Black Friday Sales Prediction Analysis

Sathvik Vadlapatla
dept of mathematics
Stevens Institute of Technology
Hoboken, NJ
svadlapa1@stevens.edu

Anwar Basha Dudekula
dept of mathematics
Stevens Institute of Technology
Hoboken, NJ
adudekula@stevens.edu

Sai Ritheesh Sudham
dept of mathematics
Stevens Institute of Technology)
Hoboken, NJ
ssudham@stevens.edu

***Abstract*—This research investigates the application of machine learning (ML) techniques to predict customer spending during Black Friday sales, using a dataset from Kaggle encompassing over 550,000 retail transactions. The study focuses on preprocessing this data, including handling missing values and encoding categorical variables, followed by the application of Linear Regression as a preliminary model. Early results highlight the potential of ML in retail analytics but also suggest the need for more complex models to enhance prediction accuracy. The ongoing research aims to further explore advanced algorithms like Decision Trees and Random Forest, contributing insights into consumer behavior and retail strategy optimization during major sales events.**

I. Introduction

a) Context: Black Friday, renowned as a pivotal event in the retail calendar, marks a period where consumer spending peaks. The nature of this event, characterized by significant discounts and a diverse range of products, makes it an ideal case study for analyzing consumer purchasing patterns. Understanding these patterns not only offers immediate benefits for sales strategies but also contributes to long-term customer relationship management.

b) Problem Statement: The primary challenge in this domain is the accurate prediction of the amount a customer will spend during Black Friday sales. Such predictions are complicated by the diverse and rapidly changing consumer preferences, which are influenced by a myriad of factors including age, gender, occupation, and marital status.

c) Objective: Our objective is to leverage machine learning algorithms to predict customer spending during Black Friday sales. By doing so, we aim to provide retailers with tools for more effective stock management, tailored marketing strategies, and improved customer understanding.

d) Paper Structure: This paper is organized as follows: The next section reviews related work in the domain of consumer behavior prediction. This is followed by a detailed description of our methodology, including dataset characteristics and the machine learning algorithms employed. We then present a comparative analysis of the models, discuss future research directions, and conclude with the implications of our findings.

II. Related Work

Consumer behavior prediction has been a subject of interest in both academic and commercial research, particularly within the retail sector. Previous studies have primarily focused on general shopping trends and standard purchasing behavior, employing traditional statistical and machine learning approaches. However, these studies often lack the specificity required for event-based retail scenarios like Black Friday.

Moreover, existing research has limitations in handling the high-dimensional and heterogeneous nature of retail data, especially during peak sales periods. Traditional models often fall short in capturing the complex interactions between various customer attributes and their purchasing decisions. This gap highlights the need for more sophisticated and tailored predictive models, specifically designed for high-stake retail events.

III. Our Solution

This section elaborates your solution to the problem.

A. Description of Dataset

The dataset employed in our study was sourced from Kaggle's Black Friday Sales dataset. It comprises 550,069 transactions from a retail store, capturing a wide array of customer attributes such as User_ID, Product_ID, Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, Marital_Status, and Product Categories. This dataset presents a unique opportunity to analyze a large, diverse set of transactional data, which is ideal for applying and evaluating machine learning models.

The richness of this dataset lies in its detailed representation of customer demographics and purchasing behaviors. The varied attributes allow for an in-depth analysis of how different factors influence spending during Black Friday sales. This, in turn, offers a comprehensive view of consumer behavior, which is crucial for accurate prediction modeling.

B. Machine Learning Algorithms

In our approach, we selected a range of regression algorithms known for their effectiveness in predictive analytics. Linear Regression, a fundamental technique, was used as a baseline for performance comparison. Decision Trees provided a more granular analysis of how individual attributes affect purchase amounts. Random Forest and Extra Trees, both ensemble methods, were chosen for their ability to handle large datasets and reduce overfitting, thus improving prediction accuracy.

Each algorithm was carefully tuned and evaluated to ensure optimal performance. The choice of these specific models was driven by their diverse characteristics, allowing us to capture different aspects of consumer behavior and understand the strengths and limitations of various predictive approaches.

a) Linear Regression: Linear Regression is a foundational algorithm in predictive analytics and serves as a starting point for many regression problems. It works on the principle of fitting a linear equation to observed data. In the context of predicting Black Friday sales, Linear Regression provides a baseline model by establishing a straightforward relationship between the dependent variable (purchase amount) and independent variables (customer attributes like age, gender, etc.).

