Black Friday dataset - A Random Forest Regression Approach

Motivation

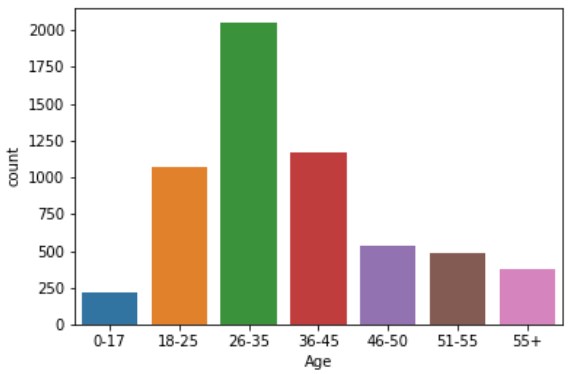
In the USA, "Black Friday" is the name used to describe the day after Thanksgiving and is regarded as the holiday season opening. This period of the year is considered of major importance for the American economy since it is when up to 30 percent of annual retail occurs.[1] Because of its impact, retailers use data from this particular day for sales strategies.[2] In this context, the Black Friday dataset was examined and a **predictive analysis of the amount of purchase** was done based on customers personal characteristics.

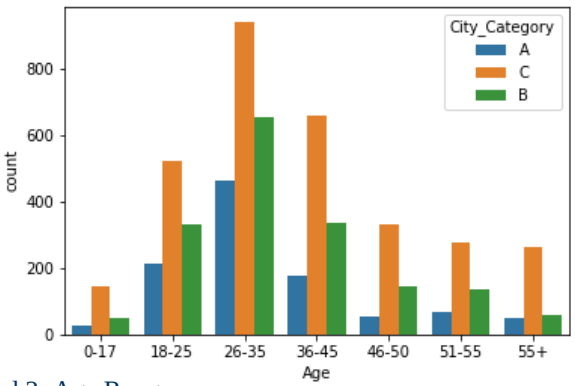
Exploratory Analysis

The first aspect that was analysed were the variables contained in the dataset, 12 in total. The first variable, **User_ID**, provides us an identification for each customer. In total the dataset gathers observations related to purchases of 5891 people.

As regards to **gender** distribution of the customers, the numbers call attention: 72% of customers were male. According to Swilley and Goldsmith, in "Black Friday and Cyber Monday: Understanding consumer intentions on two major shopping days", the variable "gender" does not moderate differences between attitudes toward shopping and intentions to shop during Black Friday[2]. That said, the unexpected difference between the amount of male and female customers might be explained by the studied store being part of a specific retail sector.[3]

In respect to the **age** distribution, it was observed a peak in the amount of customers between 26 and 35 years old, corresponding to 35% of the total, followed by the age range between 36 and 45 years old (20%). This behaviour is present across genders and among customers from the city category C. The age range between 18 and 25 years old is the second more important for the city category A and as important as the range 36 and 45 years for the city category B.





Figures 1 and 2: Age Range.

[1] https://nrf.com/media-center/press-releases/nrf-forecasts-holiday-sales-will-increase-between-43-and-48-percent
[2] SWILLEY, E., GOLDSMITH, R. E. "**Black Friday and Cyber Monday: Understanding consumer intentions on two major shopping days**", Journal of Retailing and Consumer Services 20 (2013) 43–50.
[3] https://www.statista.com/statistics/640185/black-friday-cyber-monday-wish-list-us-by-gender/

Although a precise confrontation of the age information with official data was not possible, the observed peak in age range between 26 and 35 years old is in agreement with the data provided by the National Retail Federation for 2018.[4] Analysing the variable age against the total amount purchased per customer, the age range in question also presented the highest average.

The **occupation** of customers was encoded in the dataset, ranging from 0 to 20. It was observed a peak in customers from the occupation number 4 (13% of total), followed by occupations number 0 (12%), and 7 (11%).

