Data Analysis and Price Prediction of Black Friday Sales Using Machine Learning Techniques

Erfan Panahi

January 2025

Combined Introduction and Explorations

Black Friday has become a cornerstone of consumer culture, representing a significant spike in sales and offering unique insights into shopping behavior. With the rapid adoption of e-commerce, analyzing Black Friday sales has become crucial for businesses aiming to optimize strategies and cater to customer preferences. The **Black Friday Sales Dataset** on Kaggle (link) provides a rich source of transactional data, encompassing demographic attributes, product categories, and purchase amounts.

This project seeks to investigate patterns in this dataset, focusing on key factors influencing purchase behavior, sales distribution across product categories, and predictive modeling of customer spending. By leveraging machine learning techniques, the goal is to provide actionable insights that can enhance decision-making for retailers.

Dataset Overview

The **Black Friday Sales Dataset** contains sales transaction data from a retail store during the Black Friday sales event. It includes 550,069 rows and 12 attributes, such as:

User_ID: Unique ID of the customer.

Product_ID: Unique ID of the product.

Gender: Gender of the customer (M/F).

Age: Age group of the customer.

Occupation: Occupation of the customer (masked).

City_Category: Category of the city (A, B, C).

Stay_In_Current_City_Years: Number of years the customer has stayed in the current city.

Marital_Status: Marital status of the customer (0 = single, 1 = married).

Product_Category_1/2/3: Categories of the product purchased.

Purchase: Amount of purchase (in dollars).

The dataset is publicly available on Kaggle and has been used in various studies to predict customer purchase behavior during Black Friday sales. The dataset contains both categorical and numerical variables, making it suitable for exploratory data analysis (EDA) and predictive modeling.

Early exploration reveals several interesting features:

1. Missing data in some product categories (e.g., Product_Category_2 and Product_Category_3).
2. The presence of categorical variables like Gender and City Category, which need encoding for analysis.
3. Potential correlations between demographic variables and purchase patterns, warranting deeper investigation.

Research Questions

Based on preliminary observations, the following research questions will guide this study:

1. **What demographic factors most significantly influence purchase behavior during Black Friday?**
    - **Preliminary Analyses:** Exploratory data analysis will involve visualizing customer attributes and their purchase distributions using count plots, bar plots, and heatmaps. Relationships between demographics and purchase amounts will be analyzed.
    - **Proposed Tests:**
        - ANOVA to identify significant differences in purchase behavior across age groups.
        - T-tests to compare spending patterns between genders.
2. **What are the most effective predictors for forecasting purchase amounts?**
    - **Preliminary Analyses:** Regression models (Linear, Ridge, and Random Forest) will be trained to identify and rank significant predictors. Cross-validation will ensure model robustness.
    - **Proposed Tests:**
        - Variance Inflation Factor (VIF) for assessing multicollinearity.
        - Model evaluation using Mean Squared Error (MSE) and feature importance rankings.
3. **How do product categories influence sales trends, and what combinations drive the highest purchases?**
    - **Preliminary Analyses:** Exploratory data analysis will involve visualizing customer attributes and their purchase distributions using count plots, bar plots, and heatmaps. Relationships between demographics and purchase amounts will be analyzed.
    - **Proposed Tests:**
        - Chi-square tests to assess associations between product categories and customer demographics.
        - Cluster analysis to find patterns in product combinations.

Exploratory Data Analysis and Visualizations

To better understand the dataset, several visualizations were created to explore key attributes and their relationships:

1. **Total Purchases by Gender (Figure 1):** A bar chart displays the total purchase amount for male and female customers. It highlights that **male customers contribute significantly more to overall sales** during Black Friday compared to female customers. This insight suggests that retailers should target male audiences with tailored marketing strategies to maximize revenue.

Figure 1. **Total Purchases by Gender**

2. **Purchase Distribution by Age Group (Figure 2):** A box plot illustrates the distribution of purchase amounts across various age groups. It reveals that the **26–35 age group has the highest median purchase amount**, while younger customers (18–25) tend to spend less. The broader spread in the 36–45 age group indicates greater variability in spending behavior. These insights can help tailor promotional offers to specific age groups.

