BLACK FRIDAY SALES PREDICTION

Akashdeep Balu, Bharath Janapareddi, Kavya Nagaraju, Priyanka Patil, Tejaswi Gundapaneni

ABSTRACT

The purpose of the present study involves creating a prediction model for the user to estimate the customer purchase during one of the busiest days, Black Friday, the day after Thanksgiving, and one of the largest shopping days in the USA. To aid retailers, we are making use of Regression models such as Multiple Linear Regression, Lasso regression, Ridge Regression, Decision Tree to predict customer purchase. R-Square and RMSE values have been used to evaluate the model performance. We have also built a "Product Recommender Model" which recommends the products to users based on the purchase habits of similar users.

Key words. Linear Regression, Ridge Regression, Lasso Regression, Decision tree, recommenderlab

1. Background Information

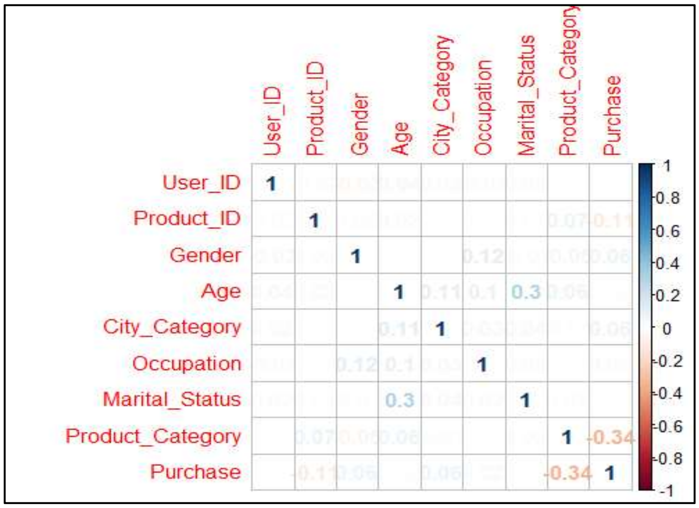
Black Friday, the day after Thanksgiving, is a term used by the retail industry in the United States that signifies the start of the Christmas holiday shopping season. This day usually attracts large number of customers and companies increase production rate based on customer buying patterns. Knowing these patterns help companies mitigate production of unnecessary items. This also helps companies to focus on needs of the target market. Finding the needs of target market depends on various factors including but not limited to age, gender etc., Breaking down each demographic further helps to determine their needs, requirements and preferences which are obtained from previous purchase patterns. The significance is to recommend the company to focus on production of products that are in high demand during the time rather than exhausting the financial resources on products that don't even make up the break-even point of sales. The Black Friday Dataset has taken from Kaggle and now has 55,0068 records and 12 attributes namely User_ID, Product_ID, Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, Marital status, Product_Category_1, Product_Category_2, Product_Category_3 and Purchase. As part of Dataset Cleaning and Pre-Processing, two variables namely 'Product_Category_2' and 'Product_Category_3' has been excluded as it had missing values and had no effect on the other variables. The Black Friday clean Dataset has 55,0068 records and 10 attributes. Below are the attributes in Data:

Variable Name	Data Type	Definition
User_ID	Integer	User Id
Product_ID	Factor	Product Id
Gender (M/F)	Factor	Sex of user
Age (7 levels)	Factor	Age
Occupation (0-20)	Factor	Occupation (Masked)
City_Category (A/B/C)	Factor	Category of the city (A, B, C)
Stay_In_Current_City_Years (5 levels)	Factor	Number of years of stay in the current city
Marital_Status (0/1)	Factor	Marital Status
Product_Category (1-20)	Factor	Product Category (Masked)
Purchase	Integer	Purchase Amount for User (Target Variable)

1.1 Exploratory Data Analysis (EDA)

As a part of EDA, the retail store has 5,891 shoppers making purchases. 3,631 products sold in the store. total revenue of $5,095,812,742. 72% of shoppers are men and the remaining 28% of shoppers are women. The skewness of the gender of shoppers might lead to a larger total revenue for male than that of female. **Correlation matrix** between the variables in the dataset is shown in the Figure 1.

