Black Friday Sales Predictions

Alisia-Nadia Sarb, Ana-Maria Bulat, Oana Totia, Minh Cao

Abstract: Black Friday has been considered as a shopping holiday — a consumption ritual since 1952[1]. It is one of the most popular retail and spending event in the United States, where shops are crowded because of the red marked price– discount.
This study focuses on finding what influences people to buy more on Black Friday and also predicting the amount people will spend this year on Black Friday. In order to find the answer to these questions, we've applied different models that can predict the sales along with an analysis that includes visualization techniques. The study revealed that the city and the years lived there as well as the occupation and the age of a person play an influential role in the amount a person is willing to spend. The study also shows that, at least in our case, the Extra Tree Regression model has the best accuracy in predicting future sales.

Keywords: Black Friday, Prediction models, Visualization, Factors, Regression, Machine Learning.

1. Introduction

According to Investopedia[2], Black Friday has two relevant meanings. In history, Black Friday was a stock market catastrophe that took place on September 24, 1869, when the price of gold plummeted, and the markets crashed, because of rampant speculation. Black Friday started in the USA and has been widely spread to other countries. Retailers see this event as the time to make more money, so the information about the sales has become very vital for the shops in order to maximize the profit. Conversely, many shop owners take it as a sign of trouble if predictions are unable to meet the expectations.

Being well-prepared and understanding the customers' needs for Black Friday will bring more profit for them. Therefore, the aim of this project is applying several techniques in order to analyze consumers' shopping habits that become essential for retailers to develop business strategies. Furthermore, one of the solutions for owners to tackle either the sales will increase or decrease, is applying a predictive model that can prognosticate future sales based on customers' spending from last months. The paper is build up on the sample transactions made in a retail store that includes demographic information about the customers. The major problem is that the existing data-set contains several inconsistent data such as missing values or wrong information. The general overview and data pre-processing will be further discussed in Section [3.1] with the intent of making more sense out of it.

In order to get an accurate prediction, the following are done: data manipulation, applying visualization techniques that helps to understand the customer behaviour and selection of the right algorithm. Some questions involved in this case study are: "Will data pre-processing and visualization technique increase accuracy?" and "Which algorithm is better and efficient for such model?".

This study is organized as follows: Section[2] underlines the methods utilized in order to analyze the data-set and implementation of the machine learning algorithms that predict the sales in Black Friday, later on, in Section[3], the data is manipulated with the help of various visualization techniques that prioritize the importance of the features that influence the sales. Section[4] demonstrates machine learning algorithms applied along with its accuracy and performance. By the end of the report, Section[5] and Section[6] discusses the findings and final conclusion reached, based on the whole research observations.

1.1. Research question

In order to make easier to answer the main question, four sub questions were created.

– What are the determinants that influence the purchases on Black Friday?

1. Which gender spend more?

2. Is there a difference between married and unmarried people in terms of amount spent?

3. Who spend more, new residents or those with more years lived in the City?

4. Does people age and occupation affect the amount they spend?

2. Methods

Statistical methods has an important role in machine learning. We use it to gain a deeper understanding about the data, clean and prepare it for machine learning algorithms. Our target variable is purchase amount. As we are dealing with a regression problem, machine learning methods such as Extra Trees Regression 4.3, Random Forest Regression 4.4, Decision Tree Regression 4.5 and XGBoost Regression 4.6 were implemented. Then using Root Mean Squared Error(RMSE) to compare them (see section 4.1).