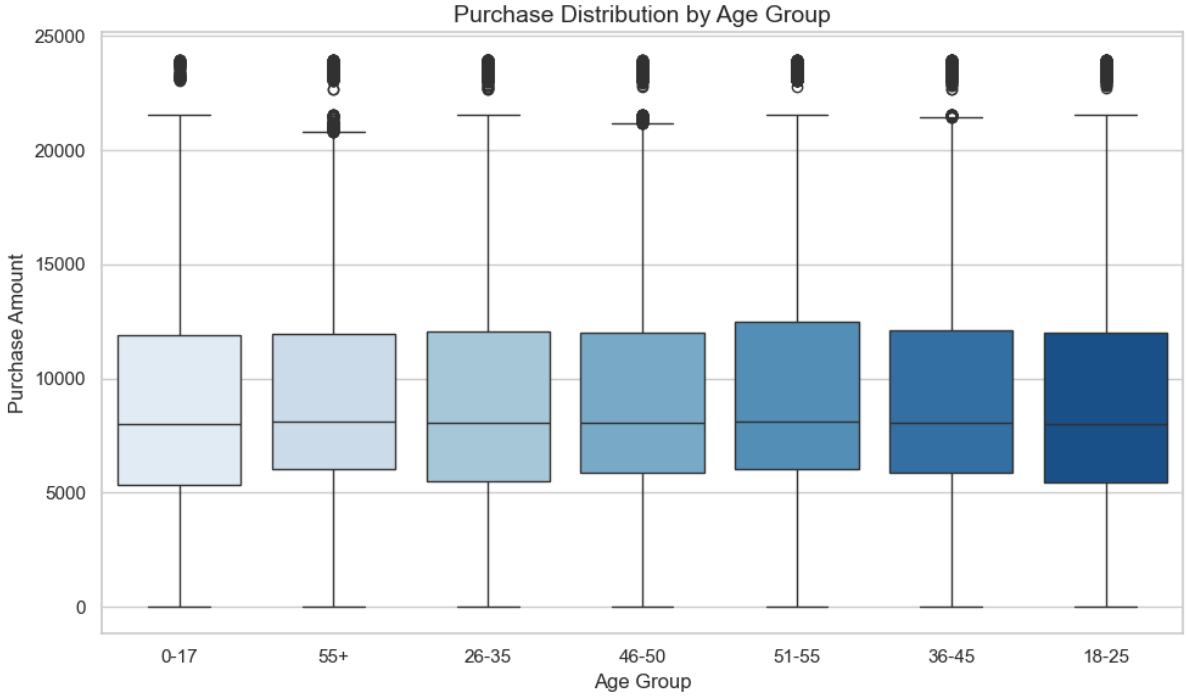

Figure 2. **Purchase Distribution by Age Group**

3. **Correlation Heatmap of Key Variables (Figure 3):** The heatmap illustrates how key numerical variables relate to each other. In this dataset, Purchase does not exhibit strong correlations with other numerical variables, indicating the potential need for feature engineering. For example,

interaction terms or transformations might enhance predictive modeling. While the direct relationships are weak, this analysis highlights the independence of certain predictors, which can be beneficial in regression modeling.