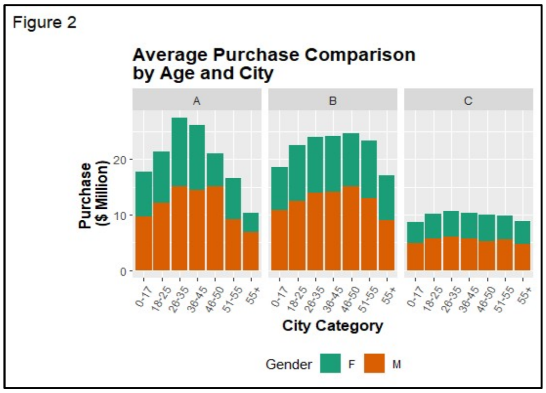
Figure 1



Gender	Number of Distinct Shoppers
F	1666
M	4225

Number of distinct shoppers based on gender is shown above. Analyzing the average unit price of each product category, the maximum average price is $19,675.57 for products in Category 10 while the minimum price is $722.40 for category 13. Category 5 sells the most products, 150,933 items, while category 9 sells the least number of products, 410 items. By investigating the different product categories at the retail store and the revenue per shopper in each product category, on average, product categories 1, 5, and 8 are the most popular product categories amongst shoppers. **A hypothesis being done on data set: The average shopper purchase of people living in City A is more than those living in cities B and C.**

Figure 2



The Figure 2 depicts purchasing habits/revenue by age and gender, male shoppers buy more than their female counterparts in all age groups and all cities. The data portrays a trend of purchasing habits - men purchase significantly more than women in age groups 18-25, 26-35 and 36-45. Across age groups, shoppers in City C purchase the least regardless of age. In response to the above-mentioned Hypothesis, the variables Shopper Purchase, City_Category, Gender, and Age are selected to conduct a two-sample t-test. Based on average purchase per shopper, the hypothesis is People living in City A spend more than their peers in City B and C. The results yield that shoppers in city A did not spend significantly more than their peers in city B but did spend more than peers in city C. Shoppers in city B spent significantly more than their peers in city C. Therefore, the shopping habits of customers in city A are significantly different from the remaining cities. The results also indicate shoppers in city C spend significantly less than cities A and B.

2. Methodology

Backward selection is a process which begins with the full least squares model containing all p predictors, and then iteratively removes predictors one at a time. It continues until the stopping rule is reached. This process is used to identify the actual predictors effecting the customer purchase in this analysis. Multiple predictors often cause more problem in regression due to overfit, on the other hand, including less predictors causes underfit. Instead of specifying which variable should be used, we let the algorithm do the job. We identified age, gender, city category, occupation, marital status and product category are the predictors plays an important role in predicting the customer purchase. Dataset is divided into a 7:3 (70 as training and 30 as testing). Below are the regression models to achieve our goals. Our response is purchase amount. Predictors are age, gender, city category, occupation, marital status and product category.

2.1 Multiple Linear Regression

Linear Regression is an analysis that explains the variance in the response variable. The Predictors mentioned above have no or little multi-collinearity between them. Knowing the relationship between the independent and dependent is linear, then this algorithm is the best because it's the least complex when compared to other algorithms. But in the model, R-Squared value is 0.6412231 on the test data, so we move to Ridge, Lasso and Decision Tree algorithms.

2.2 Ridge and Lasso Regression

When alpha=0 then a Ridge regression model is fit, and when alpha=1 then a Lasso model is fit. The Grid was defined to choose the wide range of values ranging from λ = 10^10 to λ = 10^−2, essentially covering the full range of scenarios from the null model containing only the intercept, to the least-squares fit. The ridge model was fit using the glmnet function. Further, by performing the cross-validation on the model using the cv.glmnet function we found the lamda for which CV error is minimum on training data and that value was taken as 's' to predict the purchase amount on the test matrix. Clearly, we can see that there is an improvement in both RMSE and R-Square values on test data which means lasso model now is able to predict much closer values to actual values.

2.3 Decision Tree

Decision Tree solves the problem by transforming the data into tree representation. It can be used to solve both regression and classification problems (CARTs). At each step in the algorithm, a decision tree node is split into two or more branches until it reaches leaf nodes and it chooses a feature that best splits the data with the help of two functions: Gini Impurity and Information Gain. Pruning (using rpart package) is done in the model to reduce the chances of overfitting the tree to the training data and reduce the overall complexity of the tree. It performs greedy search of best splits at each node. This is particularly true for CART based implementation which tests all possible splits. For a continuous variable, this represents 2^(n-1) - 1 possible splits with n the number of observations in current node. It can have poor prediction accuracy for responses with low sample sizes.

