Black Friday data analysis and Prediction

***Abstract*— The Friday following Thanksgiving is referred to as "Black Friday" in American slang. It traditionally signals the beginning of the holiday shopping season in the US. Many retailers open early, often as early as midnight or even on Thanksgiving, and frequently have heavily advertised bargains with steep discounts. Data from customers who purchased products on this day might be examined to provide a quick statement of their preferences for specific products. Here, we have examined the data tuples of customers together with their purchasing criteria and quantity. This data is examined and forecasted solely to offer clients special discounts on goods based on their preferences and purchasing power. Data analysis currently makes use of data science with machine learning (ML), which teaches the computer using models and offers better accuracy than business intelligence [1].**

Keywords—Black Friday, machine learning, random forest, visual analytics, deep learning

I. Problem Statement

The goal of the current study is to investigate the traditions around collective consumption on Black Friday, the day following Thanksgiving and one of the busiest shopping days in many nations throughout the world.

There were just tiny companies back then; there were no supermarkets or department stores. The shop proprietors were familiar with their patrons' purchasing habits, likes, and dislikes. It was nearly hard to get to know the clients and their tastes as the small business expanded into enormous franchises with hundreds of outlets across the nation. Without a proper understanding of their consumer base, many businesses struggle to meet the demands of their patrons. Prediction models are therefore required to comprehend customer preferences better.

With a focus on analysing costumers habits and preferences this research aims to answer the following questions:

- What is the relationship between factors that affects the purchases of a customer?
- Which factors affect the purchases the most?
- Can we make a reliable prediction model that will help us predict purchases ?

The structure of this study is as follows: To gain a deeper understanding of the data, a data exploration phase was carried out in conjunction with visuals. A data pre-processing method is then applied to the quantitative data. The final section shows how a prediction model was developed using a random forest. In the final section, we will compare the accuracy of our trained and tested model to other models from related projects that used the same dataset.

II. State of the Art

Black Friday has been the biggest shopping day of the year for the past 20 years. Crowds are inundating retail stores, and many items are marked down significantly. Over time, trends for this significant shopping day have emerged. Although the amount of online shopping has massively increased and has given businesses even more insights into their customers and their preferences, this study has not yet been effectively applied, and there has been comparatively little research on sales forecasting and prediction. Visual analytics is very adequate to this situation and the amount of data held can facilitate the buying and selling process for both sides of the table.

Two papers [2] [3] were analyzed to learn about gaining knowledge using visualizations integrating with human thinking on familiar collection of purchases.

Sumit Kalra [2] used some visual analytic techniques and trained four different models to compare their accuracy, the visualization section wasn't the main part of this study, the main visualized and studied attributes were the gender and age factors which can be optimized given all the different attributes provided by the Black Friday dataset. The study focused more on training and testing phase, 5 cases were conducted using different splitting strategies to compare the results and see which splitting strategy is the best fitted.

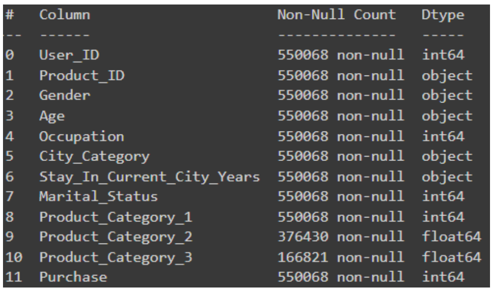
Ching-Seh Mike Wu [3] focused on the data distribution study by visualizing correlation between attributes, data cleaning was a prioritized step as the main focus of this research is to get the highest accuracy score instead of exploring the data in depth. In the pre-processing phase, the author focused on diversifying different regression and classification models for the best accuracy score. This study helped selecting the type of models to be considered for a better purchase price prediction, complex models aren't fitted for the aim and the objectives of the conducted research.

III. Properties of the Data

For the sake of this project, we used a data shared by a retail company "ABC Private Limited", this business wants to comprehend how customers behave when making purchases (more specifically, how much they spend) in relation to various goods belonging to various categories. the data contains purchase summaries of various customers for selected high-volume products in a month period. Customer demographics are also present in the data. The particular version for this study was published on Kaggle in a delimited file containing 517,401 rows of twelve categorical text columns, The purchase column is the target value we wish to predict.

The retailer will be able to create personalized offers for customers against various products with the aid of this dataset, which can now be used to train a supervised machine learning algorithm to predict the purchase number of customers against various products.

