Black Friday Sales Prediction

Arun Reddy Nalla
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
nalla2@uwindsor.ca

Darun Arumugham
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
arumugh@uwindsor.ca

Siddhartha Pitchika
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
pitchik@uwindsor.ca

Varshitha Seralathan
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
seralatv@uwindsor.ca

Abstract— **Black Friday kicks off the Christmas shopping season in the United States. On Black Friday, major retailers like Amazon, Walmart, and others attract buyers by providing deals and promotions across a variety of product categories. Electronics, clothing, kitchen appliances, and decor are among the product categories. Various researchers have conducted sales forecasting studies. This data analysis is used to deliver discounts on various products and goods. We employed four different models in order to analyze and forecast sales. The dataset Black Friday Sales Dataset, which is accessible on Kaggle, was utilized for analysis and prediction. Linear regression, Decision Tree Model, Random Forest Regressor, and XG Boost are the prediction models employed. The Root Mean Squared Error (RMSE) is a performance assessment metric. The XG Boost outperforms the other models by having the lowest RMSE score.**

Keywords - Regression, Linear Regression, Decision Tree Regressor, Random Forest Regressor, XGBoost Regressor, Root Mean Squared Error, Data Analysis

I. INTRODUCTION

"Black Friday Sales" is one of the biggest sales that happens yearly in North America. The day after Thanksgiving has been considered the beginning of the United States Christmas shopping season since 1952, although the term "Black Friday" did not become widely used until recently. Many stores offer huge sales on Black Friday and open very early, such as at midnight, or even start their sales on Thanksgiving. The major challenge for a Retail store or eCommerce business is to choose product prices such that they get maximum profit at the end of the sales.[1]

Retail business store a summary of the purchases made by several clients last month for a few high-volume products. Along with consumer demographics (age, gender, marital status, city type, and length of stay in the present city), product specifics (product id, product category), and total purchase amount from the previous month are also included in the data collection.[2]

Despite the data available, it is difficult for the business to predict the items that would be in demand and create appropriate offers.

II. MOTIVATION

The main motivation of this project is to bridge the gap between business profits, product demand, and offers. Due to the high demand for offers and discounts, businesses in an attempt to still maximize their profit sometimes manufacture poor-quality products. Products of poor quality are manufactured in large quantities and more often than not lead to large amounts of waste. By forecasting the purchases expected on black Friday for a particular business, we can solve this problem and help businesses to target their demographics, tailor offers, and maximize profits.

III. RELATED WORK

An extensive amount of research is done utilizing a variety of approaches to analyze and predict sales. Many approaches have been put forth by numerous researchers to do this. We will summarize a few machine-learning techniques in this section.

Different Machine learning techniques have been used by Ramasubbareddy S. et al. [3] to anticipate sales. The Black Friday Sales Dataset is the dataset used for experimentation that was taken from Kaggle. The system is implemented using the following algorithms: Rule-Based Decision Tree, Decision Tree, Random Forest, Ridge Regression, and Linear Regression. As a performance evaluation metric, Root Mean Squared Error is used. According to RMSE, the model is better when the RMSE is lower. Therefore, with an RMSE rate of 2291, Rule-Based DT beats other machine-learning algorithms.

A prediction model has been put forth by C. M. Wu et al. [4] to examine past consumer spending patterns and forecasts future spending patterns. Black Friday Sales Dataset from analyticsvidhya is the dataset that was referenced. They have used different machine learning models namely, XGBoost, Linear Regression,

MLK classifier, Decision Tree, Decision Tree with bagging, and Deep Learning model using Keras. The models utilized are assessed using the performance evaluation metric Root Mean Squared Error (RMSE). Instead of using complicated neural network models, simple models like linear regression can be used to address simple problems like regression.

A sales prediction model has been put forth by Odegua and Rising [5]. K-Nearest Neighbor, Random Forest, and Gradient Boosting are the machine learning models that were chosen for implementation. Data Science Nigeria provided the dataset for experimentation as part of machine learning competitions. Mean Absolute Error is the performance evaluation metric that is utilized (MAE). With an MAE rate of 0.409178, Random Forest surpassed the other algorithms.

IV. PROPOSED MODEL

A Kaggle dataset [6] is used to represent the data stored by a business. This sample dataset is stored in a MongoDB database. This stored data is then loaded into Jupiter notebook by connecting with MongoDB.