2.4 Regression Model Results

R-Squared value is 1% when product category is excluded and could see huge transformation in R-Squared value when product category is included. Product category is the most important predictor among all the predictors. There is a slight improvement in the performance of regression models from Linear to Decision Tree models. We have found the Black Friday dataset performs the best on Lasso Regression with an accuracy of 0.6412237.

Regression Models	**TRAIN**		**TEST**	
	RMSE	**R-square**	**RMSE**	**R-square**
Decision Tree	3081.557	0.6237949	3073.325	0.625291
Linear	3016.13	0.6396003	3007.286	0.6412231
Ridge	3016.13	0.6396003	3007.282	0.641235
Lasso	3016.131	0.6396003	3007.282	**0.6412237**

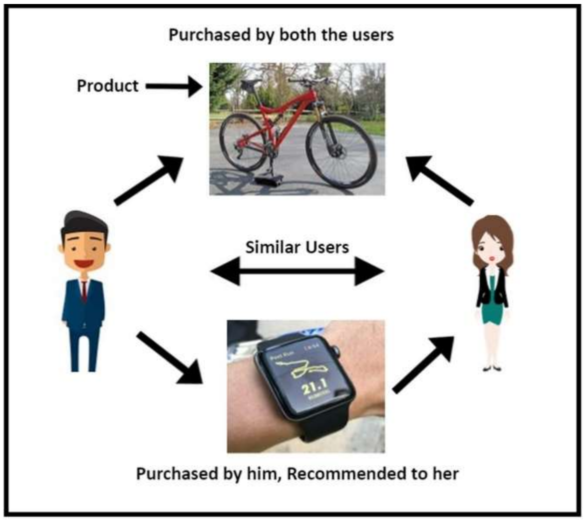
2.5 Product Recommendation

Today recommender systems are an accepted technology used by market leaders. Recommender systems apply statistical and knowledge discovery techniques to the problem of making product recommendations based on previously recorded data. Such recommendations improve the conversion rate by helping the customer to find products he/she wants to buy faster, promote cross-selling by suggesting additional products and improve customer loyalty through creating a value-added relationship. Recommender systems are categorized basically into two categories: content-based approaches and collaborative filtering. In this project, we are focusing on similar users and hence collaborative filtering of recommender algorithms. The R extension package "*recommenderlab*" provides a general research infrastructure for recommender systems.

2.5.1 Collaborative filtering Approach

Collaborative filtering (CF) uses given rating data by many users for many items as the basis for predicting missing ratings or for creating a top-N recommendation list for a given user, called the active user. The set of users U = {u1, u2,.., um} and a set of Products P = {p1, p2,.., pn} are considered. Ratings are stored in a m × n user-item rating matrix coc_mat = Cmxn where each row represents a user u[j] with 1 ≤ j ≤ m and columns represent products p[k] with 1≤ k ≤ n. Cmxn represents the rating of user uj for product Pj . Rating scale is "1", if user would buy a product, else "0". The model used in the process is Recommender. Recommender model has many methods named IBCF, LIMBF, RANDOM, POPULAR and UBCF. Our basic idea for recommending product is based on similarity between users, whose purchasing habits are more like object user, the products which are not yet purchased by the object user are recommended as referred in Figure 3.

Figure 3

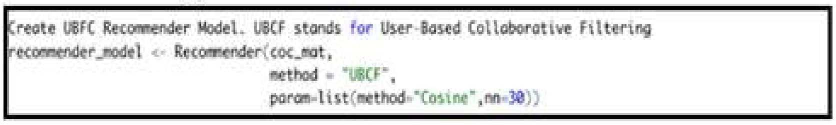

Purchased by both the users
Product
Similar Users
Purchased by him, Recommended to her

