Black Friday Sale Analysis and Prediction

Tenzin Palbar
20210165
School of Computing
Dublin City University
Dublin, Ireland
tenzin.palbar2@mail.dcu.ie

Shankar Aniruddh Tejomurtula
20210689
School of Computing
Dublin City University
Dublin, Ireland
shankar.tejomurtula2@mail.dcu.ie

Shubham Verma
20210906
School of Computing
Dublin City University
Dublin, Ireland
shubham.verma3@mail.dcu.ie

Gelmis Bartulis
20213041
School of Computing
Dublin City University
Dublin, Ireland
gelmis.bartulis2@mail.dcu.ie

***Abstract*—Black Friday is one of the major holiday sale event organized by retailers right after the day of the Thanksgiving. This is considered to be a major profiting event leading analysis of consumer preferences and routines toward shopping whether it is physically in the mall or via online services a valuable attribute towards sale marketing activities. Understanding customer buying patterns can enable retailers to provide personalized discounts and target prospective customers. This study used publicly available historic dataset consisting of anonymized customer details along with their purchased item information. An exploratory data analysis was performed on this dataset to understand the attribute characteristic and purchase correlation. Data Mining processes have been applied and comparison of multiple prediction models (LR, Lasso Regression, Ridge Regression, ElasticNet regression, XGBoost, RF and DT) were performed to provide an efficient model that could forecast customer's future spending. The study used three evaluation methods (MAE. R-squared, RMSE) to assess the model performance on preprocessed data and visualized the result for better understanding.**

***Index Terms*—Prediction, Sale, Exploratory data analysis, Data Mining, Machine Learning**

I. Introduction

Black Friday Sale is one of the biggest shopping holiday event organised every once a year, observing immense consumers purchase flow [1] compared to any other ordinary day. As current retailer business acquired an enormous repository of data and the volume and variety of data is expected to grow further in an exponential manner. The availability of these data, especially on sale event like Black Friday, gives the opportunity to imply data mining techniques to obtain valuable insights from customer purchase patterns to sale estimation. With the gaining popularity of online shopping, the e-commerce industry can also make use of an intelligent prediction model for sales with the highest possible level of accuracy and reliability. Forecasting sales that happens every year can provide ways for companies to handle the resources during the sudden rise of sales and increase profit. One of the major objectives of this study is to understand and compare the sales prediction model performance implemented using data mining techniques using consumer description and purchase transaction history.

This paper is structured as follows: Section II summarizes the related work on Black Friday sale prediction. Section III described the dataset that has been used. Section IV describes the Data Mining Methodology utilized, followed by each stage as subsection. Finally, conclusions on the study by summarizing the findings are presented in Section V.

II. Related work

Data Mining and Machine Learning (ML) methods have been widely used in different fields to capture patterns and valuable insights that can be explored for individual domain benefit. In the Retail industry, customer behavior and sales patterns have been analyzed using the Data Mining process in multiple research studies and in real world applications using diverse datasets. Although data with time series framework has been widely used by retailers to analyze their quarterly or annually growth and fall trend. The data mining processes, and technology oriented to transactional-type data not having a time series framework has seen an immense growth. The process and benefits of integration of this practice for predicting methods which includes using approaches to handle multicollinearity, dimension reduction and selection on time series data has been described [2]. Wu et al. [3], presented a performance comparison of seven different ML models for customer spending prediction using the same dataset [4]. They concluded that complex models like neural network (NN) are an overkill for simple problems like regression. And simpler models along with proper data cleaning perform well for the regression, with Extreme Gradient Boosting (XGBoost) showing the best performance among seven models. However, the missing data were handed by replacing it with zero values which is not considered as good practice for efficient model development. Kalra et al. [5], also used same dataset [4] and performed prediction for different distributions of training and testing data (50:50, 70:30, 30:70) using XGBoost, tfidftransform, both combination and extra trees regressor. In the preprocessing stage, they replaced the missing values by a number 999 to avoid data redundancy which is a unique method of data handling. Duong Trung et al. [6], applied Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting and XGBoost on the same dataset with different splitting scheme (70:30, 80:20, 90:10). They concluded that the presence of noise in the data outweighed the performance of SVM and underlined the robustness of Gradient Boosting in general and XGBoost performance. All the above three mentioned papers used one evaluation metrics Root Mean Squared Error

