money during the Purchase.
* In order to do so, we have carried out Univariate, Bivariate and Statistical Analysis to understand Customer Behaviour.

## **1.2 Motivation**

* In order to compete with Online Shopping Platforms, Brick and Mortar based Retailers need to figure out how to boost Sales during the most important Shopping Day of the Year.
* By understanding the Purchase Patterns of the Customers Retailers can provide improved Service Quality
* Improve Staffing and Inventory of the Retail Store.
* Increase Revenue and Sales.

# **Dataset**

## **2.1 Data Collection**

The data comes from a competition hosted by Analytics Vidhya(<https://www.kaggle.com/mehdidag/black-friday/home>). The Dataset comprises of 550000 observations about the Black Friday in a retail store, it contains various kinds of variables either Numeric or Categorical in nature. The dataset contains 2 columns with missing values:

1. 166986 observations missing in column ‘Product\_Category\_2’.
2. 373299 observations missing in column ‘Product\_Category\_3’.

## **2.2 Data Description**

**Table 1 DATA VARIABLES**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Type** | **Subtype** | **Description** | **Segment** |
| User\_ID | Integer | Discrete | To be converted as Categorical Variable since it is an ID Variable | Customer |
| Product\_ID | Categorical | Discrete | Product ID | Product |
| Gender | Categorical | Nominal | Sex of User | Customer |
| Age | Categorical | Ordinal | Age in Bins | Customer |
| Occupation | Categorical | Nominal | Occupation listed as numbers(Masked) | Customer |
| City\_Category | Categorical | Nominal | Category of the City(A,B,C) | City |
| Stay\_In\_Current\_City\_Years | Categorical | Ordinal | Number of Years Stay in Current City. | City |
| Marital\_Status | Categorical | Nominal | 0 – UnMarried, 1- Married | Customer |
| Product\_Category\_1 | Categorical | Nominal | Product Category(Masked) | Product |
| Product\_Category\_2 | Categorical | Nominal | Products may belong to other category also(Masked) | Product |
| Product\_Category\_3 | Categorical | Nominal | Product may belongs to other category also(Masked) | Product |
| Purchase | Integer | Continuous | Purchase Amount in Dollars(Target Variable) | Product |

## **2.3 Data Pre-processing**

Most of the raw data contained in any given Dataset is usually unprocessed, incomplete, and noisy. For, example, the dataset may contain:

* Fields that are obsolete or redundant;
* Missing values;
* Outliers;
* Data in a form not suitable for the data mining models;
* Values not consistent with policy or common sense.

In order to be useful for data mining purposes, the databases need to undergo  
pre-processing, in the form of *data cleaning* and *data transformation*.

## **2.4 Handling Missing Values**

The Literature (Shmueli, 2010) suggests that if only 5% of Data values are missing from a dataset of 30 variables, and the missing values are spread evenly throughout the data, almost 80% of the records would have at least one missing value. Therefore, we chose to drop the columns ‘Product\_Category\_2’ and ‘Product\_Category\_3’. Both the columns have more than 30% of values missing.

## **2.5 Handling Outliers**

Outliers are extreme values that go against the trend of the remaining data. Identifying  
outliers is important because they may represent errors in data entry. Also, even if an  
outlier is a valid data point and not an error, certain statistical methods are sensitive  
to the presence of outliers, and may deliver unreliable results.

For initial Outlier detection we have used the IQR method for Outlier Detection. The quartiles of a data set divide the data set into the following four parts, each containing 25% of the data:

• The first quartile (Q1) is the 25th percentile.

• The second quartile (Q2) is the 50th percentile, that is, the median.

• The third quartile (Q3) is the 75th percentile.

Then, the IQR is a measure of variability, much more robust than the SD. The

IQR is calculated as IQR = Q3 − Q1, and may be interpreted to represent the spread

of the middle 50% of the data.

A robust measure of outlier detection is therefore defined as follows. A data

value is an outlier if:

a. it is located 1.5(IQR) or more below Q1, or

b. it is located 1.5(IQR) or more above Q3.

For example, suppose for a set of test scores, the 25th percentile was Q1 = 70

and the 75th percentile was Q3 = 80, so that half of all the test scores fell between 70

and 80. Then the interquartile range, or the difference between these quartiles was

IQR = 80 − 70 = 10.

A test score would be robustly identified as an outlier if

a. it is lower than Q1 − 1.5(IQR) = 70 − 1.5(10) = 55, or

b. it is higher than Q3 + 1.5(IQR) = 80 + 1.5(10) = 95.

## **Exploratory Data Analysis**

### **5 Point Summary**

**Table 2 PURCHASE Five Point Summary**

|  |  |
| --- | --- |
| Statistical Parameter | Purchase |
| Mean | 9333.859 |
| Standard Deviation | 4981.02 |
| Median | 8062 |
| Minimum | 185. |
| Maximum | 23961 |

In the distribution, the mean (µ = 9333.859) was higher than the median, indicating a positive skew.