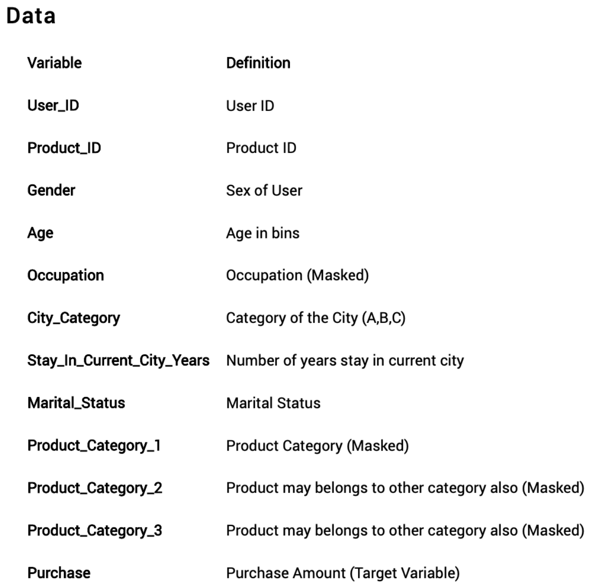
3. Analysis

3.1. Data Set

3.1.1. Data Overview

The data that is going to be used in the project is provided by Analytics Vidhya and can be accessed from the following URL[3], it represents sales transactions made during the last month.

The data-set has 550K records, that represents general information of 5891 buyers including 6 customers' features (gender, age, occupation, marital status, living city and years that they have stayed in the city), product details(product id and product category) and total amount spent during last month(Purchase). The date set was split into a train set and a test set. The test set contains 20% of the data-set.



Data

Variable	Definition
User_ID	User ID
Product_ID	Product ID
Gender	Sex of User
Age	Age in bins
Occupation	Occupation (Masked)
City_Category	Category of the City (A,B,C)
Stay_In_Current_City_Years	Number of years stay in current city
Marital_Status	Marital Status
Product_Category_1	Product Category (Masked)
Product_Category_2	Product may belongs to other category also (Masked)
Product_Category_3	Product may belongs to other category also (Masked)
Purchase	Purchase Amount (Target Variable)

Figure 1: Data Records

3.1.2. Importance of Data Pre-Processing

In real-world datasets are always present incomplete and inconsistent data points that may contain many errors. Converting these data into a format that the predictor can understand is called pre-processing. Data pre-processing is a data mining technique that involves transformation of raw data into an understandable format.

For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. According to [4] the steps taken into consideration are:

- Data Cleaning: Data is cleansed through processes such as filling the missing values with calculated mean, median, mode or zero value of entire rows values of that particular column or delete the row of the data having the missing values. It will have a negligible effect on getting accurate or predicting the output.
- Data Transformation: Modification of categorical or text values to numerical values.

3.1.3. Data Preparation

Initially, the data was not compatible to be passed through the machine learning algorithms since the algorithms require their input to be numerical.[5]

After analyzing the data, we discovered that some features like age were categorical and product category has missing values, that we fulfilled with zero. We also converted categorical data to numbers. The following Table 1 demonstrates an overview of the data mapping and transformation.

Feature	Initial data type	Mapping
Gender	F	0
Gender	M	1
City	A	0
City	B	1
City	C	2
Married	M	0
Unmarried	U	1
Age	0-17	17
Age	18-25	25
Age	26-35	35
Age	36-45	45
Age	46-50	50
Age	51-55	55
Age	55+	57
City Years	1	1

Table 1: Data mapping

3.2. Purchase Amount Analysis

Purchase amount is the most important variable to predict the Black Friday sales for a retail shop. For this reason this paper presents an analysis of the distribution of the data based on the following observations.

- Purchase Amount by Gender
- Purchase Amount by Marital Status
- Purchase Amount by Occupation

- Purchase Amount by Age
- Purchase Amount by City Category

Table 3: Purchase amount in correlation with other parameters.

3.2.1. Purchase amount by gender

The data was modified as mentioned previously so that ”F” is represented as ”0” and ”M” as ”1”.