The dataset used in machine learning algorithms needs to be balanced. There should be an equal number of samples in each class; otherwise, the classification or prediction will favor the class where the data is skewed. We investigated the correlation between attributes and, more specifically, with the target value, to eliminate any possibility of imbalanced data. The figure below shows the description and the non-null count of each column, a very large number of null values is present within the product_category_2 and product_category_3 which shows that the costumer is more interested in the product_category_1 purchase. We can also notice the presence of different data types such as integers, objects and floats.



```
#   Column                      Non-Null Count   Dtype
--  ------                      --------------   -----
0   User_ID                     550068 non-null  int64
1   Product_ID                  550068 non-null  object
2   Gender                      550068 non-null  object
3   Age                         550068 non-null  object
4   Occupation                  550068 non-null  int64
5   City_Category               550068 non-null  object
6   Stay_In_Current_City_Years  550068 non-null  object
7   Marital_Status              550068 non-null  int64
8   Product_Category_1          550068 non-null  int64
9   Product_Category_2          376430 non-null  float64
10  Product_Category_3          166821 non-null  float64
11  Purchase                    550068 non-null  int64
```

Figure 1 - Summary of the columns

This data mainly shows the factors affecting the costumer purchase on the black Friday. It is mainly divided by 4 main factors, the customer level, showing the costumer financial interests of the customer. The store level, mainly shows the average pricing of products which can limit the purchasing power of the customer, and which product category is open to a broader part of the society by the pricing factor. The city level which can hugely affect the purchasing amount and budget. And finally, the product level which is divided in 3 categories in our dataset for an easier data analysis.

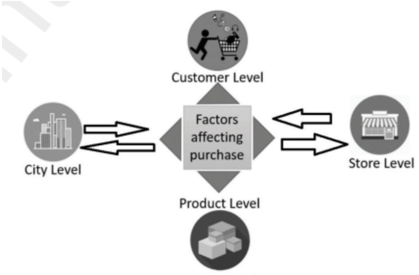
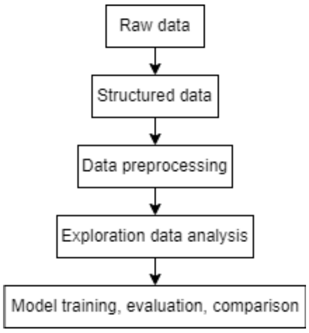


Figure 2 - Factors affecting purchases on Black Friday [4]

IV. Analysis

A. Analysis approach

The dataset we are working with uses quantitative data for exploratory analysis. Exploratory data analysis employs a visual method to analyze data sets and highlight their key features. Numerical data, such as the iris dataset and scorecard data, are involved in quantitative data. Python, pandas, matplotlib, numpy array, seaborn, and python notebook are the tools used to support analysis.


Raw data
Structured data
Data preprocessing
Exploration data analysis
Model training, evaluation, comparison

To structure the data, box plot were used as a way to check data quality and discover any type of imbalance, as they are efficient in detecting outliers, peculiar distributions and erroneous valus. Missing values were investigated by returning the 'isnull' function in Python. However, we decided to leave the data as it is knowing that it is very important to understand which category are the clients more interested in, null values show the inactivity of the costumer towards a certain product category and this information is important in order to classify products for their popularity in the black Friday market.

Another way we assessed data distribution was through the use of histograms. These visualizations helped identifying the most popular categories of each attribute and get a better look at the data before the analysing phase.

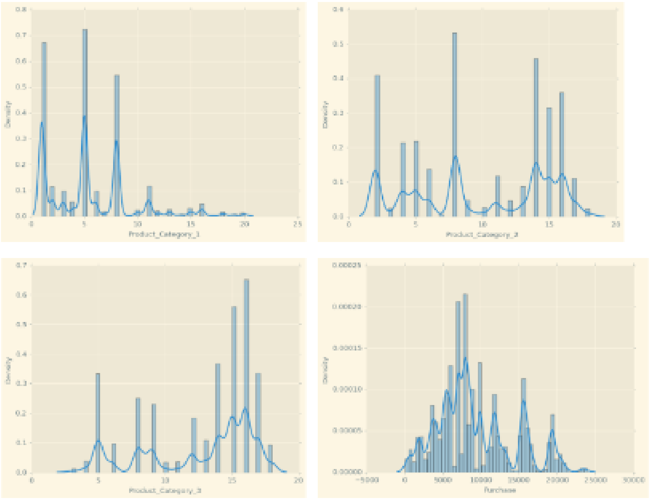


Figure 3 – histograms of product categories and purchases

B. Analysis process

After running a brief health check to the data and checked its accuracy in terms of the information it provides about the customers, the analysis phase was conducted. For the analysis section, Anaconda Navigator with Jupyter Notebooks was used. CSV files can be edited and read using the Pandas library. The data is visualized using Seaborn and matplotlib using histograms, boxplots, bar graphs, scatter plots, etc. Because data have numerical values, quantitative analysis is the type used. Using a graphical presentation, it has been successfully discovered how factors like gender, marital status, and occupation influence a customer's likelihood to make a purchase.