Data exploration is performed to understand the provided dataset. Based on the exploration, data preprocessing is performed to identify the features to be used, remove any null values, normalize the data, and perform any encoding if required.

Finally, data is split into training and test data for accuracy. Data is further explored to identify various patterns in the dataset that provides some insights into the influence of some factors on the outcome. Multiple regression models are trained on the training data set. These trained models are evaluated using the test dataset obtained in the previous steps.

The accuracy of the models built is evaluated using the Root Mean Square Error (RMSE) scores. The best-performing model is used to forecast purchase behavior based on customer details. This would help to create personalized offers and maximize profits for the business.
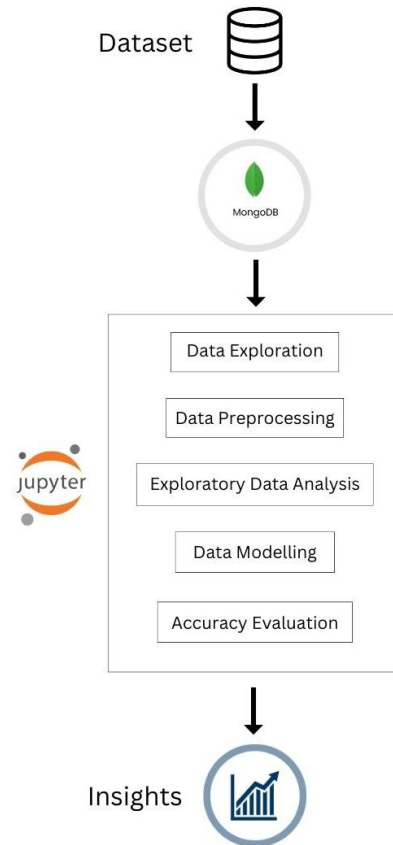

Fig 1: Proposed System Model

V. IMPLEMENTATION

A. Data Exploration and Analysis

The dataset is explored to find interesting relationships and hidden insights using the various attributes.

A boxplot and violin plot representation of the purchases attribute shows the price range that the customers typically spend more in. It shows that the customers are willing to purchase items mostly of high value from the business. Based on the Black Friday Sales dataset [6], the purchase distribution obtained is shown in Fig 2.
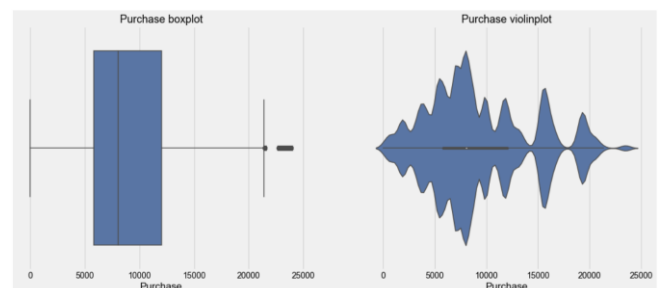

Fig 2: Purchase Distribution

To understand some of the features of the business' target demographics, the gender attribute was analyzed. It became very evident that male customers are the

majority of the customers having a ratio of 75 to 25 percent of females. Even though the purchase amount falls within the same range, male customers have a clear majority in the number of purchases. This analysis shows that both the products and campaigns should focus more on male customers.
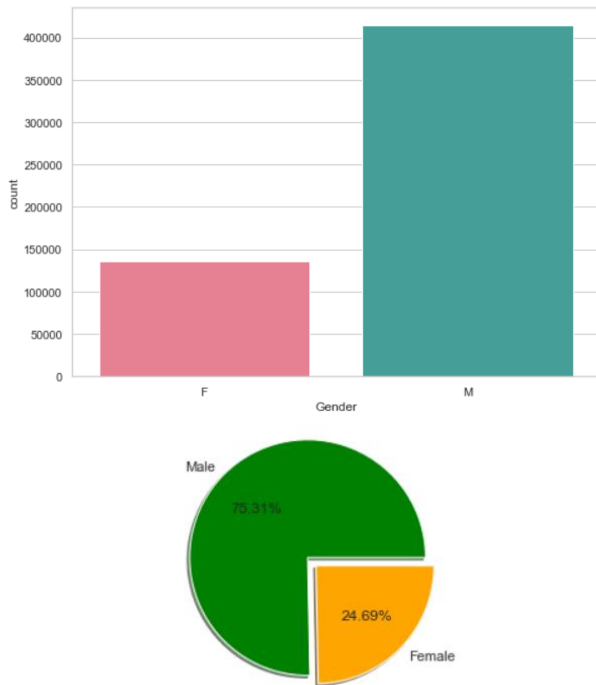

