**Black Friday: Insights on Customer Purchases**

**Hrishikesh Mahajan Jainil Parikh**

**Department of Computer Science Department of Computer Science**

**Binghamton University Binghamton University**

1. **INTRODUCTION**

Black Friday is an informal name for the Friday following Thanksgiving Day in the United States, which is celebrated on the fourth Thursday of November. 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. Black Friday has routinely been the busiest shopping day of the year in the United States at least 2005, and possibly longer. Since the start of the 21st century, there have been attempts by retailers with origins in the United States to introduce a retail "Black Friday" to other countries around the world. In several countries, local retailers have attempted to promote the day to remain competitive with US-based online retailers.

Black Friday is a shopping day for a combination of reasons. As the first day after the last major holiday before Christmas, it marks the unofficial beginning of the Christmas shopping season. In order to take advantage of this, virtually all retailers in the country, big and small, offer various sales including limited amounts of doorbuster/door crasher/door smasher items to entice traffic. On April 23, 2014, Black Friday joined a growing list of ICANN top-level domains (such as—traditionally—.com, .net, and .org). According to 2018 data the average global growth of interest in Black Friday sales was 117%. That is why it is critical and profitable day for the retail sector. It also creates the new competition for the retailers and to keep up with this, retailers need to dive deeper into advanced Black Friday marketing strategies.

Through our project, we plan to demonstrate an industry followed standard CRISP-DM (Cross Industry Process for Data Mining) procedure in order to understand the existing trend of data, analyze the results according to the data clusters and to generate predictions for customer purchases. This is interesting because, to keep up with the competition, retailers need some insights and cluster of customers to cluster them into groups and to target them. So, retailers can use different marketing strategies to attract different types of customers. With our findings, we represent a view of how a retailer can choose from products that will be most popular, advertise them to the right group and also include multiple offers and deals to maximize the profit.

1. **TECHNICAL APPROACH**

The dataset we worked upon is freely available on Kaggle. Figure1 represents a general view of the data. We follow the CRISP-DM methodology, which provides a structured approach to planning a data science project. It is a robust and well-proven methodology and also used as a standard in the industry.

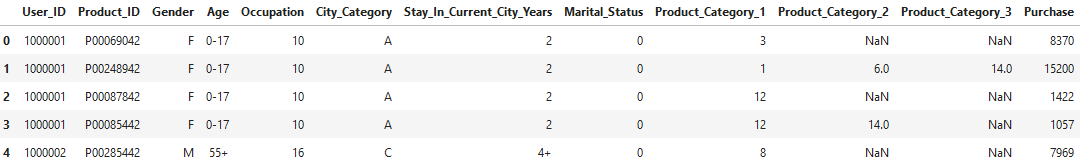


Figure 1: Description of the dataset

We plan to use multiple supervised machine learning models like Decision Tree, Linear Regression, Random Forest and XG Boost Regressor and evaluate the outcomes one by one.

1. **RESULT AND DISCUSSION**

We describe our results and findings in terms of the CRISP-DM process as explained below.

* 1. Business Understanding: To keep up with the competition, retailers need some insights and cluster of customers to cluster them into groups and to target them. Retailers can use different marketing strategies to attract different types of customers.
  2. Data Understanding: A brief analysis on the dataset reveals that we have about 5,50,000 transactions made by 5,891 users for over 3600 unique products. We find that some columns have a lot of missing data, some fields are categorical, some fields have the quartile values as integer and columns like UserID which do not provide us much information except the total count of customers. To have a better understanding of the existing trend, we formulated six questions which we answer by the graphs in Figure2. The questions are:
     1. Which age group has done more transactions?
     2. Have males done more transactions than females?
     3. Do users’ occupation have any relation with number of transactions?
     4. Do users who are living in city for more than 1 year have effect on number of transactions?
     5. Does category of city have any impact on purchase quantity?
     6. Does marital status have any correlation with Black Friday purchase value?