We started by comparing the male vs female presence in the data, and by the total purchases

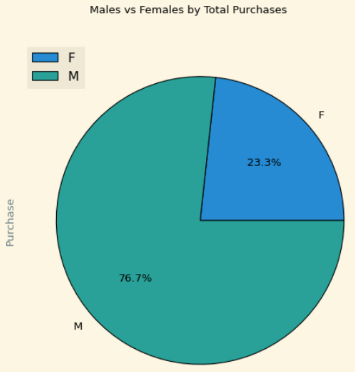


Figure 4 – Male vs female by total purchases

The result was unexpected, women tend to be more active in terms of purchases, but the male dominance is clearly visible in the graph by 52% difference rate. The male dominance shows that men are more likely to buy products during discount opportunities and they spend more money.

The occupation plays a major role in term of the buying capacity, it can permit the user to be more free in terms of choices, which can heavily affect the market, we compared the occupation categories in terms of the money spent on purchases.

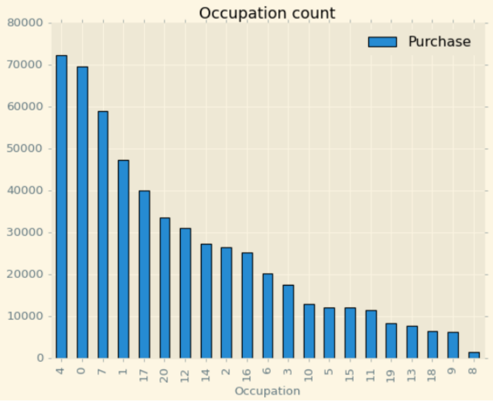


Figure 5 – Histogram of occupation count

The three most dominant occupation sectors are 4, 0 and 7. The occupation 4 has reached 70000 next to the occupation 0 and the occupation 7 reaching 60000. The occupation 8 is nearly inactive due to the lack of financial freedom.

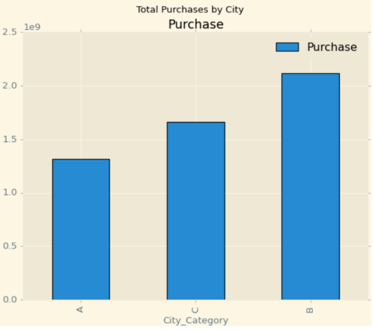


Figure 6 – Histogram of city category

The most dominant City in terms of purchases is the city B followed by the city C. This distribution has to do with the city size and population number. But the factor that need to be looked in terms of a city population is whether a citizen is new to the city or not and if this information affects the purchase activity during the black Friday.

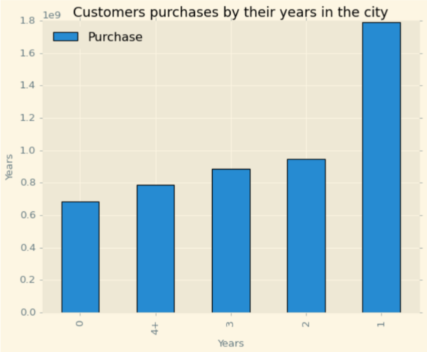


Figure 7 – Histogram of purchases by their years in the city

As we can see in the plot, this factor indeed is important to visualize as it has a repercussion on the activity. We notice that the clients who spend 1 year in a city are the most active, after this the activity is stable in terms of purchases and decreases by almost 50% in the following years.

Marital status can also affect purchases as single individuals behave differently from married couples, We used a pie chart to visualize this feature and it showed that single people are more likely to buy products and have more interest in black Friday with a 20% difference.

These graphs gave us a clear idea about the costumer’s habits and how they financially interact with the ABC retail company during the black Friday.

The age distribution plays a major role in defining which section of the costumers is the most active and which one is not, this data can help selecting niche of the products that should be focused for future sales and black Friday events.

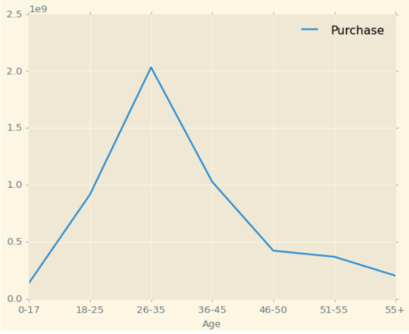


Figure 8 – Histogram of age distribution

From the graph above, we can se that the most active clients are between 26 and 35 years old. Costumers below 18 and above 55 are less active.