### **Bivariate Analysis**

#### **2.6.2.1 AGE-PURCHASE**

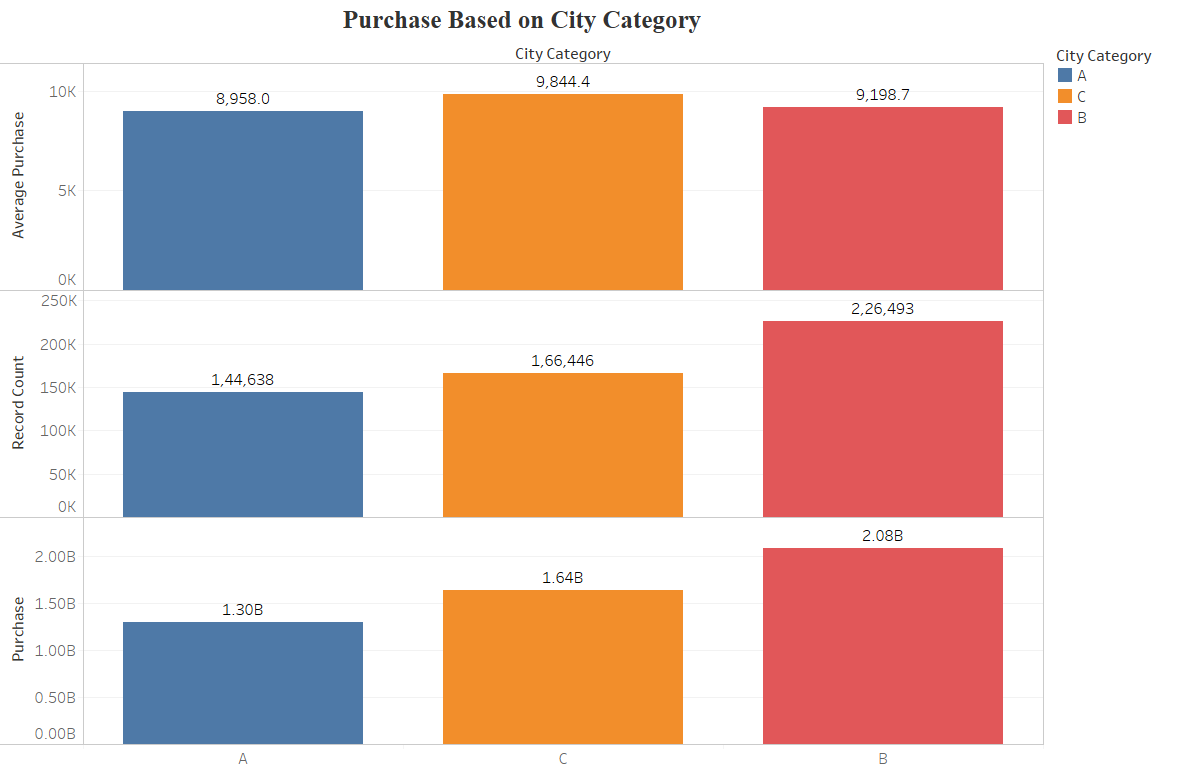


**Figure 1: AGE V/S PURCHASE**

**Observations:**

* 40% of the shopping i.e., almost 2 Billion $ is spent by the age group **’26-35’**.
* **‘0-17’** age group account for the least Sales i.e., 132 Million $ only.
* Interestingly, all the age groups spend on an average approx. 9000 $

#### **2.6.2.2 CITY – PURCHASE**

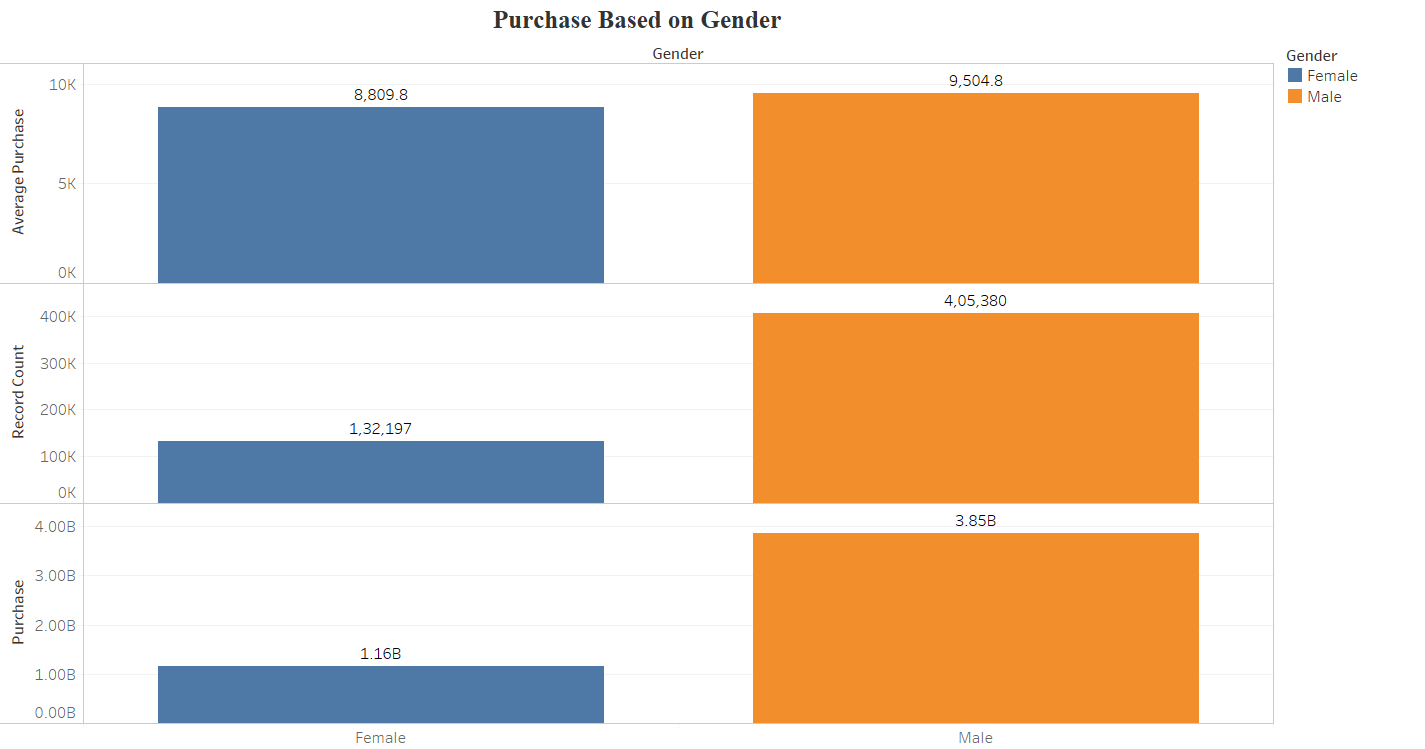


**Figure 2: CITY V/S PURCHASE**

**Observations:**

* City-C accounts for highest Purchase and City-A for the lowest amongst all the cities.
* City-B on an average spend 9844 $

