**PURCHASE PREDICTION AND PRODUCT RECOMMENDATION ON BLACK FRIDAY DATASET**

**Akashdeep BaluHC6943, Bharath JanapareddiHC4765, Kavya NagarajuHC3344,**

**Priyanka PatilHC6653, Tejaswi GundapaneniGZ5527**

**ABSTRACT**

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.

A hypothesis is done on the dataset: The average shopper purchase of people living in City A is more than those living in cities B and C.

Further, we predicted the “Purchase” amount for a user using various supervised machine learning techniques, and the variables were selected using backward selection. Multiple Linear Regression, Lasso Regression, Ridge Regression & Decision Tree were built to see which model works best in predicting the Purchase amount. R-square and RMSE values were used to evaluate the model performances.

Alongside this, we have also built a “Product Recommender Model” which recommends the products to users based on the purchase habits of similar users.

**BACKGROUND**

The Black Friday Dataset 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. For easy interpretation, Product\_Category\_1 was subsequently renamed Product\_Category.

The Black Friday clean Dataset has 55,0068 records and 10 attributes. Below are the attributes in Data:

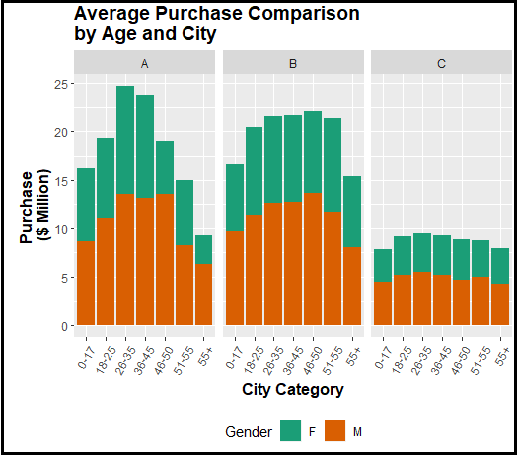
|  |  |  |
| --- | --- | --- |
| Variable Name | Data Type | Definition |
| User\_ID | Integer | User Id |
| Product\_ID | Factor | Product Id |
| Gender | Factor | Sex of user |
| Age | Factor | Age in binary |
| Occupation | Factor | Occupation (Masked) |
| City\_Category | Factor | Category of the city (A, B, C) |
| Stay\_In\_Current\_City\_Years | Factor | Number of years of stay in the current city |
| Marital\_Status | Factor | Marital Status |
| Product\_Category | Factor | Product Category (Masked) |
| Purchase | Integer | Purchase Amount for User(Target Variable) |

As part of **Exploratory Data Analysis,**

* 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.

**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.**

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.

 The graph depicts purchasing habits by age and gender, male shoppers buy more than their female counterparts in all age groups and all cities. Shoppers in City B purchase more than shoppers in cities A and C in all age groups except 26-35 and 36-45 years old. 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.

**METHODOLOGY**

We have performed a Backward selection process

Multiple predictors often cause more problems in regression due to overfitting, on the other hand, including fewer predictors causes underfit. Instead of specifying which variable should be used, we let the algorithm do the job.

As a result, the algorithm yielded the following variables: Age, Gender, City\_Category, Occupation, Marital\_Status & Product\_Category.

**Regression Models:**

We have divided the dataset into a 7:3 ratio. I.e., 70 as training and 30 as testing.

**Multiple Linear Regression:**

We have implemented the multiple linear models with the variables obtained from the above-mentioned backward selection.

**Accuracy from 1% to 64%**

We initially implemented the model with the predictors as Age, Gender, Marital status, Occupation, City category and got an accuracy of 1%

Further, we added the Product category to the above predictors, and we found that model performance was improved drastically I.e. from 0.01 to 0.64. We further, tried removing all the predictors except the Product category and found the result as 0.63. From the above trials, we found that the target variable Purchase is highly dependent on the Product category.

**Ridge and Lasso Regression:**

We first created a training matrix and testing matrix in the ratio of 7:3

*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 = 1010 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 lambda 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.

**Decision Tree:**

The decision tree algorithm works by splitting the dataset recursively, which means that the subsets that arise from a split are further split until a predetermined termination criterion is reached. At each step, the split is made based on the independent variable that results in the largest possible reduction in the heterogeneity of the dependent variable. Also, pruning is done to reduce the chances of overfitting the tree to the training data and reduce the overall complexity of the tree.

**Model Results:**

Although there was no major improvement in the performance of regression models from Linear to Decision Tree, there is a very slight increase. We found that the Black Friday dataset performs the best on Lasso Regression with an accuracy of 0.6412237. Age, Gender, Marital status, Occupation, City category, and Product category to determine the Purchase amount and to see which regression models perform best on the Black Friday Dataset.