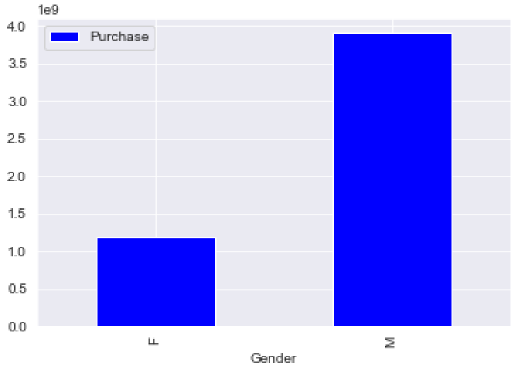


Diagram 1: Purchases by gender

Easily noticeable, there is a substantial difference between males and females when it comes to the purchase amount. In this case it is interesting to see the number of males and females in the study.

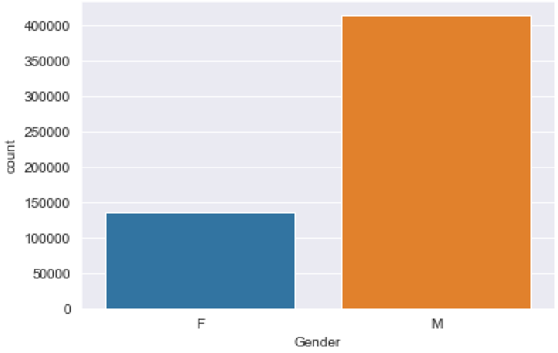


Diagram 2: Gender count

It was noticed that the number of males in the study is about three times larger than the number of females so a further analysis of per-head purchase amount by gender is conducted by comparing the purchase amount average of the two categories. In both cases the average is approximately 9$ so the per-head purchase amount is constant.

3.2.2. Purchase by marital status

For further analysis, it is easy to see the most popular products among customer with age range from 18 to 45 are still in category 1, 5, 8 even if he/she is married or not.

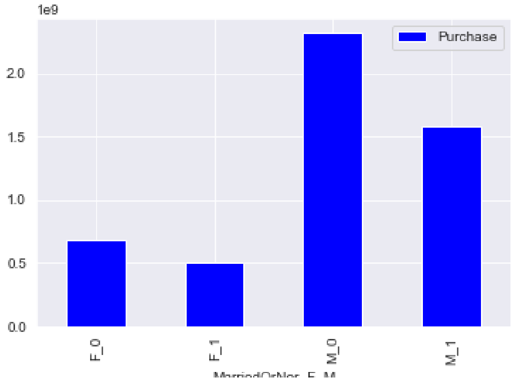


Diagram 3: Purchase by Marital status

3.2.3. Purchase amount by occupation

The demographic information presented by the data contains the occupation of the client witch is represented as a number from 0 to 20. The real meaning of this number is only known by the ABC Private Limited retail shop.

During this part of the analysis we tried to see if people tend to shop more or less based on their occupation category.

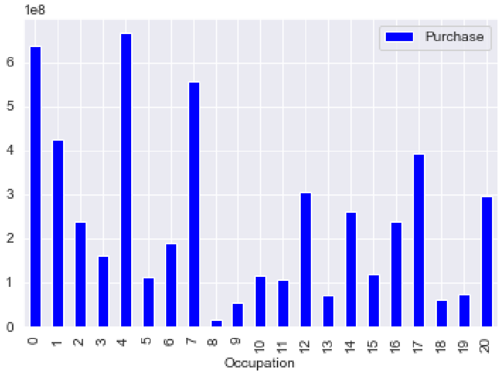


Diagram 4: Purchase by occupation

3.2.4. Purchase amount by age

The age of the client is also a factor of interest. Looking to the diagram above, we can see that most of the costumers are in age range between 26-35 and mostly single.

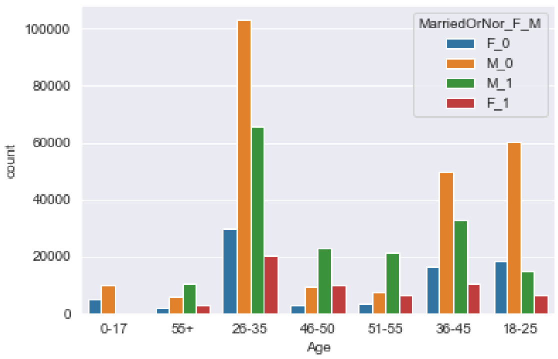


Diagram 5: Age - Marital status count

Also, we can observe that total amount spent in purchase is in accordance with the number of purchases made, distributed by age.