Fig 3: Gender Analysis

On analyzing the marital status of the purchases, initially, it shows that unmarried customers make more purchases. On further analysis, it can be seen that both unmarried and married customers spend the same amount of money on average. So, the business isn't affected much by this attribute.
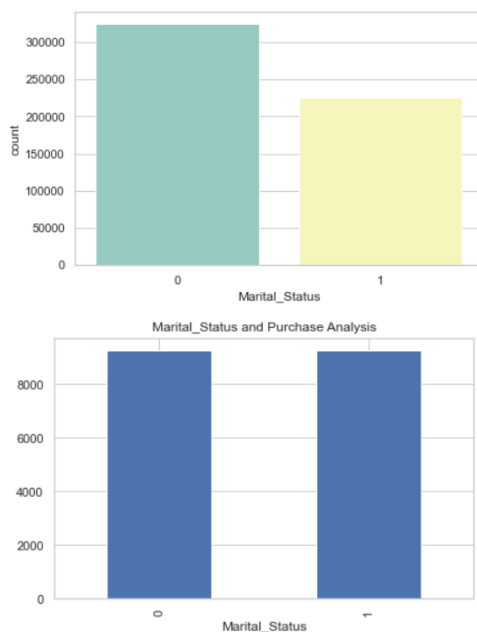

Fig 4: Marital Status Analysis

Other than the personal characteristics of the customer, other factors also provide some insights into the purchase patterns. One such factor is the customer's occupation. Customers from certain occupations have shown to be purchasing more in the black Friday sales than others. These occupations are represented by categories 0 to 20. Businesses can use this insight to produce more products related to their best-selling occupations or take more steps to expand to other occupations.
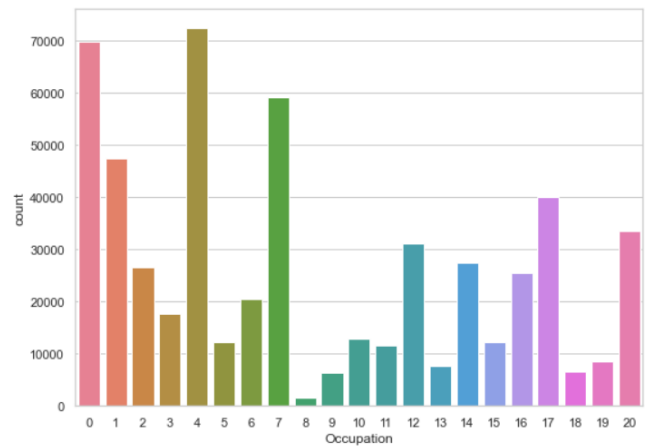

Fig 5: Occupation Analysis

Businesses also have influences from other factors such as the location of the store or the people living in certain places having a higher purchase trend than others. On analyzing the location of customers, it initially seemed like city B had the highest number of purchases. However, the average purchase amount spent is the maximum for city C. The business can further analyze what causes the difference in their different locations and how to maximize their sales in all cities.