(RMSE), however our paper will use more than one evaluation method to understand our model performance. Similar to our dataset, [7] used "Big Mart" an American based company dataset that includes item and outlet information. Models such as Linear Regression (LR), Ridge Regression, Decision Tree (DT) and XGBoost were implemented to forecast sales volume with XGBoost producing the lowest MAE and RMSE values. Cheriyan et al. [8], used dataset with attributes like Category, City, Type of items, Quantity, Quarter, Sales Revenue, Year and implemented Generalized Linear Model (GLM), DT and Gradient Boost Tree (GBT) with GBT stood out as the pioneer model. Among major relationships in Data mining techniques which are widely used in producing useful insights. Rajagopal et al. [9], performed demographic clustering technique for customer clustering using IBM Intelligent Miner to identify high-profit, high-value and low-risk customers of an organization retail smart store data. This technique is highly useful in providing further improvement in services to any specific group of customers. Association rules technique which is widely used by retailers to increase their sales is another technique of data mining. Shah et al. [10], on identifying the disadvantage of classical Apriori (candidate set generation), presented an enhanced and efficient Apriori algorithm that generates association rules that associate the usage pattern of the clients for a particular data. Gurnani et al. [11], evaluated and compared various ML models, namely, ARIMA, Auto Regressive Neural Network (ARNN), XGBoost, SVM, and Hy-brid Models like Hybrid ARIMA-ARNN, Hybrid ARIMA-XGBoost, Hybrid ARIMA-SVM and STL Decomposition (using ARIMA, Snaive, XGBoost) to forecast sales of a drug store company called Rossmann. STL Decomposition outperformed all the individual and hybrid models giving best forecasting accuracy outlining decomposition techniques being better than hybrid techniques for some applications. With the advancement of the knowledge of NN capabilities in multiple domains, comparison of traditional prediction techniques with NN has been performed to study the possible sale forecasting and insight retrieval enhancement. Chu st al. [12], studied the linear and nonlinear models in regard to sales prediction using seasonal ARIMA, regression, and feedforward neural networks and discovered that with seasonal adjustment to the data (deseasonalized time series data), NN can significantly improve forecasting performance. In this paper, we will be utilizing the data mining techniques that includes data exploration, pre-processing (feature selection, transformation, and handling missing values) and compare the models from past work against different evaluation methods.

III. Dataset

The Dataset was retrieved from kaggle (public dataset platform) [4]. After some research we came to know that the same dataset has been used for a competition hosted by Analytics Vidhya [13]. As no information about the dataset was provided by the author in the kaggle platform, we found the following description from a research paper [6] specifying the intention behind the release of the dataset. A retail company (ABC Private Limited) wants to understand the customer purchase behavior specifically the purchase amount against various products categories. They have shared a purchase summary of various customers for a selected high volume product from last month. They want to build a model to predict the purchase amount of customers against various products which will help them to create a personalized offer for customers against different products. The following table contains more details about the dataset attributes.

Variable	***Definition***
User_ID	User ID
Product_ID	Product ID
Gender	User's gender or sex
Age	Age in bins
Occupation	Occupation (Masked)
City_Category	Category of the user's city
Stay_In_Current_City_Years	User stay duration in the current city
Marital_Status	Marital status
Product_Category_1	Product category (Masked)
Product_Category_2	Product category (Masked)
Product_Category_3	Product category (Masked)
Purchase	Purchase amount (Target Variable)

Table 1. The Black Friday Dataset attributes [4] [6]

IV. Proposed Methodology

For our study, we chose CRISP-DM methodology which stands for Cross Industry Standard Process for Data Mining. It is a six-phase process model guidelines for planning, organizing, and executing data mining projects. The six-phases can be observed in the figure CRISP-DM.

