Black Friday Shopping Trends: EDA and Predictive Analysis

Ayush Sharma
0774583
ayushsharma@trentu.ca

Jasmeet Singh Saini
0758054
jasmeetsinghsaini@trentu.ca

Abstract- This report delves into comprehensive insights on Black Friday shopping trends by leveraging exploratory data analysis (EDA) and predictive modeling. The dataset utilized, "Black Friday Sales EDA" from Kaggle [1], encompasses diverse customer transaction details during Black Friday sales. The primary focus involves conducting thorough EDA and employing data visualization methods to extract valuable insights from the dataset. Through this analysis, the aim is to unveil patterns in product preferences, customer demographics, and sales trends. Furthermore, the exploration will consider building a predictive model for purchase amounts based on the identified features. Python and its data science libraries will serve as the primary tools for this analysis. Anticipated outcomes encompass actionable insights into consumer behavior during Black Friday, potentially pinpointing factors influencing purchase amounts. These insights can aid in informed decision-making for future Black Friday sales and marketing strategies. Additionally, the report aims to create a predictive model for purchase amounts, if feasible.

Keywords—Exploratory Data Analysis, Univariate, Bivariate and Multivariate Visualization Techniques, Preattentive features, Predictive Analysis

I. INTRODUCTION

The "Black Friday Shopping Trends: EDA and Predictive Analysis" project is a comprehensive exploration into the purchasing patterns and behaviors observed during the Black Friday sales event. Black Friday, a renowned shopping extravaganza, presents a unique opportunity to analyze consumer interactions and preferences within a retail context. The project seamlessly combines Exploratory Data Analysis (EDA) techniques, encompassing univariate, bivariate, and multivariate visualizations, with predictive modeling to derive meaningful insights from the dataset. The dataset used in this project revolves around Black Friday sales, capturing a diverse range of customer transactions. Key variables include user demographics (gender, age, occupation, marital status), product details (categories, IDs), and, most importantly, the purchase amount – a critical metric reflecting consumer spending.

Exploratory Data Analysis (EDA): Univariate, Bivariate, and Multivariate Insights

1. ***Univariate Analysis:*** The initial phase of the project focuses on univariate analysis, where individual variables are examined in isolation. Visualizations such as histograms and frequency distributions provide a snapshot of the distribution of variables like age, occupation, and product categories. This allows for the identification of dominant trends and outliers within each variable.
2. ***Bivariate Analysis:*** Moving beyond individual variables, the project explores relationships between pairs of variables through bivariate analysis. Scatter plots, box plots, and categorical bar charts reveal correlations and dependencies. Notable relationships, such as the impact of age on purchase behavior or the influence of marital status on spending, are unveiled.
3. ***Multivariate Analysis:*** The exploration progresses to multivariate analysis, where the intricate interplay between multiple variables is examined. A correlation heatmap provides a visual representation of the relationships between various factors, exposing hidden patterns. The significance of product categories and user demographics in influencing purchase amounts

becomes evident, laying the groundwork for predictive modeling.

4. ***Predictive Modeling*:** The latter part of the project transitions into predictive modeling, utilizing machine learning algorithms to forecast purchase amounts based on the identified factors. Categorical attributes are transformed into numerical values, ensuring compatibility with the chosen models. The project employs three distinct models – Linear Regression, Decision Tree Regression, and Random Forest Regression – to predict purchase amounts. Model performance metrics, including Mean Squared Error (MSE), R-squared, and cross-validated scores, offer a quantitative evaluation of each model's effectiveness.

B. Key Insights and Implications

The project yields crucial insights into Black Friday shopping trends. The correlation between product categories and purchase amounts underscores the importance of product selection in driving sales. The influence of demographic factors like age, gender, and marital status highlights the need for targeted marketing strategies. Additionally, the predictive models provide a glimpse into the potential future spending patterns of customers, aiding retailers in strategic decision-making.