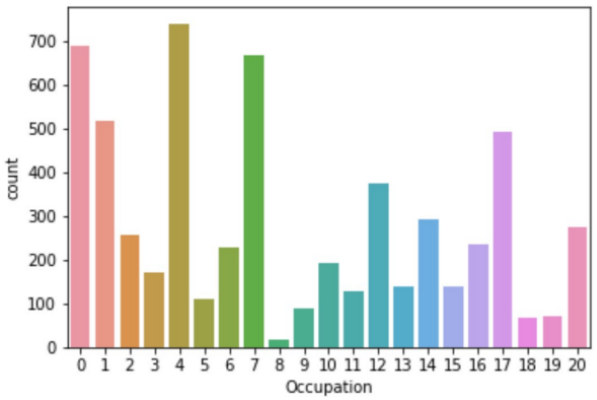
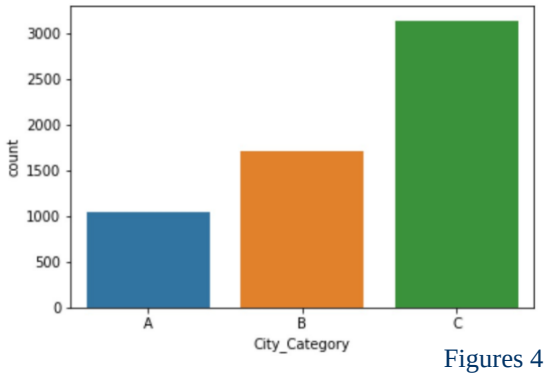
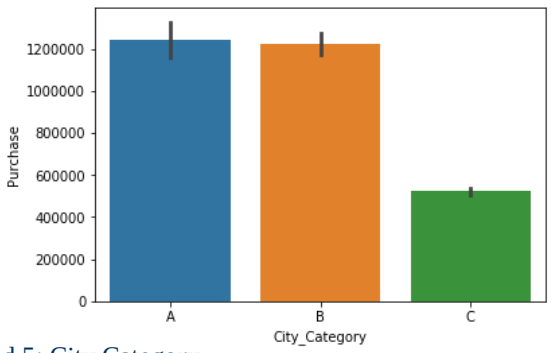


Figure 3: Occupation.

The **city category** (A, B, and C) seems to play an important role in respect to sales. Although the higher amount of clients being from cities categorized as C (53%), customers from cities categorized as A and B presented a higher average in the amount purchased per customer. This behaviour can be explained by the tendency of a larger income in bigger cities (A and B).





Figures 4 and 5: City Category

The variable **Stay_In_Current_City_Years** is divided in five levels: 0, 1, 2, 3, 4+. It was observed that most part of customers moved quite recently and stayed in the current city for up to two years (67% of total). As regards to the **marital status**, most part of customers (58%) were single.

In the dataset it is also possible to notice four variables related to products: Product_ID and Product_Category (from 1 to 3). Analysing the **Product_ID** variable it was possible to notice that there are 3623 different products available. As regards to the

[4] https://cdn.nrf.com/sites/default/files/2018-11/Natural-Insight-In-Store-Holiday-Shopping-2018.pdf

product categories, many observations presented missing data (69.44% of observations for categories 2 or 3). It was chosen to focus on the total purchase per client and remove the product information from the analysis.

Finally, the variable **Purchase** was analysed. It was noticed that the total amount purchased per customer presented an extremely skewed distribution. Such behaviour of the variable was already expected, since it is known[5] that monetary amounts are often log-normally distributed. As anticipated, the logarithm of the variable indeed follows a bell-shaped curve that can be compared with a Gaussian, with standard deviation equal the mean of the distribution (figure 6). In this context, the transformed variable, log-purchase, was considered for further investigation.

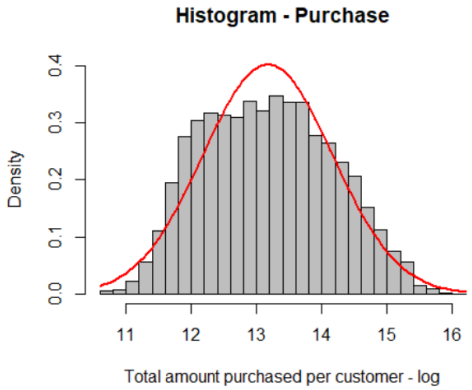


Figure 6: Log-Purchase against a Gaussian.

