Black Friday Sales Analysis

IST 565 Data Mining

Team Members: Rebecca Karunakaran

**Introduction**

**Black Friday** is an informal name for the Friday following [Thanksgiving Day in the United States](https://en.wikipedia.org/wiki/Thanksgiving_(United_States)), which is celebrated on the fourth Thursday of November and is the beginning of America's [Christmas shopping season](https://en.wikipedia.org/wiki/Economics_of_Christmas) since 1952. Many stores offer highly promoted sales on Black Friday and open very early, such as at midnight, or may even start their sales at some time on Thanksgiving.

I have chosen the “ Black Friday Sales Data “ from Kaggle ([https://www.kaggle.com/mehdidag/black-friday#BlackFriday.csv](https://www.kaggle.com/mehdidag/black-friday)) for the data mining project because it is a sample retail transactions containing a variety of data ( discrete and continuous) and offers an opportunity to apply various data mining algorithms to analyze the data and to be able to train the model to predict future Black Friday purchases.

**Dataset and Preprocessing**

The data set is fairly large with 12 variable and about 530K observations containing the purchase pattern for buyers at a retail store - what products were bought and how much was spent. What is not specified is the cost of the product and quantity purchased.

**Data Dictionary**

|  |  |  |
| --- | --- | --- |
| **Column** | **Description** | **Sample Data** |
| User\_ID | Unique number for User | 1000001 |
| Product\_ID | Unique number for Product | P00000142 |
| Gender | Male or Female | M' or 'F' |
| Age | Age Bracket | 0-17","18-25 |
| Occupation | Occupation Category Number | A, B, C |
| City\_Category | City Category | 55.5 |
| Stay\_In\_Current\_City\_years | Years lived in the City | 0,1,2,3,4+ |
| Marital\_Status | 0-Not Married, 1-Married | 0 or 1 |
| Product\_Category\_1- Clothes | Category ID | 1 to 18 |
| Product\_Category\_2- Electronics | Category ID | 2 to 18 |
| Product\_Category\_3- Home Goods | Category ID | 3 to 18 |
| Purchase | Dollar Amount spend |  |

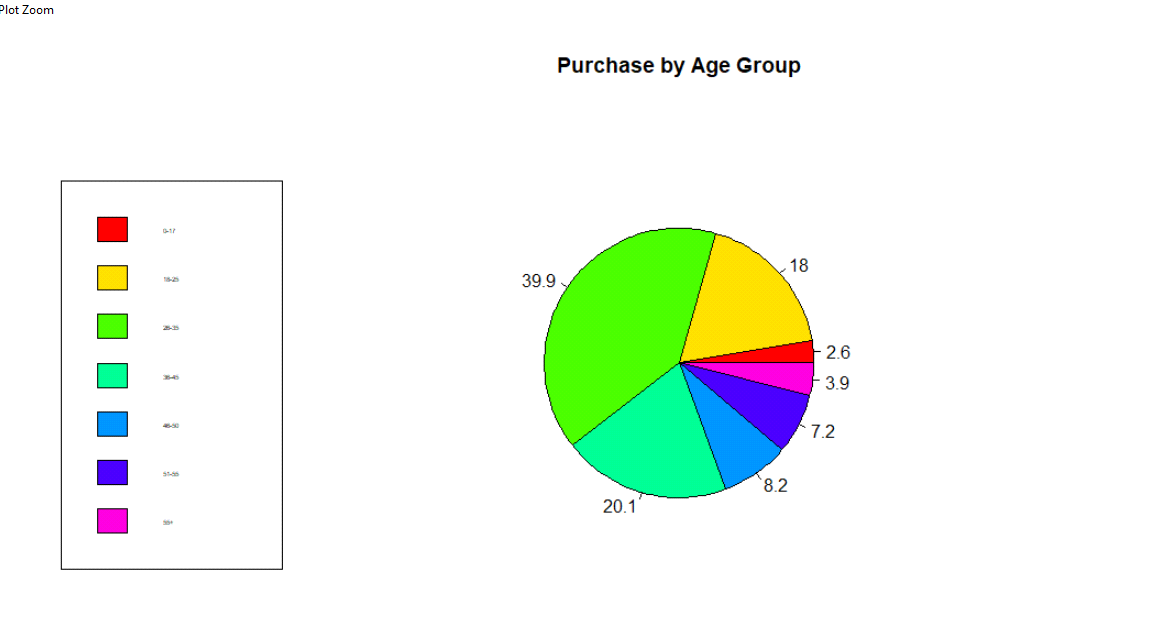
Using the head() and str() function, I was able to review the data and check for duplicates, and fortunately it is a very clean dataset.