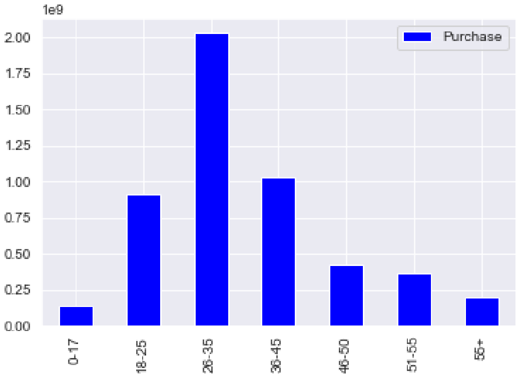


Diagram 6: Purchase by age

3.2.5. Purchase amount by years lived in city-x

Based on purchase amount per city category, we can assume that category C is a bigger shop or placed in a city with more population.

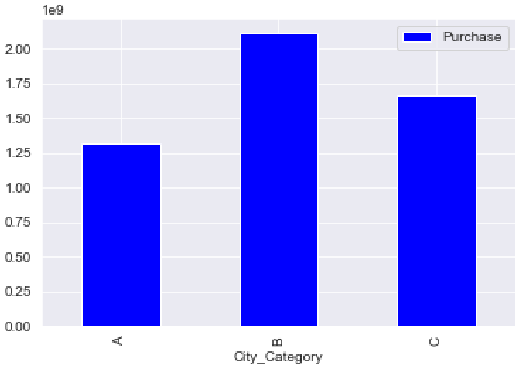


Diagram 7: Purchase by city category

Also further analysis has shown that customers that are new residents tends to spend more money.

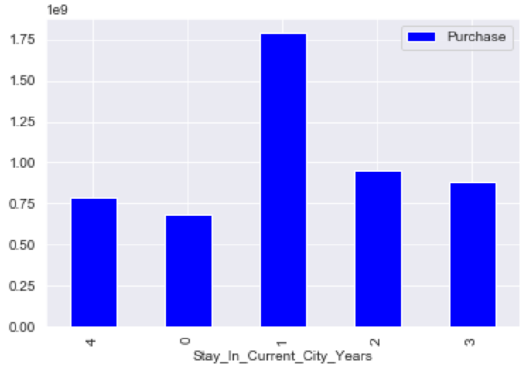


Diagram 8: Purchase by years lived in City X

4. Machine Learning Techniques

4.1. Performance Measure

In order to give an idea of how much error the system makes in its predictions with weight for large errors, the Root Mean Square Error(RMSE) was used.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}$$

Figure 2: RMSE Function

4.2. Liniar Regression

It is used to determine the extent to which there is a linear relationship between a dependent variable and one or more independent variables.[6]

The RMSE for the model was 4627.

And the accuracy was: 15.25%

This model did not have a good result but we decided to keep it in order to show that not all the models are suitable for all the date-sets.

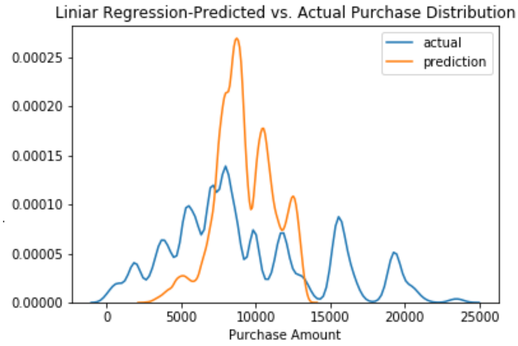


Diagram 9: Liniar Regresion

4.3. ExtraTree Regression

Extra-trees implements a meta estimator that fits a number of randomized decision trees on diverse sub-samples of the dataset and applies average to increase the predictive accuracy and handle over-fitting.[7] The RMSE for the model was 2231. And the accuracy was: 80.29%