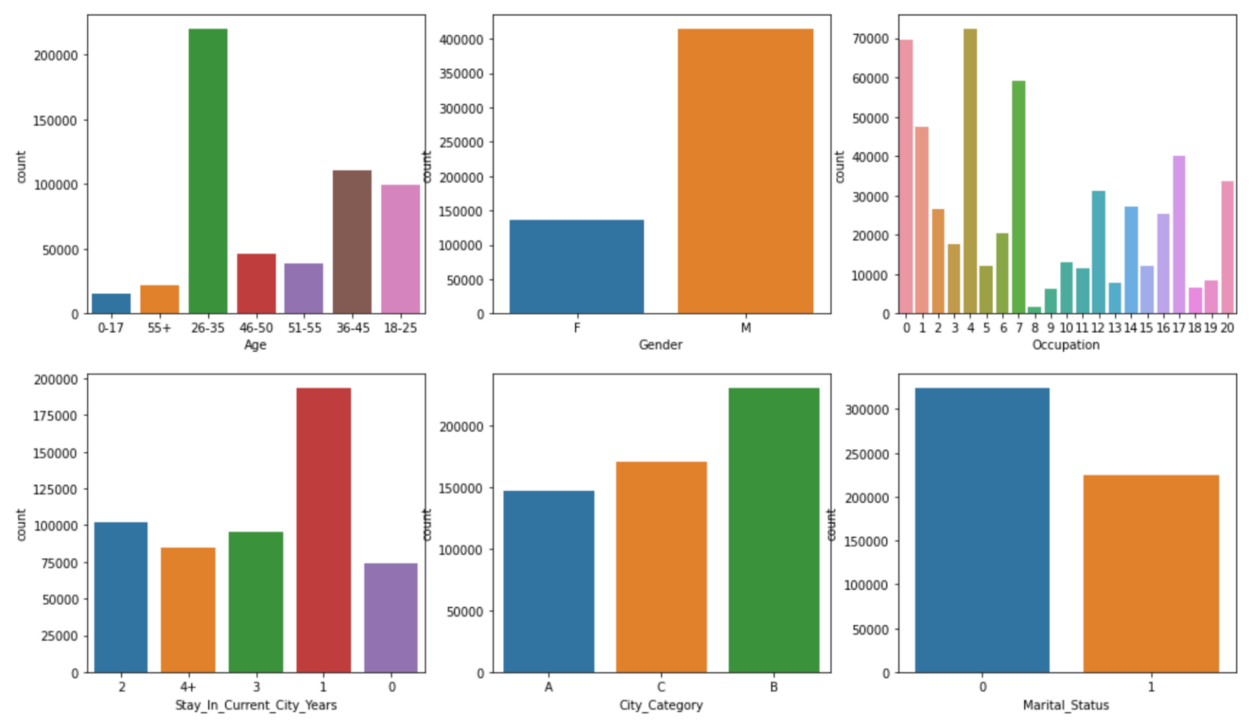


Figure 2: Understanding general trend of Data

Answers:

1. In chart 1, 26-35 age range have done more transactions.
2. In chart 2, Male users have done more transactions.
3. In chart 3, Users whose occupation are 0 and 4 have done more transactions.
4. In chart 4, Users who stayed in the current city for 1 year have done more transactions.
5. In chart 5, Users from city category B have done more transactions.
6. In chart 6, Users who are not married yet have done more transactions.
   1. Data Preparation: For our Machine Learning models to predict the consumer spend, we need to have a well-formatted, structured and cleaned data. In the process of Data Wrangling, we identified that Product\_Category\_2 had about 30% missing data. So, we used random sampling to fill the missing values. Figure3 shows the distribution of data before and after the imputation.

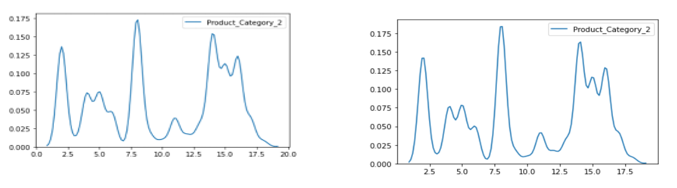


Figure 3: Distribution of Product\_Category\_2 before (left) and after (right) random fill.

The same could not be done for Product\_Category\_3 because it had almost 60% missing data and applying random fill, changed the distribution of data significantly.

* 1. Modeling: Since our dataset provides us with a separate test set to evaluate our results, we preferred to use Supervised Machine Learning models in order to gain better accuracy of the prediction. We used train\_test\_split method to divide all our data into two sets: training and testing set. We also use the StandardScaler to scale our data to make each and every data point just as important as the other one. Figure4 describes the code to split and scale the data.

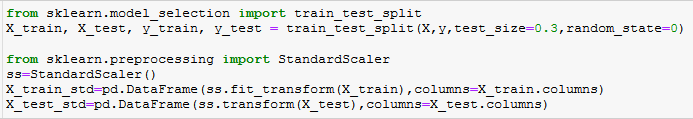


Figure 4: Splitting and Scaling training and test sets.

We use the R^2 and the RMSE metric scores to evaluate the accuracy of our predictions. The RMSE indicates the absolute fit of the model to the data–how close the observed data points are to the model’s predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit. R-squared scores range from zero to one, with zero indicating that the proposed model does not improve prediction over the mean model, and one indicating perfect prediction. For every algorithm, we tweaked around with the parameters and set the optimum values for the parameters and calculated the metric scores.

1. Linear Regression: This algorithm gave us particularly low metric scores for both, training data as well as testing data. Figure5 shows the R^2 and RMSE scores for the two sets.