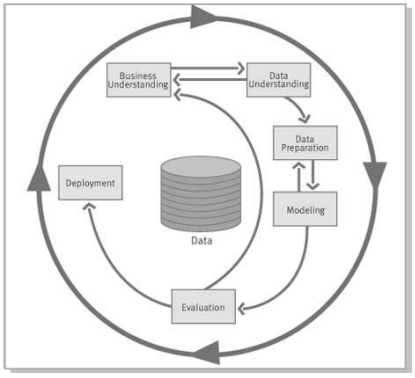

Figure 1. The Data Mining Process, according to the CRISP-DM methodology (image obtained from http://www.crisp-dm.org) [14].

A. Business Understanding

Black Friday Sale event have been widely observed not only in the US, as well as all around the world. These kinds of event now last for weeks to increase their sale in both physical and online stores. As the customer range and spending increases on every occurrence of these events, the retailer tends to adopt different methods to understand customer purchase preferences and patterns. Data mining techniques have been widely used in the retail domain. However, as consumer spending and retail sale amount differs significantly between normal days and

Black Friday Sale event due to the price discount, a Black Friday prediction model can be used to better understand consumer expectations by using consumer details and their purchase patterns.

B. Data Understanding

The Black Friday dataset consists of twelve attributes, out of which only one attribute (Purchase) is continuous and all the other remaining attributes have categorical value except Age. As the dataset consists of consumer information and product description, unique identification is assigned to each consumer and product categories. Out of the total 550068 rows of transaction, there were 5891 unique User ID and 3631 unique Product ID. It was also observed that two attribute columns Product_Category_2 and Product_Category_3 have total missing rows of 32% and 70% respectively. We removed Product_Category_3 because of the high percentage of missing values. After analysing the missing value distribution with respect to other attributes, we found that the missing values have no dependency on other attributes or to itself, hence we concluded that it is missing completely at random (MCAR). The Purchase attribute statistical characteristics (Min: 12, Max: 23961, Mean: 9264, Std: 5023) illustrate considerable difference in min and max range. We used the boxplot method to visualise the data distribution and notice few point values outside the maximum (Largest values within 1.5 times interquartile range (IQR) above 75th percentile). It could be considered as outliers, data points which differ significantly from other observations. However, the total count of values was more than a thousand and beside the purchase variable can range from low as 12 to high as 23961, we concluded that there is no outlier in this attribute. Since that dataset mainly consists of nominal variables, we used a Chi-Squared test to determine the existence of any relationship between variables. With p-value less than significance level of 0.05 for all variables, we concluded that there is a high association between the variables.

C. Data Preparation

For missing value treatment, we compared different methods for missing value handling, and it's effectiveness on the model performance. Initially, we check if there is any pattern in the missing and non-missing row value distribution of the two product category attributes to the distribution of other attributes of the overall dataset and find similarity. Thus, the distribution of the missing values would probably be the same as the distribution of the non-missing values of the same column. Since the missing data mechanism is MCAR, we imputed it with the distribution it follows to maintain the overall distribution using the imputing technique of probability distribution of the non-missing data applied to the missing data. However, as the attribute value may actually not have values at all, we use the basic method of imputation which is replacing missing values with zero and median value, which is actually not the best practice and compared model performance with all three transformations. Surprisingly, there was no significant difference in the model performance. We convert some attributes into numeric values such as Age in bin replaced with median value, Stay_In_Current_City_Years value 4+ to 4, Gender values (F, M) to (0, 1), City_Category values (A, B, C) to (0, 1, 2). As our model cannot handle categorical data, we one-hot encoded all categorical attributes. We performed the Ordinary least squares (OLS) regression statistical method to estimates the relationship between all dummy encoded independent variables to the dependent variable and observed few variables with no significance $(P > |t|) > 0.05$ to the result. If we remove attributes with no significance and both Alaik Information Criterion (AIC) and Bayesian Information Criterion (BIC) decreases, this signifies that the quality of the result has improved by removal of the attributes. However, for our dataset, removal of the insignificant attribute did not decrease the AIC and BIC value thus displaying no improvement of the result quality.