**Descriptive Analysis**

Visualizing data can give us interesting nuggets –

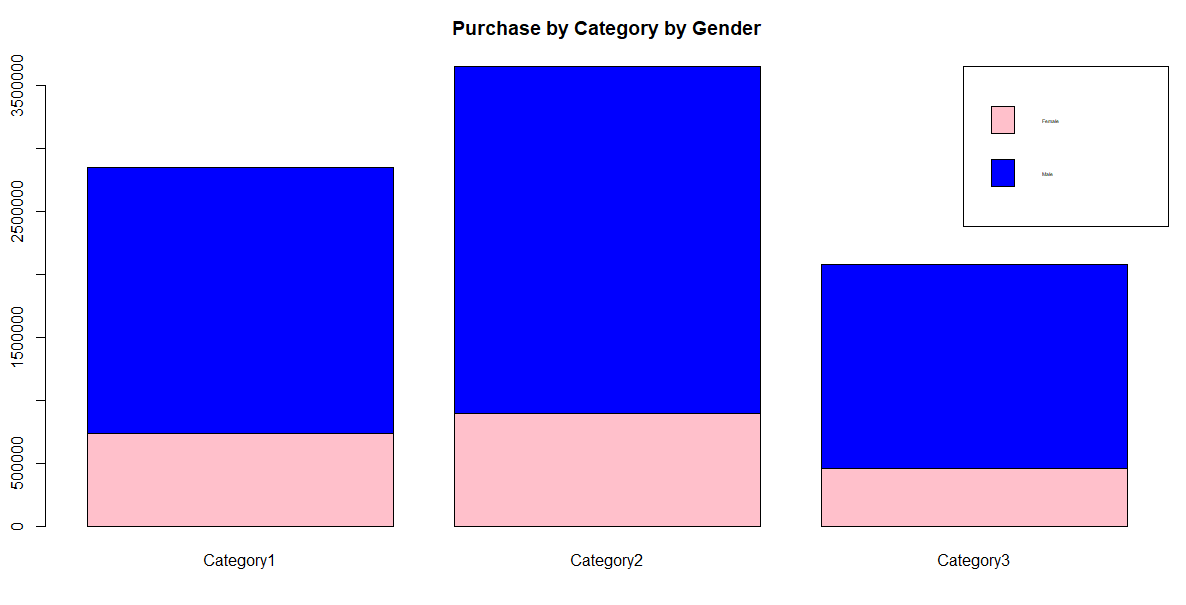
**Purchase habits based on the Age Bracket:** We can see that the majority of the buyer were in the Age Bracket 36-45 followed by people within the Age Bracket 26-35. This makes sense because these are age groups that employed and have spending power.

**Age**



**Purchase habits based on Gender:** The bar chart analysis for Gender also shows spending habits on Men versus Women. Male buyers were clearly the biggest spender which is a little surprising because people generally think that Women are big on shopping! This may be due to Online shopping where Men tend to dominate compared to Women. Also, the Category 2 which is “Electronics” had the greatest number of buyers which is not a surprise since the majority of buyers were Men.