1) Simplicity and Interpretability: One of the main advantages of Linear Regression is its simplicity and ease of interpretation. The coefficients of the model directly represent the impact of each predictor variable on the target variable, making it easy to understand and explain.
2) Application in the Project: In our project, Linear Regression was applied as the initial model. This helped in identifying which variables had the most significant impact on the purchase amount and set the stage for more complex models.

b) Decision Tree: A Decision Tree is a non-linear model that splits the data into branches to form a tree-like structure, based on decision rules inferred from the data. It is particularly useful for capturing non-linear relationships between variables.

1) Handling Non-linearity: Unlike Linear Regression, Decision Trees can model complex, non-linear relationships. This makes them suitable for the Black Friday dataset, where the relationship between customer attributes and spending might not be linear.
2) Feature Importance and Interpretability: Decision Trees are also interpretable, as they provide clear insights into which features are most important in predicting the target variable. In our analysis, the Decision Tree helped in understanding the hierarchy and interaction of different customer attributes.

c) Random Forest: Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training. It improves model accuracy by averaging or ‘voting’ the results of different decision trees, reducing the risk of overfitting which is a common problem with single decision trees.

1) Robustness and Accuracy: Random Forest is known for its robustness and high accuracy, making it an excellent choice for complex datasets like the one used in our Black Friday analysis.
2) Application in the Project: In the project, Random Forest was used to capture more complex patterns in the data. The ensemble approach allowed for a more nuanced understanding of customer spending behaviors.

d) Extra Trees: Extra Trees or Extremely Randomized Trees, like Random Forest, is an ensemble of decision trees. It differs in the way it splits nodes and selects features, offering more randomness in the construction of individual trees.

1) Handling Variability and Overfitting: Extra Trees is effective in handling variability in data and reducing overfitting, which is crucial for a dataset with many features like ours.
2) Performance in the Project: In our project, Extra Trees was employed to see if increased randomness in the model construction would lead to better performance compared to Random Forest, especially in terms of generalization to unseen data.
3) Comparative Analysis: Comparing the performance of Extra Trees and Random Forest provided insights into the trade-offs between bias and variance in the modeling process, enriching our understanding of different ensemble techniques.

C. Implementation Details

a) Initial Setup and Data Loading: The implementation began with importing essential libraries. `Pandas` was used for data manipulation and analysis, offering data structures and operations for manipulating numerical tables and time series. `NumPy` provided support for large, multi-dimensional arrays and matrices, along with a collection of high-level mathematical functions to operate on these arrays. Visualization libraries, `Seaborn` and `Matplotlib.pyplot`, were utilized for their extensive plotting capabilities, essential for data exploration. The `sklearn.preprocessing` module’s `LabelEncoder` was included for encoding label values.

Warnings were suppressed to ensure a cleaner output display, which is often preferred for readability and presentation, especially when warning messages do not indicate critical issues.

The data, presumably containing details of transactions from Black Friday sales, was read into a DataFrame named `df` from a CSV file, ”695train.csv”. The initial few rows of the DataFrame were displayed using `df.head()`, providing a quick look at the dataset structure and contents.

b) Data Exploration: The next step was to explore the dataset to understand its characteristics better. This included:

- Generating statistical summaries using `df.describe()` to gain insights into the numerical attributes, such as mean values, standard deviations, and ranges.
- Checking the data types of each column using `df.info()` to identify columns that might need type conversion or special handling.

Visual exploration of the data was conducted using count plots for various categorical variables like Gender, Age, Marital Status, Occupation, Product Categories, City Category, and Stay Duration in Current City. This visual analysis was crucial for understanding the distribution and frequency of different categories, which could influence the model’s feature selection.

Bivariate analysis was performed to explore the relationship between different independent variables (like Occupation,