#### 2.6.2.3 GENDER – PURCHASE

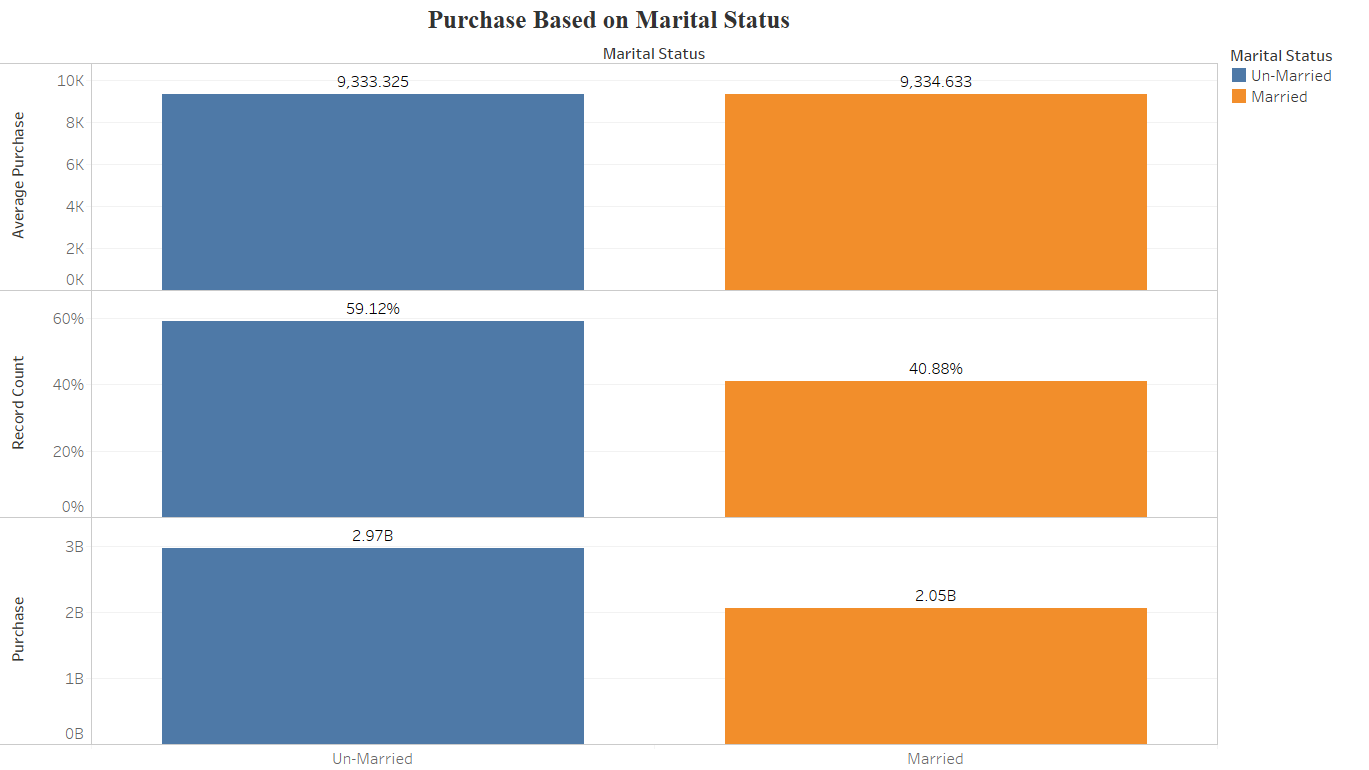


**Figure 3: GENDER V/S PURCHASE**

**Observations:**

* Male Shopper account for almost 75% of the Purchases both in terms of Frequency and Spending.
* Male Shoppers on an average spend almost 700$ more than Female Shoppers.

#### 2.6.2.4 MARITAL STATUS – PURCHASE

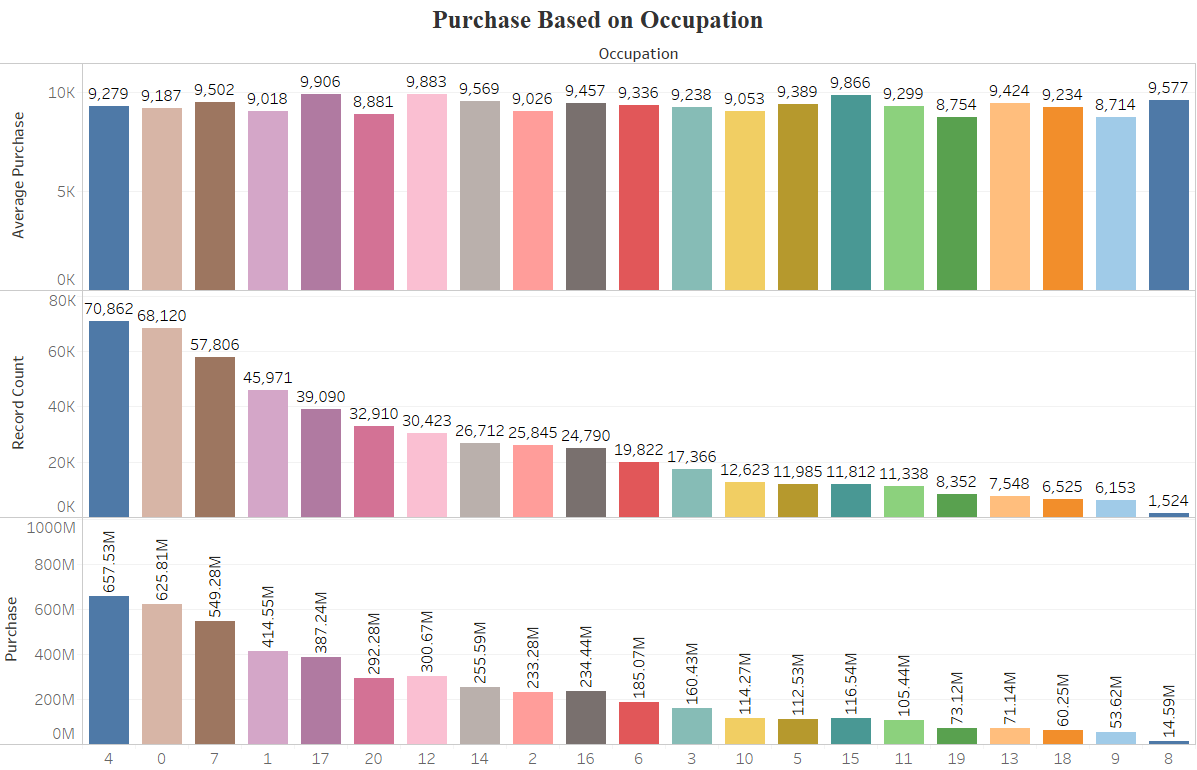


**Figure 4:MARITAL STATUS V/S PURCHASE**

**Observations:**

* Unmarried people account for an excess of 1 B$ Purchase than Married people.
* Interestingly, Average Purchase committed by Married and Unmarried individuals is the same i.e., approx. 9333$