D. Modeling

After data preparation, the purpose of this research is to understand the customer characteristic and predict purchase during the Black Friday public holiday. In this case, modelling aids our research by analysing the gathered data using a data mining algorithm. This then allows us to see, which model would be the most applicable for the dataset and would suit the best in order to get the best results possible. Therefore, the analysis is based upon several different models that have outlined the best possible scenario for predicting sales. These models helped to analyse and manage the data correctly and allowed further analysis which has been conducted using several different parameters and characterising them. Once the models have been developed, they were used to predict the sales in regard to the product category. There were different models used to create a better understanding of data and the prediction capabilities. The models that have been used were LR, XGBoost, Random Forest Regression (RFR) and Decision Tree Regression (DTR). These models in the latter stage have been evaluated by error differences and compared to each other. The error values are mean values for predicting errors, such as: Mean Absolute Error (MAE), Mean Squared Error (MSE), R-Squared (R^2), and Root Mean Squared Error. These error modelling techniques allow us to define which of the models would be mostly accurate at predicting the sales, by product. After evaluating which of the models would be the fittest for the dataset, then it is possible to implement more features regarding that particular model and extract a better amount of information because the model would be adjusted to the dataset more.

- Linear Regression
  Linear Regression (LR) allows to accurately predict the values that would be essential within this research, by comparing two different (predictor and criterion) variables. LR can be represented as the following notation, in which the two compared variables have been used as

two different product categories. Hyperparameter optimization allowed us to choose the variables that were used for the prediction. The chosen variables were all of the features except Product_ID and User_ID. When the data is trained and tested using that model, there is a possibility to compare that model to others using different evaluation metrics: Mean Absolute Error (MAE) R-squared (RS), these metric parameters can be seen in the table below.

$$Y = a + bX$$

- XGBoost
  XGBoost (XGB) is short for eXtreme Gradient Boosting package. The package contains a linear model and a tree learning algorithm. The model belongs to the Ensemble learning section, which is composed of multiple classification algorithms that provide better prediction accuracy. This model has grown in popularity quickly since it has been developed for the following reasons: it can handle missing variables, partially removes outliers and quick and efficient processing. Within this analysis, the research has been done using LR with gradient boost. In order to retrieve the best results possible while using this model, hyperparameter optimization has been initialised and the model has been fitted with the 'Product_Category_2' parameter.

$$J_m(\phi_m) = \sum_{i=1}^{n} L(y_i, \hat{f}^{(m-1)}(x_i) + \phi_m(x_i)).$$

- Decision Tree Regressor
  This model is focused on eroding a large dataset into more manageable and smaller subsets. Decision Tree Regressor (DTR) is a tree-like structured model that creates multiple subsets from the dataset and creates a prediction based on several different input [15]. This is one of the strongest data mining tools because of it's wide properties.

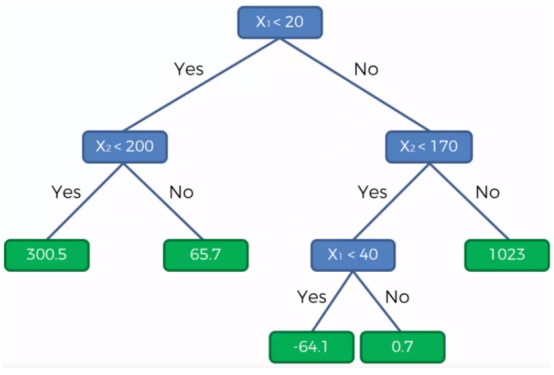

Figure 2. An example of a Decision Tree in progress [17].

The model allows for fast variable processing and distribution, although the model is memory and resource wasteful. Below Figure 4 is an example of a DTR in progress. The node displays the way the actions have been chosen. In this analysis, the regressor has been fitted with the dataset obtained and has been evaluated.
Below Figure 2 is an example of a DTR in progress. The node displays the way the actions have been chosen. In this analysis, the regressor has been fitted with the dataset obtained and has been evaluated.

- Random Forest Regressor
  Within this research Random Forest Regressor (RFR) was also used in order to obtain an accurate comparison with other models. This model belongs to the supervised learning class. It creates a forest with multiple Decision Trees, where each tree contains a decision node that then decides which tree will develop into a branch up until the end of the algorithm's lifecycle. The model has been proposed by Breiman et al. [16] and has been claimed to be one of the most effective models for predictive analysis. In this research this model has been compared with other models in order to outline the best model for this particular prediction.

$$RFfi_i = \frac{\sum_j normfi_{ij}}{\sum_{j \in all\ features, k \in all\ trees} normfi_{jk}}$$

E. Results and Analysis

After implementing LR, Lasso Regression, Ridge Regression, ElasticNet Regression, XGBoost, DTR and RFR models based on product category hyperparameter, below are the evaluation results. We have evaluated each model based on their performance with a specific variable in order to see which one of them would perform the best and be the fittest in regard to the dataset within the research.