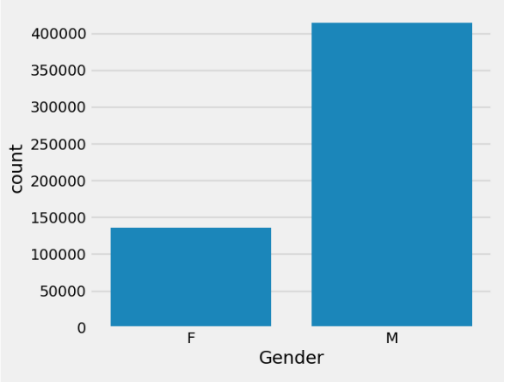

Fig. 1. Distribution of Gender variable

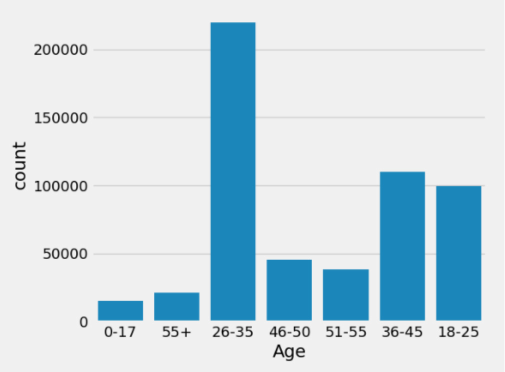

Fig. 2. Distribution of Age variable

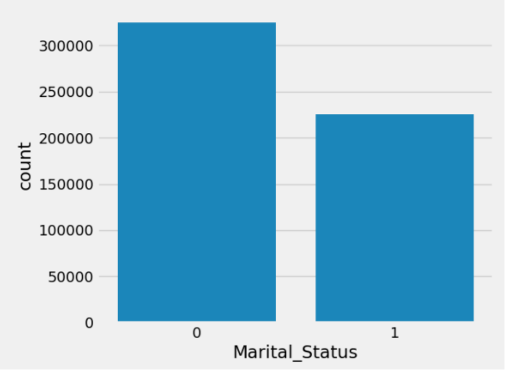

Fig. 3. Distribution of Marital variable

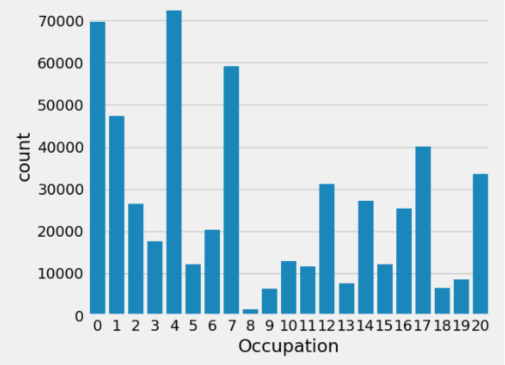

Fig. 4. Distribution of Occupation variable

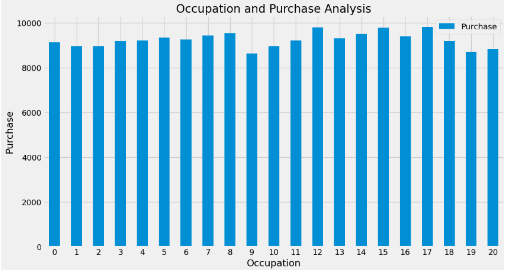

Fig. 5. Occupation and Purchase nalysis

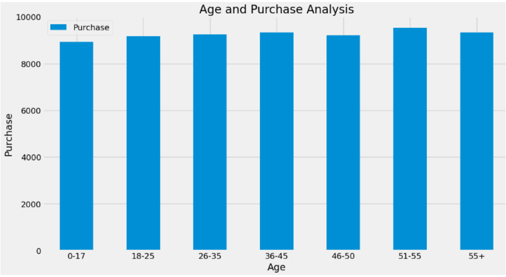

Fig. 6. Age and Purchase Analysis

Age, Gender) and the dependent variable 'Purchase'. Bar plots were used to visualize these relationships, aiding in understanding which factors might have a more significant impact on purchase amounts.

c) Data Cleaning and Preprocessing: Data cleaning focused on handling missing values in 'Product_Category_2' and 'Product_Category_3'. Missing values were filled with a placeholder value of -2.0 and converted to the float32 type. This approach is a common practice in data preprocessing, especially when dealing with missing numerical data that needs to be imputed.

The 'Gender' column was transformed from categorical ('F' and 'M') to numerical (0 and 1) format, facilitating its use in the machine learning models.

Label Encoding was applied to the 'Age', 'City_Category', and 'Stay_In_Current_City_Years' columns, converting them from categorical to a numerical format. This transformation was necessary since machine learning algorithms typically require numerical input.

A correlation heatmap was generated to visualize the correlation between different numerical features. This step is vital in identifying multicollinearity or potential predictors that have a strong correlation with the target variable.

d) Feature Selection and Model Training: The DataFrame was split into features (X) and the target variable (y). Features like 'Product_ID' were dropped, and 'Purchase' was set as the target variable. The remaining features were scaled using `StandardScaler` to normalize the data, ensuring that all features contributed equally to the model's performance.