**Gender**



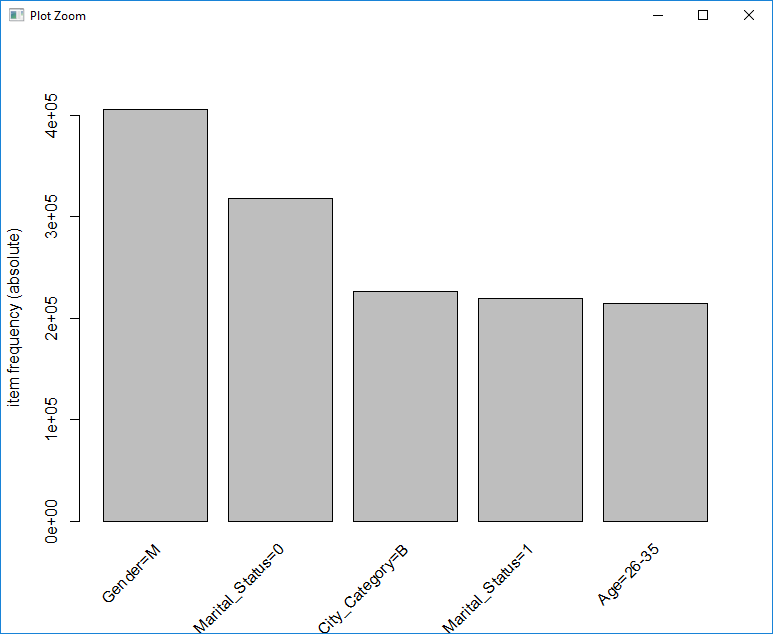
**Association Rules Mining**

Association Mining searches for frequent items in the data-set and will be an excellent algorithm to study the frequent item sets along with the association rule. Using the Apriori Algorithm, one can get some meaningful insights into the association rules for purchase categories.

**Preprocessing**

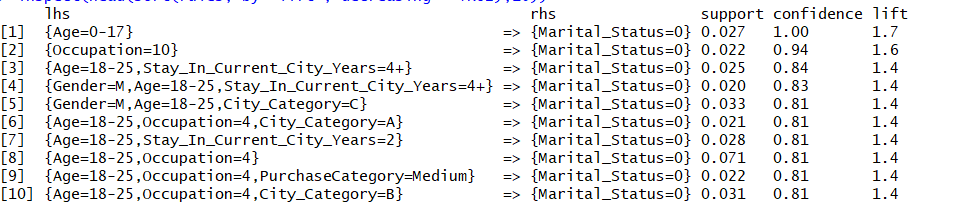
Since Association Rules work on discrete data, I had to first convert the numeric data to nominal data. I also created a nominal variable “Purchase Category” that will categorize the purchase as "Small","Medium","Large","Very Large". The dataset was then converted into transactional data in order to run the Apriori algorithm

The following chart helps us inspect the result and visualize the most common itemset in the transaction. Surprisingly we see that the majority buyers were Male (Gender = M) and Unmarried (Maritial Status = 0). I would have expected the majority of the buyer to the Female!



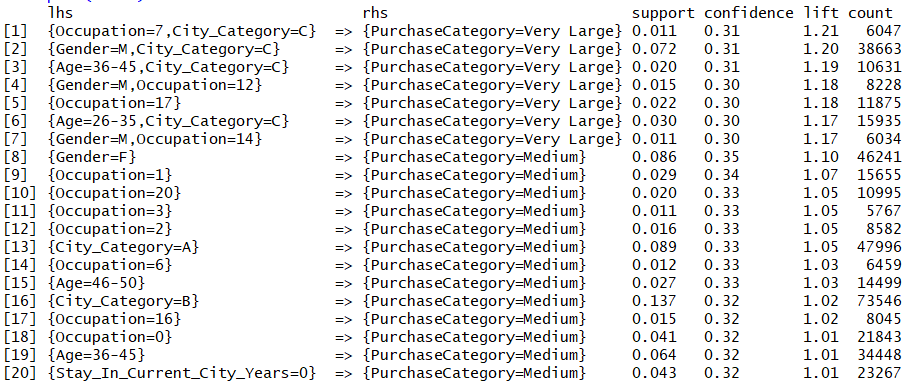
**Analysis**

With a min supp=.02 and conf=0.8, I got about 81 Rules. Here is a sample of the rules, sorted with Confidence is descending order –



We notice some interesting rules like people in the Age Bracket of 0-17 have a high probability of being not married since the confidence is 1.0. Similarly, people with Occupation 4 in the Age Bracket of 18-25 have a higher probability of being not married based on the high confidence and lift.

