*Back In Black*

A Thumbayil, S Bansal and S B Iqbal

PGP-DSE Chennai Batch – July ’18, GreatLearning

**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.**

1. 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)

* 1. 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 Logistics of the Retail Store.
* Increase Revenue and Sales.

1. Data
   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’.
   1. Data Description

TABLE I. DATA VARIABLES

|  |  |  |  |
| --- | --- | --- | --- |
| SI NO. | VARIABLES | TYPE | REMARKS |
| 1. | User\_ID | Integer | To be converted as Categorical Variable since it is an ID Variable |
| 2. | Product\_ID | String | Product ID |
| 3. | Gender | String | Sex of User |
| 4. | Age | Categorical | Age in Bins |
| 5. | Occupation | String | Occupation listed as numbers |
| 6. | City\_Category | String | Category of the City(A,B,C) |
| 7. | Stay\_In\_Current\_City\_Years | Categorical | Number of Years Stay in Current City. |
| 8. | Marital\_Status | Categorical | 0 – UnMarried, 1- Married |
| 9 | Product\_Category\_1 | Categorcial | Product Category |
| 10. | Product\_Category\_2 | Categorcial | Product may belongs to other category also |
| 11. | Product\_Category\_3 | Categorcial | Product may belongs to other category also |
| 12. | Purchase | Integer | Purchase Amount in Dollars |

* 1. Data Preprocessing

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  
preprocessing, in the form of *data cleaning* and *data transformation*.

* + 1. 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.

* + 1. Handling Outliers
  1. Exploratory Data Analysis
     1. Univariate Analysis
     2. Bivariate Analysis
     3. Statistical Analysis

1. Model
   1. Feature Engineering
      1. Data Bucketing
      2. Dichotomization
      3. Target Transformation
   2. Model Development
      1. Regression Model
      2. Model Optimization
         1. Ridge Regularization
         2. LASSO Regularization
         3. Elastic-Net Regularization
2. Analysis
   1. Metrics used for Measuring ML Models.
   2. Bootstrapping Technique for estimating Test-Set Error Rate

8. Acknowledgement

9. 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