These techniques have guided us to choose the fittest model possible by looking at the MAE, MSE and R^2 score test results. They have shown, which model has the largest overall error in relation to the dataset and allowed for choosing the correct model better. The models that have been evaluated beside each other, have been Lasso regression, Ridge regression, ElasticNet regression, Linear regression, Random Forest regression, XGBoost and Decision Tree regression. By implementing and evaluating each of these models we have been able to see which one of them suits the dataset the best and which one would provide the most information, in order for us to form a hypothesis.

Below are the results of our implementation of the models that have been evaluated accordingly with every error recognition technique. These values indicate whether the model would not be suitable, more suitable or be the most suitable for our dataset. From the table below we can observe that the Linear Regression has performed the best in regard to the overall errors and the errors that have been recognised. Therefore, in more detail:

- Mean Absolute Error (MAE) allows us to see accuracy within continuous variables and displays the overall error within the analysis
- R-Squared (R^2) gives us a notion of how fit the model is in relation to the correlated variables.
- Root Mean Square Error (RMSE) is the standard deviation of predicted errors. It allows us to see how close the model prediction is to the line of best fit.

We also analyzed the omnibus, skew and kurtosis of the errors, to test if they are normally distributed and what are the chances of errors being normally distributed.

Model	Evaluation		
	MAE	***R-squared***	***RMSE***
Lasso Regression	2281.29	0.64	3013.4
Ridge Regression	2281.21	0.64	3013.4
Elastic Net Regression	2435.37	0.60	3137.8
Linear Regression	**2266.7**	**0.64**	**3000.0**
XGBoost	2398.03	0.62	3088.1
Random Forest Regression	2270.9	0.61	3114.9
Decision Tree Regression	2379.2	0.56	3332.2

Table 2. Model Evaluation.

Since all the models have been evaluated with the same evaluation methods, they helped us see which model suits the dataset more. In case of this research, Linear regression has shows to be the most fitting. The model prepared with hyperparameter optimization, which helped to reduce the loss of the variables, loss of data and allowed better fitting.

We have also converted the variables into numeric values because all the machine learning algorithms work using numeric values. This helped for further processing of the dataset. Therefore, by looking at all the models that we have tested and implemented, our analysis shows that the Linear Regression model is providing the best prediction results for train and testing data.

V. Conclusions

In our study, we have used 550,069 purchase records for the comparison of algorithms. Though the available attributes used in this analysis were sufficient for this analysis. However, attaining a dataset with greater variety of attributes could benefit sales prediction at a higher level. Missing data handling was the one of the major challenges we faced during the research as most of the previous work done with the same dataset had used approaches that could affect the data quality. In this paper, we performed statistical attributes independence analysis, variable significant estimation, used different methods for handling missing data and compared model performance with respect to each preprocessed output attribute. We observed that there was not much variation in model performance. Out of all prediction model, Linear Regression gave the MAE value of 2276.2, showcasing better performance compared to other models in this study.

As the dataset used to construct current predicting models includes many inconsistencies, masked, and missing values. In the future, using a dataset with more attributes will provide better scope for sale understanding and assist in better performing model development. Consideration of the removal of multicollinearity between independent variables by using appropriate analysis methods can enhance model performance. As the result of our finding is focused on retailer profit, using good visualization techniques and providing easy to understand model outcomes is very important for retail owner understanding. The study analysis notebook and presentation video can be found in here [18].

Acknowledgment

We would like to thank Dr. Andrew McCarren (School of Computing, Dublin City University) for teaching us the required knowledge of Data Mining and Data Analysis that we applied on this study. We would also like to take this opportunity to thank our team members for their cooperation.

References

[1] R. Laycock and C. Choi, "Black Friday statistics 2020," finder, Oct. 2020. Accessed on: April. 16, 2021. [Online]. Available: https://www.finder.com/black-friday-statistics.