Fig 6: City Analysis

Another personal factor that identifies the target demographic for many businesses is the age of their customers. For the considered sample dataset [6], the customers in the age range 26 to 35 make the greatest number of purchases by a large difference. When it comes to the amount spent on purchases, the age category of 51 to 55 spends the highest average and all the other adults spend a similar average value. The target customers are adults in the business, and it should look to maximize the purchases made by these categories.
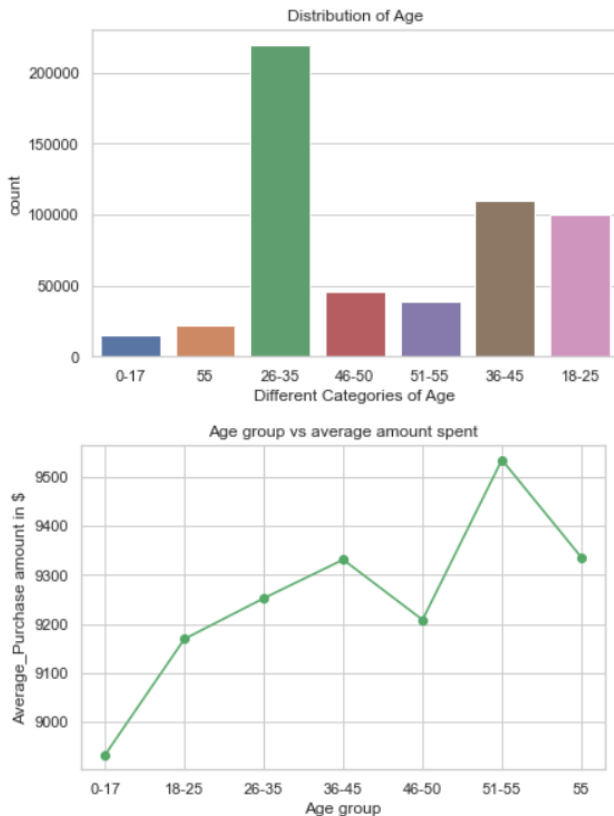

Fig 7: Age Analysis

B. Building the model

The data mining technique used for forecasting the purchase pattern on Black Friday Sales is Regression. Multiple models are trained and validated against the dataset. The models are evaluated based on Root Mean Square Error (RMSE) scores. The comparison of all the algorithms is depicted in Table 1 below.

Model	RMSE
Linear Regression	4689.610487166007
Decision Tree Model	2739.0579311299994
Random Forest Regressor	2794.786750228736
XGBoost Regressor	2520.2800251681333

Table 1: Comparison of algorithms

A range of algorithms was used in this project, including linear regression, decision tree regression, random forest, and XGB regression, in order to make the most accurate prediction possible. In this problem, the hyperparameter-tuned XGB regressor gives us the best RMSE value and r2 score.

VI. RESULT

In cases of complex data such as the Black Friday Sales data, traditional methods of analyzing don't help the business grow in revenue. Data mining techniques prove to be pivotal to the business plan considering the shopping patterns of customers.

The regression model built can now be used to make predictions and forecast the sales providing information such as the number of products that would be in demand, which products would be in higher demand, how to target their customers, how to predict the price and gain maximum profit while maintaining quality.

VII. LIMITATIONS AND CHALLENGES

There is no actual real-time data to make predictions based on the continuous input of data as it would be proprietary. Also, analysis based on more specific components such as offline vs online purchases, the actual effects of certain marketing campaigns, etc. would provide more insights. Sales forecasting is restricted to regression models since the prediction is to be made for a continuous value of purchases. Other models and more specific attribute tuning can help improve the predictions.

VIII. CONCLUSION AND FUTURE SCOPE

With the help of the analysis, mining and forecasts created, it can be used as a general framework to predict customer purchases on Black Friday Sales. It provides the business insights into the customer choice resulting in customer-centric choices while increasing profit to the business and resulting in reduced waste.

For future research, we can perform parameter tuning and apply different algorithms and mining techniques. An application can be built that continuously makes predictions based on the continuous flow of data from point of sales applications and provides a dashboard automatically for business planning.

IX. REFERENCES

[1] Aher, Amruta & Rajeswari, K & Vispute, Sushma. (2021). Data Analysis and Price Prediction of Black Friday Sales using Machine Learning Techniques.
[2] Black Friday sales prediction. Analytics Vidhya. (n.d.). Retrieved September 30, 2022, from https://datahack.analyticsvidhya.com/contest/black-friday/#About

[3] Sales Analysis on Back Friday Using Machine Learning Techniques by Ramasubbareddy S., Srinivas T.A.S., Govinda K., and Swetha E. 2021. vol. 1171, Singapore's Springer. https://doi.org/10.1007/978-981-15-5400-1_32

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

[5] Applied Machine Learning for supermarket sales prediction. (n.d.). Retrieved from https://www.researchgate.net/publication/338681895_Applied_Machine_Learning_for_Supermarket_Sales_Prediction

[6] Kaggle. (2020, May 17). Black Friday sales prediction(analytics vidhya). Kaggle. Retrieved from https://www.kaggle.com/code/vishnu691999/black-friday-sales-prediction-analytics-vidhya/data