#### 2.6.2.5 OCCUPATION - PURCHASE

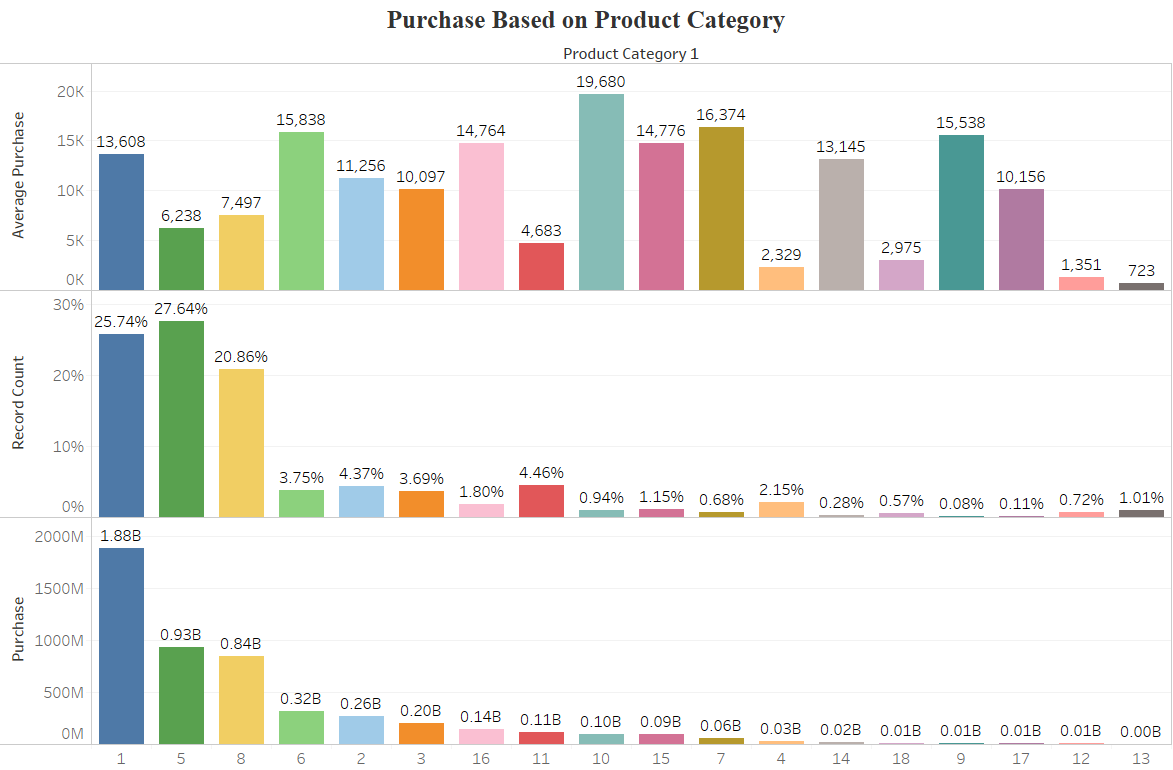


**Figure 5: OCCUPATION V/S PURCHASE**

**Observations:**

* Occupation – **4,0 and 7** are responsible for purchase above 500 M$ each.
* There are 13 Occupation Categories within which Customers spend between 100-500 M$.
* Occupation – **19,13,18,9 and 8** in total account for less than 100 M$ Purchase.

#### 2.6.2.6 PRODUCT CATEGORY 1 - PURCHASE



**Figure 6: PRODUCT\_CATEGORY\_1 V/S PURCHASE**

**Observations:**

* Product Category -1 accounts for highest Revenue and is the 2nd most popular product in terms of Sales. Also, Customers spend 4275$ more on it than Average Purchase.
* Product Category-5 is the most popular Product in terms of Sales.
* On Average, Customers are ready to spend the most on Product Category-10.
* Product Category – 12 and 13 have Low Sales and Lowest Average Spending.

### 2.6.3 Multivariate Analysis

#### 2.6.3.1 Age – City Category

**Table 3: Age – City Category**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **City Category** | | | | | |
|  | **A** | | **B** | | **C** | |
| **Age** | **% City Population** | **Avg. Purchase** | **% City Population** | **Avg. Purchase** | **% City Population** | **Avg. Purchase** |
| 0-17 | 1.73% | 8,673 | 2.33% | 8,985 | 4.16% | 9,172 |
| 18-25 | 18.68% | 8,887 | 18.75% | 9,070 | 16.91% | 9,820 |
| 26-35 | 49.81% | 8,990 | 39.63% | 9,199 | 31.77% | 9,953 |
| 36-45 | 18.07% | 9,042 | 20.58% | 9,150 | 20.88% | 10,009 |
| 46-50 | 5.16% | 8,386 | 8.79% | 9,297 | 10.31% | 9,662 |
| 51-55 | 4.13% | 9,575 | 7.70% | 9,394 | 8.54% | 9,918 |
| 55+ | 2.41% | 8,587 | 2.22% | 9,886 | 7.44% | 9,523 |

**Observations:**

* Age group **26-35**  is the most frequent shopper group.
* In City-C, except Teens all the Age Group spend above the Average.
* The Age Groups **’18-45’** account for 70% and above Product shopping in all the cities.

#### 2.6.3.2 Age – Gender

**Table 4: AGE - GENDER**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Gender** | | | |
|  | **Female** | | **Male** | |
| **Age** | **% Shoppers** | **Avg. Purchase** | **% Shoppers** | **Avg. Purchase** |
| 0-17 | 3.75% | 8,445 | 2.41% | 9,312 |
| 18-25 | 18.20% | 8,405 | 18.15% | 9,507 |
| 26-35 | 37.33% | 8,792 | 40.79% | 9,471 |
| 36-45 | 19.99% | 9,047 | 20.00% | 9,517 |
| 46-50 | 9.72% | 8,929 | 7.81% | 9,429 |
| 51-55 | 7.29% | 9,131 | 6.90% | 9,789 |
| 55+ | 3.73% | 9,120 | 3.94% | 9,557 |

**Observations:**

* Male Customers on an average spend more than Female Customers.
* Teenage Girls spend less than older Women.

#### 2.6.3.3 Age – Marital Status

**Table 5: Age – Marital Status**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Marital Status** | | | |
|  | **Un-Married** | | **Married** | |
| **Age** | **% Shopper** | **Avg. Purchase** | **% Shopper** | **Avg. Purchase** |
| 0-17 | 4.63% | 9,020 |  |  |
| 18-25 | 24.23% | 9,281 | 9.39% | 9,065 |
| 26-35 | 41.07% | 9,313 | 38.30% | 9,316 |
| 36-45 | 20.45% | 9,470 | 19.34% | 9,296 |
| 46-50 | 3.88% | 9,035 | 14.65% | 9,381 |
| 51-55 | 3.35% | 9,664 | 12.28% | 9,604 |
| 55+ | 2.40% | 9,660 | 6.04% | 9,335 |

**Observations:**

* Older people are likely to spend more irrespective of their marital status.

#### 2.6.3.4 City – Gender

**Table 6: City – Gender**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Gender** | | | |
|  | **Female** | | **Male** | |
| **City Category** | **% Shoppers** | **Avg. Purchase** | **% Shoppers** | **Avg. Purchase** |
| A | 26.33% | 8,631 | 27.09% | 9,062 |
| B | 42.73% | 8,591 | 41.94% | 9,401 |
| C | 30.94% | 9,265 | 30.97% | 10,033 |

**Observations:**

* Male Customers : *City C* customers are likely to spend the *most*, *City A* customers are likely to spend the *least*.
* Female Customers: *City B* are likely to spend the *least* whereas *City C* the *most*.

#### 2.6.3.5 City – Marital Status

**Table 7: City – Marital Status**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Marital Status** | | | |
|  | **Un-Married** | | **Married** | |
| **City Category** | **% Shoppers** | **Avg. Purchase** | **% Shoppers** | **Avg. Purchase** |
| A | 28.13% | 9,021 | 25.14% | 8,856 |
| B | 42.23% | 9,192 | 41.99% | 9,208 |
| C | 29.64% | 9,830 | 32.87% | 9,863 |

**Observations:**

* Irrespective of Marital Status, People in City C spend the most and City A the least.