The Random Forest Approach

Aiming to build a model that could predict the total amount purchased per customer considering their personal characteristics, it was chosen a Random Forest approach in Python. Considering our dataset, composed by both numerical and categorical data, a tree method appeared to be suitable. The Random Forest Regression method is a supervised learning algorithm. It constructs a forest of decision trees with 1/3 of the total variables each (in order to decorrelate them) bootstrapping the training set. The output is the mean prediction of the individual trees.[6]

Firstly, the data were grouped by User_Id, summing the amount of purchase per customer. As aforementioned, the variables related to products were removed from the analysis and the purchase values were log-transformed. In order to perform the algorithm, the Python library Scikit-learn was used. Firstly the dataset was split between train and test set in the proportion 60% and 40%, respectively, using the function “train_test_split”. The module called “RandomForestRegressor” was imported and 100 decision trees (due to computational limitations) were considered.

In order to train the model, the function “RandomForestRegressor” was used. To make predictions, the function “predict” was employed using the test set. Finally, an additional major point to be considered were the variable importances. The Python library Scikit-learn provides a function called “.feature_importances_” that returns features that contributed the most in decreasing node impurity.[7]

---

[5] ZUMEL, N., MOUNT, J. "**Practical Data Science with R**", 1st Edition, Manning Publications Co. Greenwich, CT, USA, 2014.
[6] JAMES, G. Et al. “**An Introduction to Statistical Learning : with Applications in R**”, New York :Springer, 2013.
[7] https://scikit-learn.org/stable/modules/feature_selection.html

Model Performance

In order to asses the prediction accuracy, it was considered the Mean Absolute Error (MAE) and the Mean Absolute Percentage Error (MAPE)[8]. Additionally, it was also computed the OOB score, used to estimate the R^2 on unseen data (out of bag).[9] The model presented accuracy (MAPE) of 93,51%, MAE of 0.85 (log-purchase), and OOB score corresponding to a R^2 of 0.90.

Discussion

As briefly discussed in previous sections, a graph with the variable importances was plotted. In the PhD dissertation entitled "Random Forests: From Theory to Practice", by Gilles Louppe, it is concluded that importances are a sound and appropriate criterion for assessing the usefulness of variables.[10]

As can be seen in figure 7, the occupation is the variable most relevant when analysing the total purchase per customer, followed by city category, years that the client stayed in the current city, and age range.

The variable importances achieved by the model reflect many points highlighted during the exploratory analysis. Interestingly, however, is that despite the fact that 72% of clients were men, the gender did not appear to be so relevant in determining the amount of purchase.

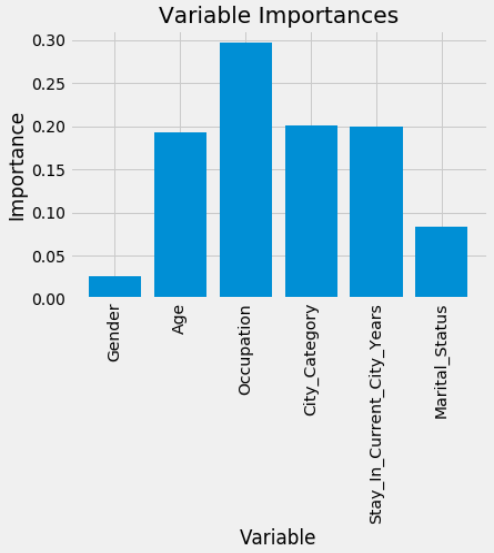


Figure 7: Feature Importances.

problem statement - Letícia
solution design - Letícia
solution development - Eduardo
data collection - Eduardo
writing - Letícia/Eduardo

[8] MYTTENAERE, A. Et al. "**Mean Absolute Percentage Error for regression models**", Neurocomputing 192 (2016) 38–48.
[9] https://scikit-learn.org/stable/modules/generated/ sklearn.ensemble.RandomForestRegressor.html sklearn.ensemble.RandomForestRegressor
[10] LOUPPE, G. "**Random Forests: From Theory to Practice**", PhD dissertation, University of Liège, 2014.