Now we want to focus on the products and deep into each category comparing the product popularity and compare different categories. The pie chart below shows the top 5 most popular products.

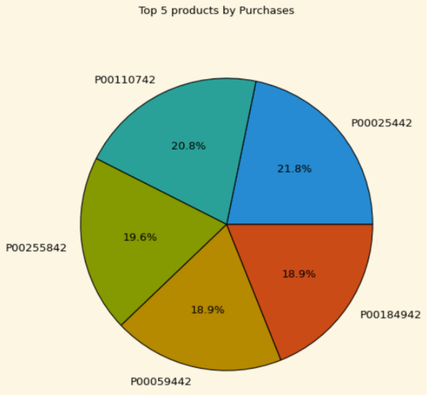


Figure 9 – Pie chart of 5 most purchased products

After visually analysed the data and made assumptions from it, we used a correlation matrix which shows the correlation between attributes, this matrix helps comparing each feature in terms of how it will affect the target value for an accurate prediction. Age, Gender and Occupation has a low impact on the Purchase amount, while product category and Product ID has a high impact on the purchases amount.

After the data exploration phase was done, the model training and testing was conducted. Research was made to select the best model to use in our dataset. We decided to use the random forest Regressor, from the work Summit Summit [2] has made using different splitting strategies, the best fitted strategy was the 70:30 as we used 70% for the training and 30% for the testing. We achieved a model accuracy of 95% on training dataset and 69% on testing dataset. We used visualization to see the compare the target feature between the true values and the predicted values.

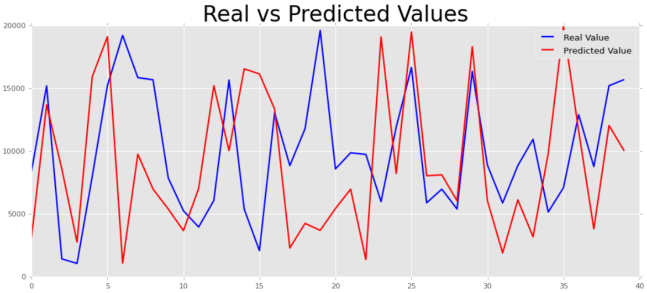


Figure 10 – Real vs predicted values comparison

C. *Analysis result*

We accurately predicted purchases using Random Forest regressor. The results were satisfying, our chosen classifier has performed decently compared to other regression models from other studies, but more improvement can be made as the model still made some errors due to the large scale of predictions, we which to obtain. The solution is to perhaps use even more data, and use less broad information to avoid confusing the prediction model.

V. Critical Reflection

We have analyzed each feature on the dataset and how strong is its relationship with the target value to see how likely it will affect the purchase amount, the results were satisfying, we surprisingly discovered that men between 26 and 35 years old are the most active buyers in the community, we also discovered that income does not define the purchase amount even though it is highly correlated to it. The conclusion is that the most affecting factors to the purchase amount are the city category feature, and the occupation. They highly affect the outcome and this was the main focus for this research

After analyzing the data using various data exploration techniques and graphs and trained a model for prediction, we see that no costumer is same as another, every client is different in terms of personal details and also preferences, the human habits are very difficult to predict as it is random and people think differently, even if we try to predict from other datasets and train an even better model, results would still be surprising and the prediction is very unlikely not be true all the way, we can see this from the outliers, some people have a high paid occupation and spend less than people with a lower income occupation. The human brain is surprisingly fascinating.

References

[1] H. Chen, R. H. Chiang and V. C. Storey, “Business intelligence and analytics: From big data to big impact,” MIS. Quarterly, vol. 36, no. 4, pp. 1165–1188, 2012.

[2] S. Kalra, B. Perumal, S. Yadav and S. J. Narayanan, "Analysing and Predicting the purchases done on the day of Black Friday," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-8, doi: 10.1109/ic-ETITE47903.2020.256.

[3] C. -S. M. Wu, P. Patil and S. Gunaseelan, "Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data," 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), 2018, pp. 16-20, doi: 10.1109/ICSESS.2018.8663760.

[4] M. H. Rehman, C. S. Liew, A. Abbas, P. P. Jayaraman, T. Y. Wah et al., “Big data reduction methods: A survey,” Data Science and Engineering, vol. 1, no. 4, pp. 265–284, 2016.