#### 2.6.3.6 CITY – OCCUPATION

**Table 8: CITY – OCCUPATION**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **City Category** | | | | | |
|  | **A** | | **B** | | **C** | |
| **Occupation** | **% Shoppers** | **Avg. Purchase** | **% Shoppers** | **Avg. Purchase** | **% Shoppers** | **Avg. Purchase** |
| 0 | 12.72% | 8,919 | 12.96% | 9,078 | 12.24% | 9,586 |
| 1 | 8.57% | 8,778 | 8.52% | 8,802 | 8.57% | 9,517 |
| 2 | 6.15% | 8,588 | 4.90% | 8,847 | 3.51% | 10,032 |
| 3 | 3.84% | 8,821 | 3.02% | 9,281 | 2.98% | 9,645 |
| 4 | 16.50% | 8,989 | 12.94% | 9,107 | 10.63% | 9,954 |
| 5 | 1.62% | 8,937 | 2.88% | 9,132 | 1.87% | 10,266 |
| 6 | 2.54% | 9,302 | 4.72% | 9,152 | 3.29% | 9,720 |
| 7 | 10.93% | 8,884 | 9.98% | 9,442 | 11.65% | 10,076 |
| 8 | 0.07% | 11,443 | 0.37% | 9,480 | 0.36% | 9,397 |
| 9 | 0.48% | 8,985 | 1.40% | 8,581 | 1.37% | 8,818 |
| 10 | 1.51% | 8,662 | 1.82% | 8,953 | 3.79% | 9,254 |
| 11 | 1.66% | 9,382 | 2.42% | 8,704 | 2.07% | 10,190 |
| 12 | 4.79% | 9,496 | 5.68% | 9,866 | 6.39% | 10,155 |
| 13 | 0.27% | 8,684 | 1.03% | 8,844 | 2.90% | 9,763 |
| 14 | 5.27% | 9,483 | 4.88% | 9,323 | 4.83% | 9,987 |
| 15 | 2.07% | 9,807 | 2.24% | 9,788 | 2.25% | 10,019 |
| 16 | 4.09% | 9,450 | 4.71% | 9,201 | 4.93% | 9,794 |
| 17 | 5.57% | 9,140 | 6.93% | 9,814 | 9.22% | 10,402 |
| 18 | 1.10% | 8,871 | 0.92% | 8,817 | 1.70% | 9,745 |
| 19 | 1.53% | 8,362 | 1.49% | 8,900 | 1.67% | 8,890 |
| 20 | 8.72% | 8,361 | 6.19% | 9,064 | 3.77% | 9,518 |

**Observations:**

* In City-C, 18 different Occupation type People spend above average i.e., more than 93%
* In City-A and City-B, not more than 6 types of Occupation Type are observed to spend above Average.

#### 2.6.3.7 CITY – STAY

**Table 9: CITY – STAY**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **City Category** | | | | | |
| **A** | | **B** | | **C** | |
| **Stay In Current City Years** | **% Shoppers** | **Avg. Purchase** | **% Shoppers** | **Avg. Purchase** | **% Shoppers** | **Avg. Purchase** |
| 0 | 16.39% | 9,029 | 12.43% | 8,947 | 12.55% | 9,900 |
| 1 | 33.30% | 8,928 | 36.04% | 9,225 | 35.69% | 9,768 |
| 2 | 18.35% | 9,015 | 18.01% | 9,251 | 19.29% | 9,900 |
| 3 | 16.86% | 8,957 | 18.53% | 9,231 | 16.20% | 9,893 |
| 4+ | 15.10% | 8,880 | 15.00% | 9,240 | 16.27% | 9,856 |

**Observations:**

* Customers who arrived recently in City-B and City-C shop less frequently than those who stayed longer (Acclimatization can be an issue).

#### 2.6.3.8 GENDER-OCCUPATION

**Observations:**

* Female Customers of Occupation category – **0,1 and 4** are likely to shop the most.
* Female Customers of Occupation category – **8,15,17,18** are likely to spend above average.
* Male Customers of Occupation category – **0,4 and 7** are likely to shop the most.
* Male Customers of Occupation category – **4,5,6,7,8,10,12,13,14,15,16 and 17** are likely to spend above average.

#### 2.6.3.9 MARITAL STATUS – OCCUPATION

**Observations:**

* On an Average, Unmarried people of Occupation – **9** spend the least and of Occupation – 15 spend the most.
* On an Average, Married people of Occupation – **2** spend the least and of Occupation – 17 spend the most.

#### 2.6.3.10 OCCUPATION – STAY

**Observations:**

* People who came recently to a City are likely to spend more on Purchase than People who have been residing in the City for more than a Year.
* Occupation – **0,1,4 and 7** are the most frequent Shoppers, irrespective of duration of Stay in a City.

#### 2.6.3.11 PRODUCT

**Observations:**

1. Irrespective of any other variable:
   * Product Category – **1,2,3,6,7,9,10,14,15** and **16** record an above Average spending by the Customers.
   * Product Category –**4,5,8,11,12,13** and **18** record a below Average spending by the Customers.
   * Product Category – **1,5,** and **8** are the most popular products i.e., these Products are bought more frequently by the Customers.
   * Product Category – **9,14, 17**and **18** are the least popular products i.e., these Products are bought less frequently by the Customers.

### Statistical Analysis

#### Chi-Square Test of Independence

H₀ : Category-1 and Category-2 are independent, or unrelated

Hₐ : Choice of Category-1 depends on, or is contingent upon, Category-2

**Table 10: Chi Square Analysis**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **AGE** | **CITY CATEGORY** | **GENDER** | **MARITAL STATUS** | **OCCUPATION** | **PRODUCT CATEGORY-1** | **STAY** |
| **AGE** | YES | YES | YES | YES | YES | YES | YES |
| **CITY CATEGORY** | YES | YES | YES | YES | YES | YES | YES |
| **GENDER** | YES | YES | YES | YES | YES | YES | YES |
| **MARITAL STATUS** | YES | YES | YES | YES | YES | YES | YES |
| **OCCUPATION** | YES | YES | YES | YES | YES | YES | YES |
| **PRODUCT CATEGORY-1** | YES | YES | YES | YES | YES | YES | YES |
| **STAY** | YES | YES | YES | YES | YES | YES | YES |

A chi-square analysis was performed to determine whether each Category was represented across all the groups proportionally to their numbers in the sample. The analysis produced a signifcant χ2 value, indicating that groups were overrepresented in any of the categories.

#### ANOVA

##### GENDER

We performed a one-way ANOVA to compare the Two group’s average Purchase on the eve of Black Friday. This analysis produced a statistically significant result (*F(1,9998)* = 47.34 , *p* < .05). Post hoc Tukey test revealed that the only significant difference between the groups was found between Male(M = 9504.77) and Female(M = 8809.76), with the Male spending more on Purchase significantly more than the Females.

##### CITY CATEGORY

We performed a one-way ANOVA to compare the Three group’s average Purchase on the eve of Black Friday. This analysis produced a statistically significant result (*F(2,9997)* =37.26 , *p* < .05). Post hoc Tukey test revealed that significant difference between the groups was found between City A(M = 8958.01), City B(M =9198.65), and City C(M = 9844.44 )with the City C Purchasing significantly more than City A and City B.

##### MARITAL STATUS

We performed a one-way ANOVA to compare the Two group’s average Purchase on the eve of Black Friday. This analysis produced no statistically significant result( *F(1,9998)* = 0.005, *ns*).

# 3. Feature Engineering

Feature Engineering played an important role in making the model more accurate, robust and stable for Modeling. We tried several novel Feature Engineering techniques that led to an increase in the model performance.