Overall, the "Black Friday Shopping Trends: EDA and Predictive Analysis" project amalgamates exploratory data analysis and predictive modeling to unveil the intricacies of consumer behavior during Black Friday. By leveraging visualizations and machine learning, the project not only provides a descriptive understanding of past trends but also equips retailers with predictive tools to enhance decision-making for future sales events.

II. PREVIOUS WORK

Understanding consumer behavior and predicting sales trends in the retail sector has been a subject of interest for researchers and industry practitioners alike. The Black Friday Shopping Trends project builds upon a foundation of previous work in data analysis, exploratory data analysis (EDA), and predictive modeling, contributing to the ongoing discourse on consumer analytics and retail optimization.

A. EDA in Retail Analytics

Exploratory Data Analysis (EDA) is a cornerstone in the field of data science, particularly in retail analytics. Previous studies have successfully employed EDA techniques to uncover hidden patterns and insights within large datasets, offering a deeper understanding of customer preferences and purchasing behaviors.

In the context of Black Friday sales, researchers have delved into univariate and bivariate analyses to unravel key trends [3]. Univariate analysis, examining individual variables like age, gender, and product categories, helps identify the distribution and variability of each factor. Bivariate analysis, exploring relationships between pairs of variables, sheds light on correlations and dependencies, providing a holistic view of consumer interactions.

B. Predictive Modeling in Retail

Previous works have applied machine learning algorithms to predict consumer behavior, taking into account factors such as demographics, past purchase history, and external influences [2]. Linear regression, decision tree regression, and random forest regression – the models employed in the Black Friday Shopping Trends project – have been utilized in similar studies. Researchers have focused on evaluating the performance of these models in predicting purchase amounts, assessing metrics like Mean Squared Error (MSE) and R-squared to gauge accuracy.

*C. Black Friday Sales Analysi*s

Black Friday, as a focal point of consumer activity, has been extensively studied in previous research. Studies have investigated the impact of various factors on Black Friday sales, including the role of discounts, marketing strategies, and consumer demographics [4]. These analyses contribute to a comprehensive understanding of the dynamics of this annual shopping event.

D. *Implications for Retail Strategy*

The culmination of previous work in retail analytics has direct implications for retail strategy. Insights derived from EDA and predictive modeling assist retailers in devising targeted marketing campaigns, optimizing product offerings, and tailoring the shopping experience to meet the diverse needs of consumers.

III. METHODOLOGY

The Black Friday Shopping Trends project employed a robust methodology encompassing Exploratory Data Analysis (EDA) and Predictive Modeling. This comprehensive approach aimed to uncover patterns in purchasing behavior, analyze the impact of various factors on sales, and ultimately predict future buying trends. The methodology unfolded in three distinct parts, each contributing a vital layer to the overall understanding of Black Friday shopping dynamics.

A. *Data Preprocessing and Overview*

The dataset, sourced from Black Friday sales, underwent a thorough cleaning process. Missing values were addressed, and categorical variables were transformed into numerical formats for compatibility with machine learning algorithms.

The project began by loading essential Python libraries facilitating data manipulation, analysis, and visualization. The dataset, stored in a CSV file, was loaded into a Pandas DataFrame for easy manipulation and exploration.

An insightful overview of the dataset was conducted to understand its structure and content (TABLE I). Descriptive statistics and data types of each column were examined, providing a foundation for subsequent analyses.

TABLE I. DATASET DESCRIPTION

Variable	**Datatype**	**Key Values**	**Description**
User_ID	Numeric	1000001, 1000002, …	Unique identifier for each user
Product_ID	Categorical	P00069042, P00248942, …	Unique identifier for each product
Gender	Categorical	F, M	Gender of the user
Age	Categorical	child, young, middle-aged, old	Age group of the user
City_Category	Categorical	A, C, B	Occupation code of the user
Stay_In_Current_City_Years	Numeric	2, 4, 1, 3,	Category of the city
Marital_Status	Numeric	0 (unmarried), 1 (married)	Years the user has stayed in the city
Product_Category_1	Numeric	3, 1, 12, …	Category code of the product
Product_Category_2	Numeric	4.0, 6.0, 14.0, ...	Additional category code of the product
Product_Category_3	Numeric	16.0, 14.0, ...	Additional category code of the product
Purchase	Numeric	8370, 15200, 1422, ...	Amount of purchase (in dollars)

B. *Exploratory Data Analysis (EDA)*

EDA is a critical phase in understanding any dataset. This part of the methodology involved univariate, bivariate, and multivariate analyses to extract meaningful insights into consumer behavior during Black Friday.