But my main goal is to use Association Rule Mining as Supervised Learning Method by forcing the target calcification attribute “Purchase Category” on the RHS of association rules to find out which attributes of the dataset contribute to a given Purchase Category. After removing redundant rules, I got 35 rules. Here is a snap shot of the rules –



A few insights from these rules-

* Rules 2, 4 and 7 indicate that Male shopper have a high probability of making Large Purchases ($25000 - $12000) whereas Rule 8 indicates that Female buyers tend to make medium purchases ($5000 - $ 8000).
* Rules 1,2,3 and 6 show that buyers that live in City Category “C” tend to make Large Purchases indicating that it may be an affluent City where as Rule 13 tells us that buyers from City Category “A” are more likely to make medium purchases.
* Rules 3 and 6 indicate that buyer in the Age Group of 26-35 and 36-45 are more likely to make Large purchases where as buyers in the Age Group 45-50 (Rule 15) are more likely to make Medium Purchases.

**Classification**

Retailer are interested in analyzing their buyer to see who they should target for future marketing and advertising. I am going to use various classification algorithm to predict the Gender of the buyer.

I am splitting 80% of the data for the training and the remaining 20% for testing.

**Random Forest**

My first choice is the Random Forest Algorithm which is based on the Ensembling technique to improve the predictive performance of Decision Tress. I will be using 2 parameters **Ntree** which is the number of trees to grow and **Mtry** which is the number of variables randomly sampled as candidates at each split. With Default parameters, I see that the model has used Ntree=500 and Mtry=2 with OOB (Out of bounds) error rate is 17.61% and Accuracy of **82.89%.** In order to increase the accuracy, I decided to find the optimal Mtry parameter value using the tuneRF () function and see that the OOB error was lowest when Mtry=4 and Ntree=600. The tuned model has an Accuracy of **85.39%** which is an improvement from the Accuracy of the default model.

**Naïve Bayes**

Naïve Bayes algorithm works well on categorical data. So, after discretization of the data, I used the klaR algorithm in the caret function to train and test the model. With default parameters, the model gave me an accuracy of **75.24%** with fL=0, usekernel = TRUE and adjust =1.

I decided to tune to model to increase the accuracy and used 3 parameters

* **usekernel**: parameter allows us to use a kernel density estimate for continuous variables versus a gaussian density estimate,
* **adjust:** parameter that allows us to adjust the bandwidth of the kernel density (larger numbers mean more flexible density estimate)
* **fL**: parameter that allows us to incorporate the Laplace smoother

I trained 4 models, the first mode did not use the kernel density estimate i.e. usekernel=FALSE while keeping fL=0 and adjust=1 and the accuracy dipped to 68.30%. The second model was trained using the kernel density estimate – usekernel=TRUE, but changing to include a Laplace smoother of 1 while keeping adjust=1 and the Accuracy was 75.24% which was what the default model had predicted. Since the Laplace smoother did not make much of a difference in the accuracy. I trained a third model with usekernel=TRUE, fL=1 and increased the adjust to 3 and the Accuracy remained at 75.24 and a fourth with the same values for usekernel =TRUE and fL=1 but increased the adjust to 5 which resulted in a slight increase in Accuracy of **75.25%.**

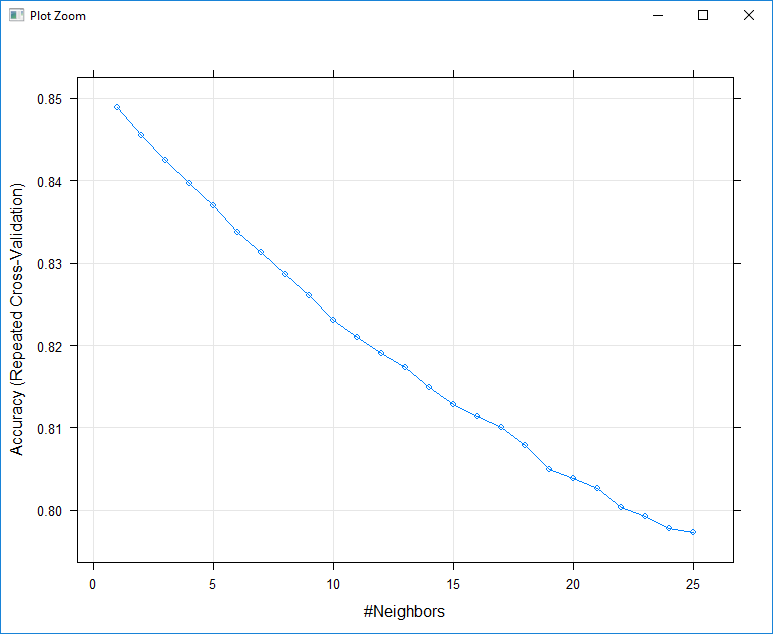
**KNN**

KNN algorithm is one of the simplest classification algorithms and it is one of the most used learning algorithms The KNN algorithm assumes that similar things exist in close proximity. It is instance based which means it simple stores the training example without doing any calculations during the training phase and the classification and prediction are delayed until new classification examples are given by compare the similarities between the test and training examples and choose the majority voted category labels in the k nearest training example.