## 3.1 Data Conversion

We made the following conversion in the Predictors:

1. **‘User\_ID’**: Used as Raw Feature.
2. **‘Product\_ID’**: Used as Raw Feature
3. **‘Gender’**: Converted to Binary.
4. **‘Age’**: Converted to Numeric.
5. **‘Marital\_Status’**: Converted to Binary.
6. **‘Occupation’**: Used as Raw Feature
7. **‘City\_Category’**: One-Hot Encoded.
8. **‘Stay\_In\_Current\_City’**: Converted to Numeric.
9. **‘Product\_Category\_1’**: Used as Raw Feature.

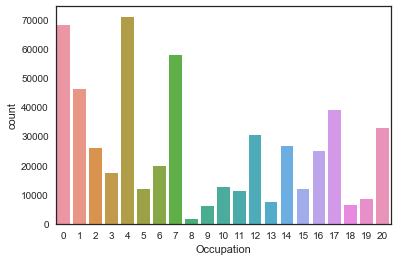
## 3.2 Data Binning

The Predictors ‘**Occupation**’ and ‘**Product\_Category\_1**’ had more than 3 categories. Since the dataset has half a million data-points, we reduced the categories for both the Predictors in order to reduce computational complexity. Also, this would help in making the model simpler to understand as there would be less variables at play.

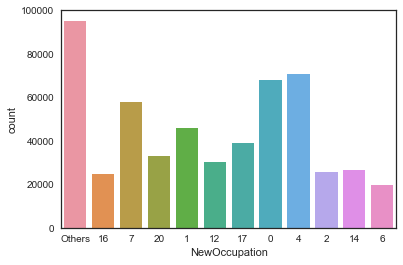
The strategy we used for binning relied on Parito’s rule i.e., categories that accounted for 80% of the data-points were kept while the other categories were placed under ‘**Other**’ category.

The results are as follows:

1. Occupation:
   1. Pre-binning categories = 21
   2. Post-binning categories = 12

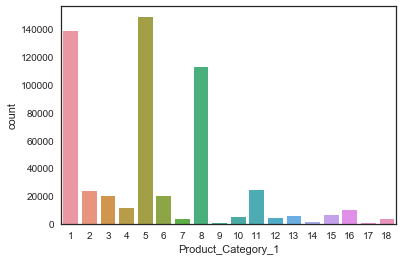


**Figure 7: Pre Binning Occupations**

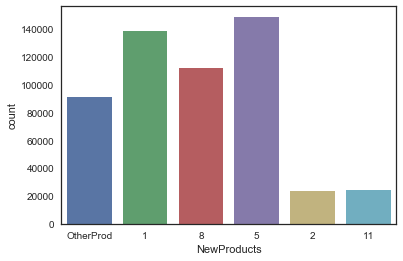


**Figure 8: Post Binning Occupation**

1. Product\_Category\_1:
   1. Pre-binning categories = 18
   2. Post-binning categories = 6



**Figure 9: Pre Binning Product Category 1**



**Figure 10: Post Binning Product Category 1**

## 3.3 Feature Creation

### 3.3.1 Based on Average Feature Purchase

We created features that would calculate the Average Purchase performed by a Predictor’s particular category. For Instance, for the Predictor ‘Gender’, this engineered feature would list the Average Purchase for ‘Gender’ = ‘Male’ as 9504.77

The following features were created:

1. **Avg\_User\_Purchase** : Average Purchase for 5891 different Users.
   1. **Avg\_ProductID\_Purchase**: Average Purchase based on 3623 different Product ID’s.
2. **Avg\_Gender\_Purchase**: Average Purchase based on both the Gender’s.

|  | **Avg\_Gender\_Purchase** |
| --- | --- |
| **Gender** |  |
| **F** | 8809.761349 |
| **M** | 9504.771713 |

1. **Avg\_Age\_Purchase**: Average Purchase based on different Age groups.
2. **Avg\_Occupation\_Purchase**: Average Purchase based on 21 different Occupation types.
3. **Avg\_City\_Category\_Purchase**: Average Purchase based on 3 different cities.

|  | **Avg\_City\_Category\_Purchase** |
| --- | --- |
| **City\_Category** |  |
| **A** | 8958.011014 |
| **B** | 9198.657848 |
| **C** | 9844.441855 |

1. **Avg\_Stay\_Years\_Purchase**: Average Purchase based on Years of Stay.

|  | **Avg\_Stay\_Years\_Purchase** |
| --- | --- |
| **Stay\_In\_Current\_City\_Years** |  |
| **0** | 9247.238625 |
| **1** | 9319.865095 |
| **2** | 9397.607316 |
| **3** | 9350.685121 |
| **4+** | 9346.370158 |

1. **Avg\_Marital\_Status\_Purchase**: Average Purchase based on Marital Status.

|  | **Avg\_Marital\_Status\_Purchase** |
| --- | --- |
| **Marital\_Status** |  |
| **Married** | 9334.632681 |
| **UnMarried** | 9333.325467 |

1. **Avg\_Product\_Category\_1\_Purchase**: Average Purchase based on 18 different Product Categories.

### 3.3.2 Based on Feature Frequency

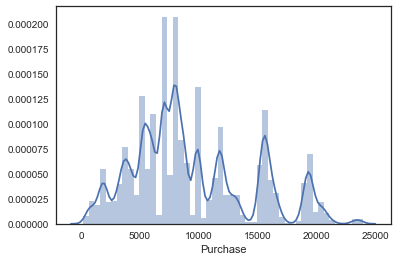
We created a feature that counts the frequency of each element for all the predictors. It was one of the most important features and helped scale the model efficiency.

## 3.4 Polychotomization

Certain Machine Learning Algorithms require predictors to be numeric. Thus, to use categorical predictors we need to recode categorical variables into one or more *flag variables*. A *flag variable* is also called as *Dummy Variable* and the process of dummification in Literature is called as polychotomization.

## 3.5 Response Transformation

In order to perform Regression Analysis, the Response variable should be normally distributed. In the current dataset, the response was almost normally distributed with some positive skewness.

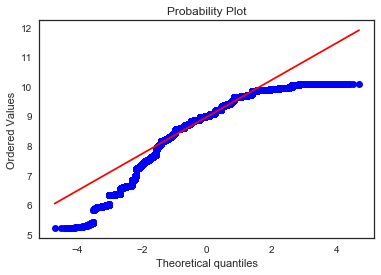


**Figure 11: Target Distribution**

In several models that we used, we transformed the Target variable in order to remove Heteroskedasticity.

In order to achieve Normality:

* We employed Log-Transformation of the Target variable.
* Observed skewness = -0.11
* The Q-Q Plot of Log-Transformed target variable indicated normality.



**Figure 12: Normal Probability Plot of log(Purchase) indicates normality.**

# 4. Model Development

## 4.1 Regression Model

The Machine Learning algorithm that is used is Multiple Linear Regression because Regression analysis helps one understand how the typical value of dependent variable changes when one of the independent variables is varied, while the other independent variables are held fixed.