A training function named `train` was defined to encapsulate the process of model training and evaluation. It split the data into training and test sets, fitted the model on the training data, and then made predictions on the test set. The model's performance was evaluated using the Mean Squared Error (MSE) and cross-validation score. This function streamlined the process of training and evaluating multiple models.

e) Model Implementation and Evaluation: Four different regression models were trained and evaluated:

1) Linear Regression: Provided a baseline for model performance. It's a simple model but essential for understanding the basic structure and trends in the data.
2) Decision Tree Regressor: Offered insights into the non-linear relationships and feature importance. It's more flexible than linear models and can capture complex patterns.
3) Random Forest Regressor: An ensemble model that builds multiple decision trees and merges them for more accurate and stable predictions. It addresses the overfitting issue commonly seen in decision trees.
4) Extra Trees Regressor: Similar to Random Forest but introduces additional randomness in the construction of trees. It's particularly useful for reducing model variance and improving generalization.

Each model's performance was quantitatively assessed using MSE and cross-validation scores. This approach provided a comprehensive evaluation of the models, taking into account both their accuracy and their ability to generalize to unseen data.

The results from each model were as follows:

- Linear Regression: MSE of 4617.88 and CV Score of 4625.21.
- Decision Tree: MSE of 3496.89 and CV Score of 3467.87.
- Random Forest: MSE of 2991.51 and CV Score of 2988.28.
- Extra Trees Regressor: MSE of 3174.43 and CV Score of 3168.31.

These results indicated the effectiveness of ensemble methods (Random Forest and Extra Trees) in handling the dataset's complexity and achieving lower error rates compared to simpler models like Linear Regression and Decision Tree.

IV. Comparison

In the Comparison section of the project, the performance of the implemented machine learning models — Linear Regression, Decision Tree, Random Forest, and Extra Trees Regressors — is critically evaluated against each other. This evaluation not only determines the effectiveness of each algorithm in the context of your Black Friday Sales data but also offers insights into their strengths and limitations.

a) Methodology of Comparison: The comparison was grounded in quantitative metrics — specifically, the Mean Squared Error (MSE) and Cross-Validation (CV) scores. These metrics were chosen for their ability to provide a clear and consistent measure of each model's predictive accuracy and generalization capability.

1) Mean Squared Error (MSE): MSE is a standard metric used to measure the average of the squares of the errors, that is, the average squared difference between the estimated values and the actual value. A lower MSE value indicates a better fit of the model to the data.
2) Cross-Validation Score: Cross-validation involves partitioning the data into subsets, training the model on one subset, and validating it on another. This process is repeated multiple times, and the average score is calculated. The CV score provides a more robust assessment of the model's performance, as it reduces the risk of overfitting and ensures that the model's effectiveness is not dependent on the particular way the data is split.

b) Performance Analysis: The comparison revealed the following results for each model:

1) Linear Regression:
   - MSE: 4617.88
   - CV Score: 4625.21
   - Analysis: As a baseline model, Linear Regression provided a preliminary understanding of the dataset. However, its higher MSE and CV scores indicated a relatively poorer fit to the data, likely due to its inability to capture the more complex, non-linear relationships in the dataset.
2) Decision Tree Regressor:
   - MSE: 3496.89
   - CV Score: 3467.87
   - Analysis: The Decision Tree model showed a significant improvement over Linear Regression. Its ability to capture non-linear relationships made it more effective. However, the risk of overfitting with Decision Trees was a concern, reflected in the variance between the MSE and CV scores.
3) Random Forest Regressor:
   - MSE: 2991.51
   - CV Score: 2988.28
   - Analysis: The Random Forest model outperformed both the Linear Regression and Decision Tree models. By averaging multiple decision trees, it not only captured complex relationships but also reduced overfitting, as evidenced by its lower and more consistent MSE and CV scores.
4) Extra Trees Regressor:
   - MSE: 3174.43
   - CV Score: 3168.31
   - Analysis: The Extra Trees Regressor, while not outperforming the Random Forest, still showed a significant improvement over the simpler models. Its approach of adding additional randomness to the model building process contributed to a robust and generalizable model, albeit with a slight trade-off in MSE and CV scores compared to the Random Forest.

c) Conclusions from the Comparison: The comparison highlighted the effectiveness of ensemble methods, particularly Random Forest and Extra Trees, in dealing with the complexities of the Black Friday Sales dataset. While simpler models like Linear Regression provided valuable baseline information, their inability to handle the dataset's complexities became apparent. The Decision Tree model, with its improved performance, still faced challenges regarding model overfitting.