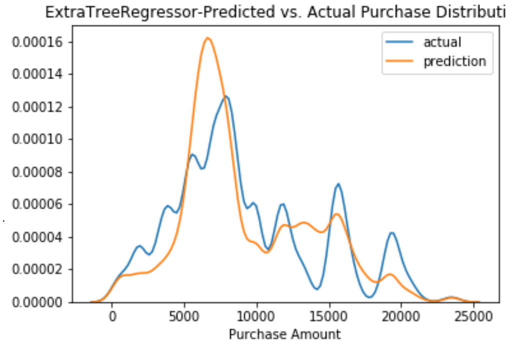


Diagram 10: ExtraTreeRegresor

4.4. Random Forest Regression

Random Forests are an ensemble of k untrained Decision Trees (trees with only a root node) with M bootstrap samples (k and M do not have to be the same) trained using a variant of the random subspace method or feature bagging method.[8] The RMSE for the model was 2956. And the accuracy was: 65.41%

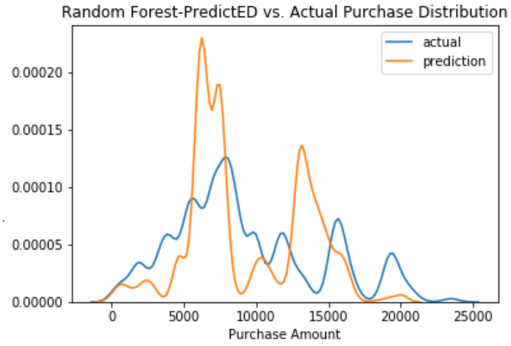


Diagram 11: RandomForestRegressor

4.5. Decision Tree Regression

Decision Trees is a non-parametric supervised learning method that create a model that predicts the value of a target variable by learning simple decision rule inferred from the data features.[9] The RMSE for the model was 2898. And the accuracy was: 66.75%

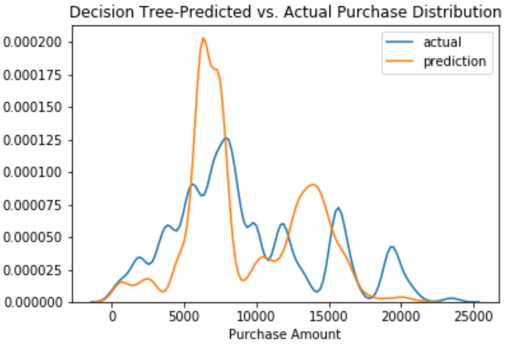


Diagram 12: DecisionTreeRegreson

4.6. XGBoost Regression

XGBoost stands for eXtreme Gradient Boosting. Gradient boosting involves creating and adding decision trees to the model sequentially.[10]

During this study we used the XGBoost model with a learning rate of 1 and a with a number of trees of 1000. The RMSE of the model was: 2796 And the accuracy was: 69.05%

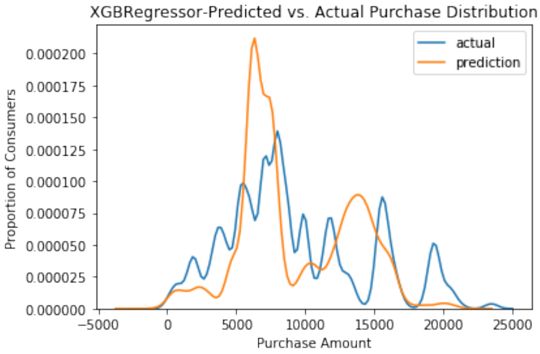


Diagram 13: XGBR

4.7. Comparison of algorithms

For visual comparison, the Diagram [13] represents the plot of implemented algorithms and their RMSE.[11]