Furthermore, Regression analysis helps in answering the below Business-related questions from a statistical perspective:

1. Is there a relationship between Purchase on Black Friday by a Customer and Predictor variables?
   * + We answer this question in the ‘Statistical Model’ section, where we analyse the *hypothesis test* on the coefficients.
2. How strong is the relationship?
   * + We shall be using the metrics RMSE to calculate the standard deviation of the response from the population regression line.
     + R2 and adjusted R2 to assess the percentage of variability in the Purchase that is explained by the Predictors.
3. Which Predictor contributes to the Purchase on the eve of Black Friday?
   * + From Statistical Model, we shall look at the p-values associated with each predictor’s t-statistic.
4. How large is the effect of each predictor on Purchase?
   * + We shall look at the value of each Predictor’s coefficient to assess the effect of each Predictor on Purchase individually.
5. How accurately can we predict the Purchase?
   * + The response can be predicted using the Regression Equation:

*Y* = *β*0 + *β*1*X*1 + *β*2*X*2 + *β*3*X*1*X*2 + ε

1. Is the relationship linear?
   * + We shall observe Residual Plots to check for non-linearity of the model.

### 4.1.1 Model Development

* + - The Dataset was divided into Train and Validation Set with Training set consisting of 70% of the Data.
    - Several Models were developed based on Feature Selection and Target Transformation.

### 4.1.2 Model Evaluation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Regression Models** | **Training Set** | | | **Validation Set** | | |
| **RMSE** | **R2** | **Adjusted R2** | **RMSE** | **R2** | **Adjusted R2** |
| **Baseline Model** | 4707.53 | 0.11 | 0.11 | 4715.49 | 0.11 | 0.11 |
| **Model 1** | 3888.17 | 0.39 | 0.39 | 3895.55 | 0.39 | 0.39 |
| **Model 2** | 4979.67 | 0 | 0 | 4984.44 | 0 | 0 |
| **Model 3** | 2903.5 | 0.66 | 0.66 | 2906.65 | 0.66 | 0.66 |
| **Model 4** | 4979.71 | 0 | 0 | 4984.36 | 0 | 0 |
| **Ridge Regression** | 2903.84 | 0.66 | 0.66 | 2906.96 | 0.66 | 0.66 |
| **LASSO Regression** | 2928.48 | 0.65 | 0.65 | 2930.12 | 0.66 | 0.66 |

**Figure 13: Results**

### 4.1.3 Model Optimization

#### 4.1.3.1 Cross-Validation

Cross-validation, sometimes called rotation estimation,or out-of-sample testing is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (called the validation dataset or testing set).The goal of cross-validation is to test the model’s ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).

In the current Regression analysis, we employed 10-fold cross-validation.

#### 4.1.3.2 Ridge Regularization

In Ridge Regression, the OLS loss function is augmented in such a way that we not only minimize the sum of squared residuals but also penalize the size of parameter estimates, in order to shrink them towards zero:

https://res.cloudinary.com/dyd911kmh/image/upload/f_auto,q_auto:best/v1543418449/eq7_ylxudw.png

Solving this for β^ *ridge* gives the ridge regression estimates β^*ridge* = (X′X+λI)−1(X′Y), where *I* denotes the identity matrix.

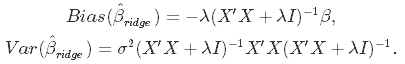
The *λ* parameter is the regularization penalty. We will talk about how to choose it in the next sections of this tutorial, but for now notice that:

* As λ→0,β^ridge→β^OLS;
* As λ→∞,β^ridge→0.

So, setting *λ* to 0 is the same as using the OLS, while the larger its value, the stronger is the coefficients' size penalized.

**Bias-Variance Trade-Off in Ridge Regression:**

Incorporating the regularization coefficient in the formulas for bias and variance gives us



From there we can see that **as λ becomes larger, the variance decreases, and the bias increases**.

#### 4.1.3.3 LASSO Regression

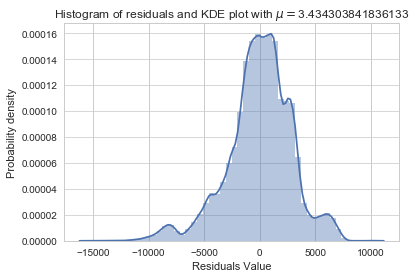
Lasso, or Least Absolute Shrinkage and Selection Operator, is quite similar conceptually to ridge regression. It also adds a penalty for non-zero coefficients, but unlike ridge regression which penalizes sum of squared coefficients (the so-called L2 penalty), lasso penalizes the sum of their absolute values (L1 penalty). As a result, for high values of λ, many coefficients are exactly zeroed under lasso, which is never the case in ridge regression.

**Model specification:**

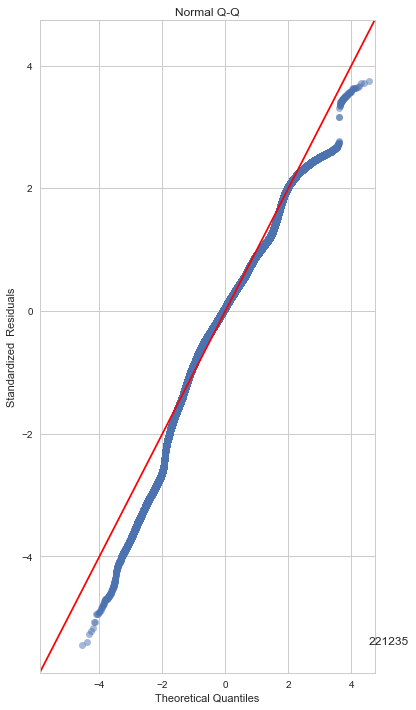
The only difference in ridge and lasso loss functions is in the penalty terms. Under lasso, the loss is defined as:

https://res.cloudinary.com/dyd911kmh/image/upload/f_auto,q_auto:best/v1543418448/eq11_ij4mms.png

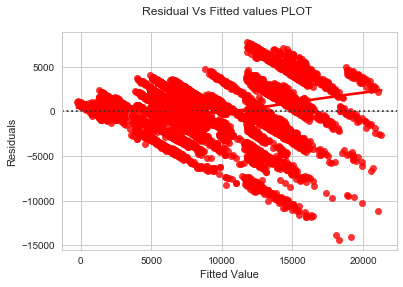
#### 4.1.3.4 Residual Plots



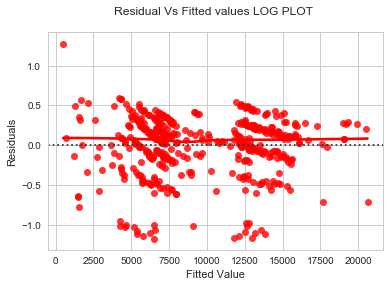
**Figure 14: Residual Distribution**



**Figure 15: Normal Probability Plot of Residuals**



**Figure 16: Visible Heteroskedasticity**



**Figure 17: Removed Heteroskedasticity**

**Observations:**

* + - It is evident from fig-14 and fig-15 that residuals do not follow a normal distribution suggesting that the model is unable to capture the pattern well.
    - From Fig-16 we can observe that non-linear pattern is not observed.
    - Fig-16 also suggest that residuals follow a funnel shape which is a strong indicator of heteroskedasticity.
    - In fig-17, after log transforming the Target, we can observe that heteroskedasticity is diminished.