To carry out this process, a subset of original dataset is chosen because it is very difficult to obtain input matrix i.e., coc_mat from the whole data. As data is very large to compute coc_mat, the process used to meet this requirement consumed more than 4 hours and still no result is produced. So, a subset of data containing 500 unique users and 2972 different products purchased by these users. We trained our recommender model using User Based Collaborative Filtering (UBCF) method and cosine method to compute similarity is

$$\mathrm{sim}_{\mathrm{Cosine}}(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|\|\vec{y}\|},$$

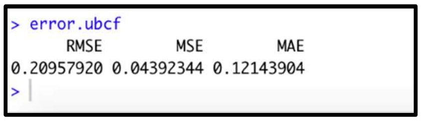
Now that our model is trained with the data and we can do recommendations for user(s). We tested it for a User_id "1000001" and the results are shown in Figure 5. There are the top 10 products recommended by

Figure 4



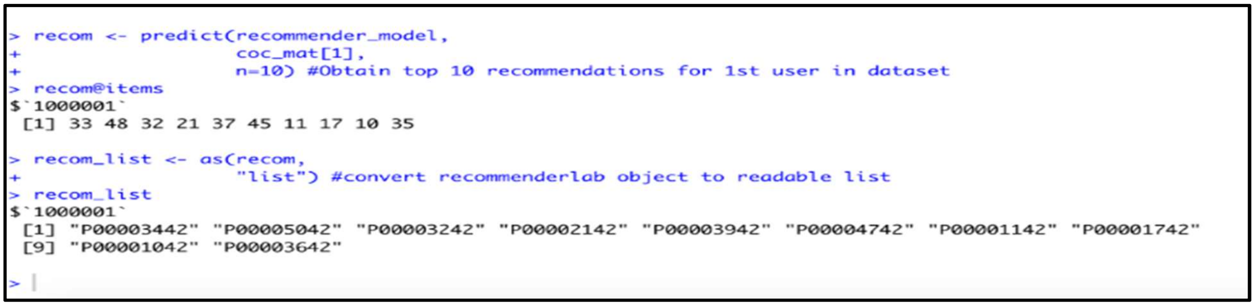
```
Create UBFC Recommender Model. UBCF stands for User-Based Collaborative Filtering
recommender_model <- Recommender(coc_mat,
                                 method = "UBCF",
                                 param=list(method="Cosine",nn=30))
```

Figure 6



```
> error.ubcf
      RMSE        MSE        MAE
0.20957920 0.04392344 0.12143904
> 
```

the model to the user. Further, the Evaluation was done by creating an evaluation scheme that determines what and how data is used for training and testing. Here we created an evaluation scheme which splits the 500 users in ration of 7:3 as training and testing. For the test set some items will be given to the recommender algorithm and the other items will be held out for computing the error. Obtained results are shown in Figure 6.



```
> recom <- predict(recommender_model,
+                  coc_mat[1],
+                  n=10) #Obtain top 10 recommendations for 1st user in dataset
> recom@items
$`1000001`
 [1] 33 48 32 21 37 45 11 17 10 35

> recom_list <- as(recom,
+                  "list") #convert recommenderlab object to readable list
> recom_list
$`1000001`
 [1] "P00003442" "P00005042" "P00003242" "P00002142" "P00003942" "P00004742" "P00001142" "P00001742"
 [9] "P00001042" "P00003642"

> |
```

Figure 5

3. Conclusion

Several Machine Learning Techniques like Multiple Linear Regression, Ridge, Lasso and Decision Trees were used. After having applied all the above models, it was observed that Lasso regression worked better than others in terms of prediction efficiency and accuracy. Significant analysis was made with the help of Root Mean Squared Error (RMSE) and R2 values. One more interesting insight is concluded that with city, gender and age, the sales rate or the number of purchases made varies. Highest purchases were made in city B, by males and by the people of the age group 26-35 respectively. We must also consider that all the variables which are there in the dataset are nit strongly co-related with the purchase variable. The Recommender model used to recommend products for a user and an evaluation scheme is also generated to evaluate the accuracy and efficiency of the model. The accuracy we acquired is creditable because the dataset was large with hidden values. Hence, from model selection and Product recommendations, with more understanding of the purchasing patterns, retailers can provide improved service quality.

4. References

[1] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0184516

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5628815/

[3] https://rdrr.io/r/utils/citation.html

[4] http://fs2.american.edu/alberto/www/analytics/ISLRLectures.html

[5] https://github.com/priyankamt/Black-Friday-Dataset