[2] T. D. Rey, C. Wells, and J. Kauhl, "Using data mining in forecasting problems," In SAS Global Forum 2013, Data Mining and Text Analytics, 2013.

[3] C. 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), Beijing, China, 2018, pp. 16-20, doi: 10.1109/ICSESS.2018.8663760.

[4] StefanDolezel, "Black Friday," 2018. Accessed on: April. 16, 2021. [Online]. Available: https://www.kaggle.com/sdolezel/black-friday.

[5] 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), Vellore, India, 2020, pp. 1-8, doi: 10.1109/ic-ETITE47903.2020.256.

[6] N. Duong Trung, D. Tan, T.-D. Luu, and H. Huynh, "Black Friday Sale Prediction via Extreme Gradient Boosted Trees," AIR-Fundamental And Applied IT Research Conference, Hu, 2019, doi: 10.15625/vap.2019.0007.

[7] G. Behera and N. Nain, "A Comparative Study of Big Mart Sales Prediction," in Computer Vision and Image Processing, Singapore, 2020, pp. 421–432, doi: 10.1007/978-981-15-4015-8_37.

[8] S. Cheriyan, S. Ibrahim, S. Mohanan and S. Treesa, "Intelligent Sales Prediction Using Machine Learning Techniques," 2018 International Conference on Computing, Electronics Communications Engineering (iCCECE), Southend, UK, 2018, pp. 53-58, doi: 10.1109/iCCECOME.2018.8659115.

[9] S. Rajagopal, "Customer Data Clustering Using Data Mining Technique," International Journal of Database Management Systems (IJDMS) Vol. 3, No. 4, November 2011, doi: 10.5121/ijdms.2011.3401.

[10] N. Shah, M. Solanki, A. Tambe and D. Dhangar, "Sales Prediction Using Effective Mining Techniques," International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 6, No. 3, 2015, 2287-2289.

[11] M. Gurnani, Y. Korke, P. Shah, S. Udmale, V. Sambhe and S. Bhirud, "Forecasting of sales by using fusion of machine learning techniques," 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), Pune, India, 2017, pp. 93-101, doi: 10.1109/ICDMAI.2017.8073492.

[12] C. Chu, G. P. Zhang, "A comparative study of linear and nonlinear models for aggregate retail sales forecasting," International Journal of Production Economics, vol. 86, Issue 3, 2003, pp: 217-231, ISSN 0925-5273, doi: 10.1016/S0925-5273(03)00068-9. [13] Practice Problem, "Black Friday Sales Prediction," April 2021. Accessed on: April. 16, 2021. [Online]. Available: https://datahack.analyticsvidhya.com/contest/black-friday.

[13] Practice Problem, "Black Friday Sales Prediction," April 2021. Accessed on: April. 16, 2021. [Online]. Available: https://datahack.analyticsvidhya.com/contest/black-friday.

[14] C. Soares, Y. Peng, J. Meng, T.i Washio, Z. Zhou, “Applications of Data Mining in E-Business Finance: Introduction,” Applications of Data Mining in E-Business and Finance, June 2008, Vol. 177, pp. 1-9, doi: 10.3233/978-1-58603-890-8-1.
[15] C. Lazăr, and M. Lazăr, “Using the Method of Decision Trees in the Forecasting Activity,” Petroleum-Gas University of Ploiesti Bulletin, Technical Series, 2015, vol. 4, pp. 41-48.
[16] L. Breiman, “Random forests,” Machine learning 45, 2001, pp. 5-32, doi: 10.1023/A:1010933404324.
[17] S. Girgin, “Decision Tree Regression in 6 Steps with Python,” Medium, 2021. Accessed 12 April 2021. [online]. Available: https://medium.com/pursuitnotes/decision-tree-regression-in-6-steps-with-python-1a1c5aa2ee16.
[18] T. Palbar, S. A. Tejomurtula, S. Verma and G. Bartulis, “Black_Friday_Sale_EDA,” GitHub repository, 2021, [Source Code]. Available: https://github.com/aniruddh1804/Black-Friday-Sales-Analysis-and-Prediction.