**Below are the results of all the Models Training and Test data.**

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

**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 can help to improve the conversion rate by helping the customer to find products she/he wants to buy faster, promote cross-selling by suggesting additional products and can improve customer loyalty through creating a value-added relationship.

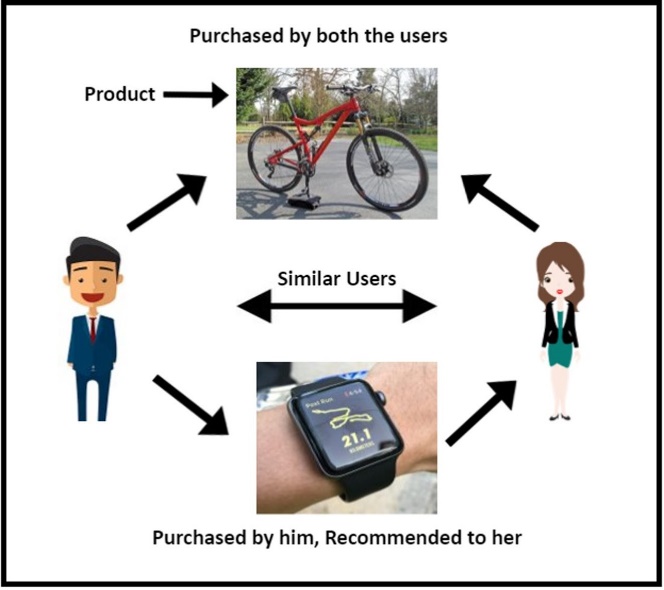
Recommender systems are categorized basically into two categories content-based approaches and collaborative filtering. In this project we are focusing on collaborative filtering of recommender algorithms. The R extension package “*recommenderlab*” provides a general research infrastructure for recommender systems.

**Collaborative filtering Approach**

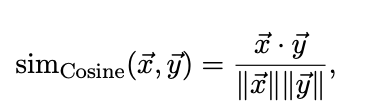
Collaborative filtering (CF) uses given rating data by many users for many items as the basis for predicting missing ratings or for creating a top-N recommendation list for a given user, called the active user. Formally, we have a set of Users U = {u1, u2, . . . , um} and a set of Products P= {p1, p2, . . . , pn}. Ratings are stored in a m × n user-item rating matrix coc\_mat = Cmxnwhere each row represents a user u[j] with 1 ≤ j ≤ m and columns represent products p[k] with 1≤ k ≤ n. Cmxnrepresents the rating of user uj for product Pj .

Rating scale is 1 if user bought a product otherwise 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 i.e., whose purchasing habits are more like object user, their products which are not purchased by the object user yet are recommended.

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 hrs 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

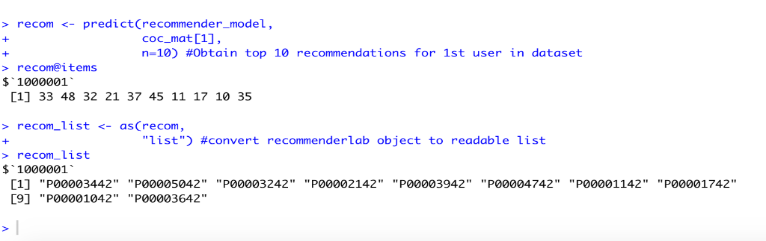
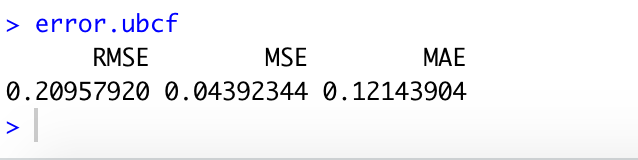




Now that our model is trained with the data and we can do recommendations for user(s). We tested it for a User\_id “1000001” and the results are shown below (Fig-1). There are the top 10 products recommended by the model to the user.

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 below (Fig-2).

Fig-1 Fig-2

**CONCLUSION:**

As a part of learning, the course concepts were implemented beginning from Data cleaning, EDA, Hypothesis testing, Model Building to Product Recommendation.

From the analysis of the hypotheses, we can provide suggestions to the retail store on which segment, demographic and city category the store can focus its marketing efforts.

After performing detailed EDA, Product\_Category\_2 and Product\_Category\_3 contained null values and were least correlated to our response variable “Purchase”. Hence, we excluded the variables.

We performed regression to predict the “Purchase” using the Multiple linear regression, Ridge, Lasso, and decision tree. From the above results, all models performed fairly well.

Hence, from model selection and Product recommendations, by understanding the Purchase Patterns of the Customers, retailers can provide improved Service Quality.

**References:**

<https://www.kaggle.com/sdolezel/black-friday>