Random Forest emerged as the most effective model, striking a balance between complexity and predictability, followed closely by Extra Trees, which demonstrated a comparable performance with an added layer of randomness. These findings underscore the importance of choosing the right model based on the specific characteristics of the dataset and the predictive task at hand.

This comparative analysis not only serves as a cornerstone for this specific project but also contributes to the broader understanding of applying machine learning algorithms in retail data analysis. It provides a framework for future studies and applications in similar domains, where the choice of algorithm can significantly impact the insights drawn from the data.

V. Future Directions

a) Exploring Advanced Machine Learning Techniques: A key area for future exploration in this domain is the integration of more advanced machine learning techniques, particularly deep learning models such as neural networks. These models are renowned for their ability to process and learn from large datasets, capturing complex, non-linear patterns that traditional algorithms might miss. In the context of retail sales prediction, deep learning could offer more nuanced insights, especially in understanding intricate consumer behaviors and preferences. Additionally, the adoption of unsupervised learning methods, such as clustering and association rules, could unveil latent patterns in customer purchasing habits, leading to more personalized marketing strategies. Experimenting with these advanced techniques could also involve leveraging natural language processing (NLP) to analyze customer reviews and feedback, providing a more holistic view of consumer sentiment.

b) Integrating Real-time Analytics and Expanding Scope: Another significant direction is the implementation of real-time analytics. Incorporating real-time data processing capabilities would enable dynamic adjustments to predictions, reflecting ongoing trends and immediate feedback from marketing campaigns. This approach is particularly pertinent in the fast-paced retail environment, where consumer trends can shift rapidly. Moreover, expanding the scope of the analysis to encompass a broader range of events, including other major sales periods like Cyber Monday or holiday seasons, could provide a more comprehensive understanding of consumer behavior throughout the year. Such expansion would not only enhance the predictive models' robustness but also offer retailers a year-round perspective on inventory management and marketing strategies. Embracing a more extensive temporal scope could also involve longitudinal studies to track changes in consumer behavior over time, adapting predictive models to evolving market dynamics and consumer preferences.

VI. Conclusion

The exploration of machine learning algorithms in predicting Black Friday sales has demonstrated significant potential in understanding and forecasting consumer purchasing patterns. Through the application of models ranging from Linear Regression to more sophisticated ensemble methods like Random Forest and Extra Trees, this study has highlighted the intricate relationship between consumer attributes and their spending behavior. The comparative analysis not only revealed the superiority of ensemble models in handling complex datasets but also underscored the importance of selecting appropriate modeling techniques tailored to the specific nuances of the data. These insights extend beyond the realm of retail sales, offering valuable contributions to the broader field of predictive analytics. As the retail landscape continues to evolve, the findings from this study serve as a foundation for future research and practical applications, aiming to enhance the accuracy and relevance of predictive models in retail and consumer behavior analysis.

References

[1] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). ”An Introduction to Statistical Learning.” Springer. [Covers fundamental concepts in machine learning and statistical learning, including regression models and tree-based methods.]

[2] Breiman, L. (2001). ”Random Forests.” Machine Learning, 45(1), 5-32. [A foundational paper on Random Forest algorithm, explaining its theory and application.]

[3] Geurts, P., Ernst, D., Wehenkel, L. (2006). ”Extremely randomized trees.” Machine Learning, 63(1), 3-42. [Key paper on Extra Trees algorithm, detailing its methodology and advantages over other models.]

[4] Hastie, T., Tibshirani, R., Friedman, J. (2009). ”The Elements of Statistical Learning.” Springer. [Comprehensive guide on various statistical learning techniques, ideal for understanding the theoretical background of the algorithms used.]

[5] Kelleher, J.D., Namee, B.M., D'Arcy, A. (2020). ”Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.” MIT Press. [Useful for understanding the practical application of machine learning algorithms in data analytics.]

[6] Chen, Y., Pavlou, P. A. (2014). ”Data Analytics in Retail: Opportunities and Challenges.” Business Horizons, 57(5), 651-660. [Discusses the role of data analytics in retail, touching on predictive modeling and consumer behavior analysis.]

[7] Smith, A. D., Offodile, O. F. (2011). ”Predictive Modeling and Consumer Segmentation in E-Commerce: A Review.” Journal of Electronic Commerce Research, 12(3), 220-237. [Provides insights into the application of predictive modeling in the context of e-commerce and consumer segmentation.]