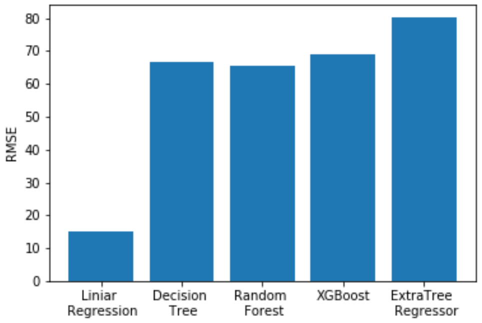


Diagram 14: Algorithms vs Accuracy rate

5. Findings

The aim of this study was to see what influences people when it comes to spending money.

One of the first things we checked when making the analysis is if the gender 1 and the marital status 2 of a person affects the amount of money spend. As it can be seen in Purchase amount by gender 3.2.1 on average, both females and males spend the same amount. However, when it comes to marital status, we've noticed that married males tend to spend more than a married woman or single people as can be seen in Diagram 3 3.2.1.

The city and the number of years a person has lived there does affect the amount a person spends. In Diagram 8 3.2.5, people who live in city B(after the data modification city 1) tend to spend more than the people from the other cities. Also, from Diagram 73.2.5, we can conclude that new residences tend to buy more things, maybe they will take the advantage of the discount on Black Friday to purchase all the things needed for new place.

Additionally, Diagram 4 3.2.3 shows that someone's job can affect the amount they will spend. The diagram reveals that people with job categories like 0, 4 or 7 tend to spend more that people with a job category like 8, 18 or 19. Since we do not know for sure the meaning of each category we can assume that people from categories like 0, 4 and 7 have a higher salary than the people from categories like 8, 18, 19.

During this study we have also found the best types of machine learning models to use for this data set. The predictive power is measured by RMSE, where the models with lower RMSE is desirable. As it is illustrated in Diagram 13 4.7, Extra Tree Regression model has higher accuracy rate and suitable for our data set, while neural network will be over-fitting and too advance for such regression problem.

6. Conclusion

This paper focuses on analyzing important factors that can influence the purchases on Black Friday using statistical methods combined with Machine learning.

Based on the findings, we concluded that while gender and marital status does not necessarily have a major impact in the spending amount, the city and the years living there as well as age and occupation make a big difference when it comes to the amount a person will spend on Black Friday. Thus, retailers can use this information to personalize customers' needs and develop suitable business decisions.

Additionally, Extra Tree Regression is the one that shows more accuracy in the prediction of the purchase amount and could be used further to prognosticate the data needed for Black Friday.

The Jupyter Notebook can be found here: https://github.com/amy131313/Black-Friday-Prediction.git

References

[1] Jane Boyd Thomas and Cara Peters. An exploratory investigation of Black Friday consumption rituals. *International Journal of Retail Distribution Management*, 39(7):522–537, 2011.

[2] Investopedia. Black Fridaty. https://www.investopedia.com/terms/b/blackfriday.asp.

[3] Analytics Vidhya. Black Friday Problem. https://datahack.analyticsvidhya.com/contest/black-friday//.

[4] https://www.datasciencelearner.com/top-steps-data-preprocessing-machine-learning/.

[5] Hugo Ferreira. 7 Steps to Mastering Data Preparation for Machine Learning with Python.

[6] https://ml-cheatsheet.readthedocs.io/en/latest/linear_regression.html.

[7] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html.

[8] https://towardsdatascience.com/decision-trees-and-random-forests-for-classification-and-regres

[9] https://scikit-learn.org/stable/modules/tree.html#regression/.

[10] https://xgboost.readthedocs.io/en/latest/.

[11] Saravana Gunaseelan Ching-Seh (Mike) Wu, Pratik Patil. Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data Ching-Seh. *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, 18-Novem(2):16–20, 2019.