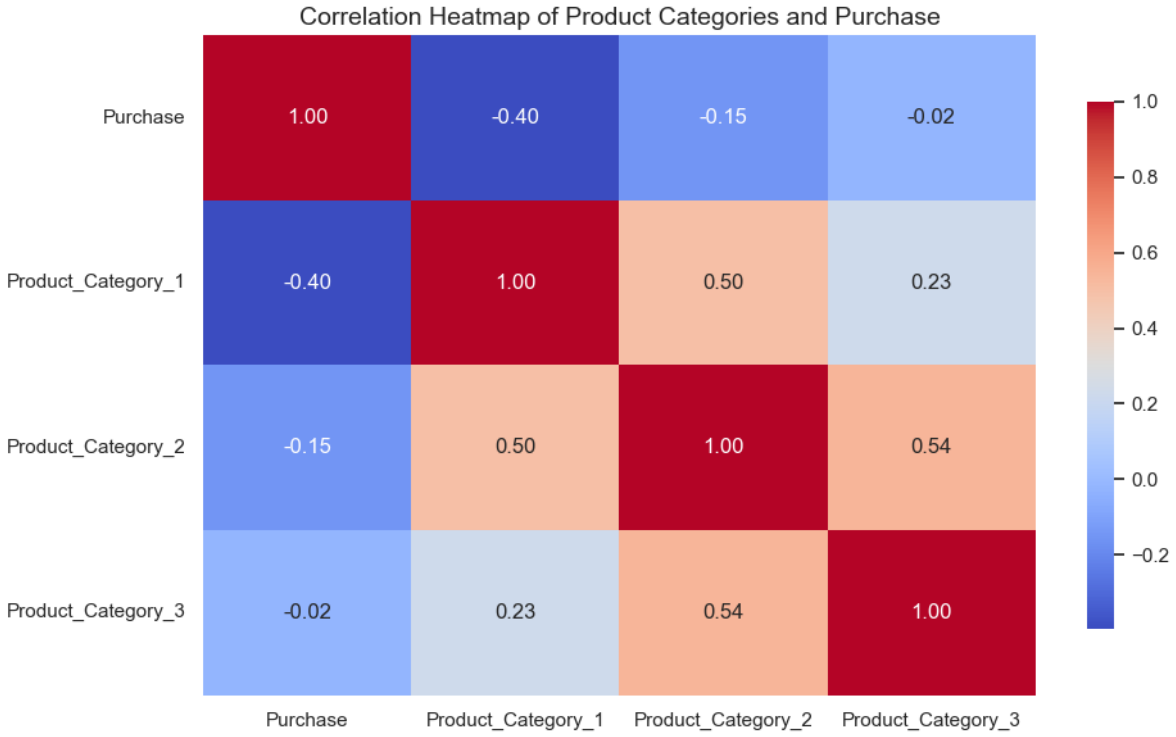

Figure 3. **Correlation Heatmap of Key Variables**

4. **Customer Count by City Category (Figure 4):** A count plot shows the distribution of customers across city categories (A, B, C). It reveals that **City B has the highest number of shoppers**, followed by City C, with City A showing the lowest count. Retailers might focus on boosting sales in City A by offering targeted promotions or discounts.

Figure 4. **Customer Count by City Category**

Proposed Methods

To address these questions, the project will employ:

1. **Data Preprocessing:**
   - Handle missing values with imputation techniques (mean/mode).
   - Encode categorical variables using label encoding or one-hot encoding.
   - Remove irrelevant columns (e.g., user_id, product_id) to prevent model bias.
2. **Exploratory Data Analysis (EDA):**
   - Use heatmaps and pair plots to visualize relationships between variables.
   - Summarize data distributions using descriptive statistics and box plots.
3. **Statistical Testing:**
   - Parametric tests (ANOVA, t-tests) to analyze group differences.
   - Non-parametric tests (Kruskal-Wallis, Mann-Whitney U) where necessary.
4. **Machine Learning Models:**
   - Implement regression models (Linear, Ridge, Lasso, Random Forest).
   - Use cross-validation to validate models and select the best performer.
5. **Evaluation Metrics:**
   - Evaluate model performance using MSE, MAE, and R^2.
   - Visualize residuals to check model assumptions.

Bonus Component

The article **"Data Analysis and Price Prediction of Black Friday Sales using Machine Learning Techniques"** highlights the superior performance of the Random Forest Regressor for sales prediction (MSE = 3062.72). As part of the bonus task, I will:

1. Recreate their Random Forest model and validate its findings using my analysis.
2. Compare it with other advanced models, such as Gradient Boosting or XGBoost, with hyperparameter tuning.
3. Investigate alternative predictors or preprocessing steps not emphasized in the original study.

Conclusion

This proposal outlines a structured approach to analyzing the Black Friday Sales Dataset. By addressing the research questions through statistical analysis and machine learning models, the study aims to provide meaningful insights into consumer behavior and sales trends. These findings will empower retailers to optimize their strategies and enhance the shopping experience during high-demand sales events like Black Friday.