## 4.2 Statistical Learning

|  |  |  |  |
| --- | --- | --- | --- |
| OLS Regression Results | | | |
| **Dep. Variable:** | Purchase | **R-squared:** | 0.653 |
| **Model:** | OLS | **Adj. R-squared:** | 0.653 |
| **Method:** | Least Squares | **F-statistic:** | 3.935e+04 |
| **Date:** | Sun, 06 Jan 2019 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 17:40:07 | **Log-Likelihood:** | -3.5381e+06 |
| **No. Observations:** | 376303 | **AIC:** | 7.076e+06 |
| **Df Residuals:** | 376284 | **BIC:** | 7.076e+06 |
| **Df Model:** | 18 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **const** | 1.134e+04 | 20.082 | 564.769 | 0.000 | 1.13e+04 | 1.14e+04 |
| **Product\_Category\_1\_10** | 6534.2674 | 50.434 | 129.561 | 0.000 | 6435.419 | 6633.116 |
| **Product\_Category\_1\_7** | 4267.2267 | 58.652 | 72.755 | 0.000 | 4152.270 | 4382.183 |
| **Product\_Category\_1\_6** | 2659.5272 | 26.192 | 101.541 | 0.000 | 2608.192 | 2710.862 |
| **Product\_Category\_1\_16** | 2026.6088 | 36.722 | 55.187 | 0.000 | 1954.634 | 2098.583 |
| **Product\_Category\_1\_15** | 2123.6187 | 45.426 | 46.749 | 0.000 | 2034.586 | 2212.651 |
| **City\_Category\_C** | 283.9126 | 10.471 | 27.114 | 0.000 | 263.389 | 304.436 |
| **Age** | 10.0330 | 0.359 | 27.939 | 0.000 | 9.329 | 10.737 |
| **Product\_ID\_Counts** | 2.5978 | 0.014 | 185.461 | 0.000 | 2.570 | 2.625 |
| **Stay\_In\_Current\_City\_Years** | 7.8901 | 3.708 | 2.128 | 0.033 | 0.622 | 15.158 |
| **Occupation\_1** | -162.6174 | 17.166 | -9.473 | 0.000 | -196.262 | -128.973 |
| **Product\_Category\_1\_3** | -2811.2377 | 26.454 | -106.270 | 0.000 | -2863.086 | -2759.389 |
| **Product\_Category\_1\_8** | -5218.7197 | 13.907 | -375.253 | 0.000 | -5245.977 | -5191.462 |
| **Product\_Category\_1\_18** | -9453.6809 | 64.223 | -147.202 | 0.000 | -9579.555 | -9327.806 |
| **Product\_Category\_1\_11** | -7742.6858 | 24.644 | -314.179 | 0.000 | -7790.988 | -7694.384 |
| **Product\_Category\_1\_5** | -6633.2756 | 12.698 | -522.406 | 0.000 | -6658.162 | -6608.389 |
| **Product\_Category\_1\_12** | -1.122e+04 | 56.755 | -197.758 | 0.000 | -1.13e+04 | -1.11e+04 |
| **Product\_Category\_1\_4** | -1.045e+04 | 33.805 | -309.155 | 0.000 | -1.05e+04 | -1.04e+04 |
| **Product\_Category\_1\_13** | -1.191e+04 | 48.513 | -245.426 | 0.000 | -1.2e+04 | -1.18e+04 |

**Inferences:**

* + - A Multiple Linear Regression analysis was conducted to examine the predictors of Purchase on the eve of Black Friday. 18 variables were simultaneously entered into the model. Together, these predictors accounted for 65.3% of the variance in the Purchase. All of these variables were significant predictors of Purchase. Product\_Category\_1\_10(β = 7234) and Product\_Category\_1\_13(β = -9136.8) were the strongest predictor of Purchase.

# 5. Recommendations:

## 5.1 Based on Descriptive Analytics:

* + - Male Shoppers are likely to spend more than Female Shoppers.
    - Older(40+) people are likely to spend more irrespective of their marital status.
    - Customers who arrived recently in City-B and City-C shop less frequently than those who stayed longer (Acclimatization can be an issue).

## 5.2 Based on Predictive Analytics:

* + - Purchase is heavily influenced by Product Category.
    - People of 60+ Age will spend as much as 600$ more than Teenagers.
    - People belonging to Occupation-1 are likely to spend less.
    - Product Category that have an average price over 9000$ are likely to influence more on Purchase.
    - City C Customers will spend 283$ more than other city Customers.
    - Strangely, Product Category-8 which is the third most selling Product contributes negatively on an Individuals Purchase.

## Based on Prescriptive Analytics:

* + - If the Price of **‘Product-5’** is increased by **5%**, **‘Product-1’** by **3%** and **‘Product-8’** by **4%** then the Revenue will increase by 150 M$ which is higher than the combined Revenue of eight lowest selling Products.

# 6 Conclusions:

* + - As revealed in the EDA, Product Categories influence Purchase the most. This inference is backed by the Regression Model.

# 7 References

Shay, M. (2015, December 16), “Black Friday: Changing, but Far from Gone,”  
National Retail Federation, available at: <https://nrf.com/news/black-friday-changing-far-gone>

Thomas, J. B. and Peters, C. (2011), “An Exploratory Investigation of Black Friday  
Consumption Rituals,” International Journal of Retail & Distribution Management, 39, 522-  
537.

Wahba, P. (2016, November 27), “About 10 Million More Americans Shopped  
Online than in Stores over Black Friday Weekend,” Fortune, available at:  
<http://fortune.com/2016/12/12/swift-new-cyber-thefts-hacking/>

Gallit Shmueli, Nitin Patel, and Peter Bruce, Data Mining for Business Intelligence, 2nd edition, John  
Wiley and Sons, 2010

Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A.(2006). Feature Extraction Foundations and Applications, Springer-Verlag Berlin Heidelberg.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), The elements of statistical learning: data mining, inference and prediction, 2 edn, Springer. URL: http://www-stat.stanford.edu/ tibs/ElemStatLearn/

G. James et al., An Introduction to Statistical Learning: with Applications in R,  
Springer Texts in Statistics, DOI 10.1007/978-1-4614-7138-7 3,  
© Springer Science+Business Media New York 2013

# 8 APPENDIX

## 8.1 Python Code For Statistical Analysis

<https://github.com/S-B-Iqbal/Capstone-Project/tree/master/Python%20Codes>

## 8.2 Python Code For Feature Engineering

<https://github.com/S-B-Iqbal/Capstone-Project/tree/master/Python%20Codes>

## 8.3 Python Code For The Models

<https://github.com/S-B-Iqbal/Capstone-Project/tree/master/Python%20Codes>