**Preprocessing**

Since knn is a distance-based algorithm it was important to convert the relevant attributes to numeric and normalize the values. Using the Caret package, I trained the model with default parameters and saw that the model had an accuracy of **83.52%** choosing k to be 5.

I used the **k** parameter which is the number of neighbors "voting" on the test examples to tune the model. Setting k=1 to 25 using 5 repeats of 10-fold cross validation and saw that the accuracy was highest when k=1 as seen on the graph below.



The tuned model on the test data had an accuracy of **84.55%** which is an improvement from the default model

**SVM**

A Support Vector Machine (SVM) is a discriminative classifier defined by a separating hyperplane. In other words, given labeled training data the algorithm outputs an optimal hyperplane which categorizes new examples.

**Preprocessing**

SVM constructs a hyperplane such that it has the largest distance to the nearest data points (called support vectors). If the dimensions have different ranges, the dimension with much bigger range of values influences the distance more than other dimensions. So, it is necessary to scale the features such that all the features have similar influence when calculating the distance to construct a hyperplane.

I test 2 SVM Kernels- Linear and Radial.

**Linear SVM**

I trained the data to use SVM Linear model with default parameter and saw that it used Cost C=1 and had an accuracy of **75.71%**

## Since the C (Cost) which used for regularization, penalty associated with misclassification is the most important parameter for Linear SVM, I used the caret package to train the model with C (0.001, 0.01, 0.1) using bootstrapping with resample vector of 5. The tuned model gave me an accuracy of **75.70%** with C=.001 which is the same accuracy of the default model. Tuning C did not improve the accuracy.

**Radial SVM**

Next, I checked SVM with Radial Kernel and tuned 2 parameters - **Cost** C (0.001, 0.01, 0.1) as well as **Sigma** (0.1,0.2,0.3) which controls the shape of the "peaks" where you raise the points and the model gave an Accuracy of **75.13%.**

**Summary**

Nothing is more important to a retailer than really knowing their customers. With data mining algorithms they can learn exactly who their best customers are, what pushes them to shop, how frequently they buy, how much they spend per order, and more.

Since the dataset had transactional information regarding purchases at a retail store, the Association Rules Mining proved to be very value data mining tool since it could identify some important rules such as the population of buyer that make Large or Medium or Very Large purchases. For a retail store, this analysis is very valuable in determining who are the big item buyers, what age bracket do they belong and where do they live. This will help with better marketing and advertising strategies to improve future black Friday sales

|  |  |  |
| --- | --- | --- |
| **Model** | **Parameter** | **Accuracy** |
| **Random Forest** | Mtry, Ntree | 85.39% |
| **Naïve Bayes** | Usekernel, Adjust, fL | 75.25% |
| **knn** | k | 84.55% |
| **Linear SVM** | C | 75.70% |
| **Radial SVM** | C, Sigma | 75.13% |

From the summary above, we see that that the Accuracy to predict the Gender of the buyer was the highest for Random Forest which is what I had expected to see because Random Forest uses Ensembling technique which improves the predictive performance. Knn had the next highest Accuracy which is not unexpected since knn tests the similarities between the test and training examples and chooses the majority voted category labels in the k nearest training example. The Accuracy with the other 3 models – Naïve Bayes, Linear SVM and Radial SVM were very close, about 75%.

In term of the execution speed Radom Forest was the fastest and SVM was the slowest which may be due to the fact that Random Forest performs well with categorical data where the data is not sparse.

In Summary, the data and its feature like high-dimensionality, sparsity, multi-view, multi-label define which algorithm will be suitable. Classifier that work well with a kind of data may perform poorly with another kind of data.

.