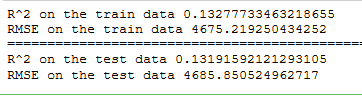


Figure 5: R^2 and RMSE scores using Linear Regression algorithm

1. Decision Tree: We used parameters for this algorithm such as max\_depth = 10 and min\_samples\_leaf = 500. This algorithm gave us good accuracy over Linear regression, about 68%. Figure6 shows the R^2 and RMSE metric scores for Decision tree.

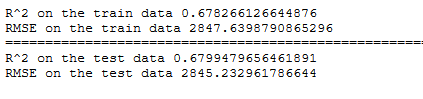


Figure 6: R^2 and RMSE scores using Decision Tree algorithm

1. Random Forest: For Random Forest regressor, we used the following parameters n\_estimators=30, random\_state=3, max\_depth=15, min\_samples\_split=100. This algorithm improved our prediction accuracy to about 72%. Figure7 shows the R^2 and RMSE metric scores for Random Forest regressor.

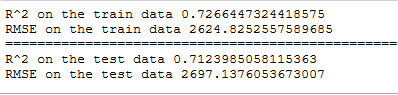


Figure 7: R^2 and RMSE metric scores for Random Forest regressor.

1. XGBoost: After tweaking, we used the following parameters for XGBoost algorithm. learning\_rate=0.05, n\_estimators=500, max\_depth=10. This algorithm gave us the maximum accuracy in training, about 80% and about 75% on the test set. Figure8 shows the R^2 and RMSE metric scores for XGBoost.

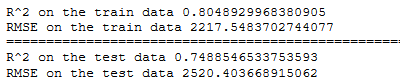


Figure 8: R^2 and RMSE metric scores for XGBoost.

1. **CONCLUSION AND FUTURE WORK**

The last part of our CRISP-DM process is evaluation which we describe in detail in this section.

1. Evaluation: In this project, we applied the CRISP-DM process model as a set of guidelines to help us plan, organize, and implement our project. We used Supervised Learning based Machine Learning algorithms to predict the amount a customer is likely to spend on Black Friday. We stand to a good result accuracy of 75% given by the XGBoost algorithm. We also note a bit of difference in the training scores and testing scores of the Random Forest and XGBoost algorithms. This is because of the fact that our models overfit the data to a certain extent. In reality, the dataset often has some degree of error or random noise within it. Thus, attempting to make the model conform too closely to slightly inaccurate data can infect the model with substantial errors and reduce its predictive power.

With our findings, retailers and shop owners have a better understanding of some questions such as What products are customers going to purchase, how much will a customer spend, which age-group is going to spend more, and more. Retailers and shop owners now have a detailed analysis report and a trained model which will help them to perform clustering on consumers, and based on the cluster’s characteristics, they can offer various types of discounts and deals.

A Future scope of our project can be such that, we use this model to analyze daily trends of purchases by customers for holidays and festivals. Our project could also be used for customers shopping for the Black Friday, online. The project can be scaled to include more dimensions and attributes. To improve more on the accuracy of the prediction, we also foresee the use of Ensemble Learning models to tackle the issue of overfitting.

1. **REFERENCES**
2. **Dataset:** [**https://www.kaggle.com/sdolezel/black-friday**](https://www.kaggle.com/sdolezel/black-friday)
3. [**https://matplotlib.org/tutorials/introductory/pyplot.html**](https://matplotlib.org/tutorials/introductory/pyplot.html)
4. [**https://seaborn.pydata.org/**](https://seaborn.pydata.org/)
5. [**https://scikit-learn.org/stable/**](https://scikit-learn.org/stable/)
6. [Black Friday 2020 Predictions from the Experts | BestBlackFriday.com](https://bestblackfriday.com/blog/ask-the-experts-2020-edition/)
7. [**An exploratory investigation of Black Friday consumption rituals**](https://www.emerald.com/insight/content/doi/10.1108/09590551111144905/full/html?casa_token=It9OdIhs0AUAAAAA:Xw7tdfGfVaDYaxAYJuiei1vZAzLTutdta3pRLXY_66sMSE-HyZXf7cdKjVPsudRy8NSzR7qIwvs-dt5yvIP1sRDR-EKtxUJEhfq4gKpPEfxxUF82Y3TF)
8. [**An analysis of consumer behavior on Black Friday**](https://thekeep.eiu.edu/fcs_fac/13/)
9. [**BLACK FRIDAY SALE PREDICTION VIA EXTREME GRADIENT BOOSTED TREES**](http://vap.ac.vn/Portals/0/TuyenTap/2020/4/18/fa7f0835392b41f4a4751e1593be09f5/8-132-Black-Friday.pdf)