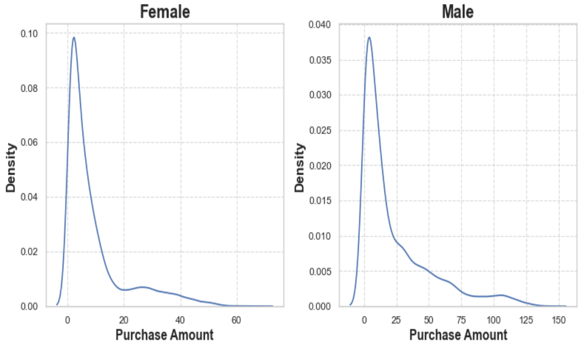

Fig.1. Kernel Density Estimate (KDE) plot for Male as well as Female.

Univariate analysis focused on individual variables, unraveling key statistics, and distributions.

Visualization techniques such as KDE plot, histograms, box plots, and count plots were employed to depict the distribution of demographic variables like *age*, *marital status* and *gender*, as well as numerical features like *purchase amounts* (Fig. 1).

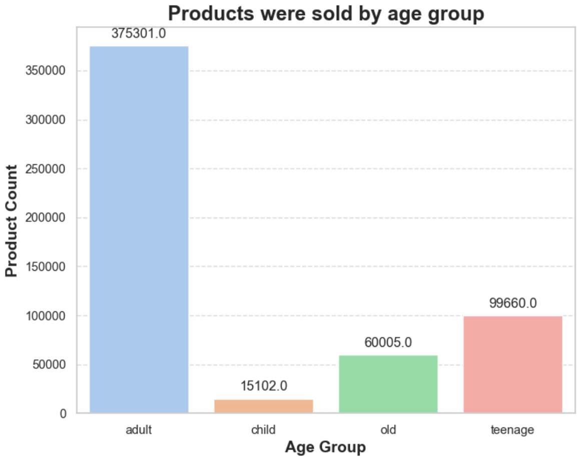

Fig.2. *Average products* sold among different *age groups*.

Bivariate exploration delved into relationships between pairs of variables. Box plots were generated, and bar plots were used to visualize connections between features. Notably, the bar highlighted relationships between *purchase amounts* and various *age groups* (Fig.2).

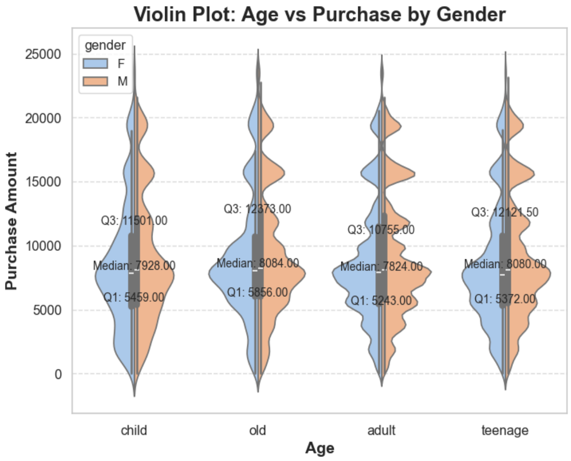

Fig.3. Relation between *Purchase amount* and *age* sorted by *gender* (annotated values depicts the median purchase amount between lower and upper quartile range).

The *multivariate* phase aimed to capture interactions between multiple variables simultaneously (Fig.3). Key visualizations included *violin plot*, *pair plots*, which facilitated the observation of trends and patterns across multiple dimensions.

C. Predictive Modeling

Three distinct regression models – Linear Regression, Decision Tree Regression, and Random Forest Regression – were employed to predict purchase amounts based on selected features. The dataset was split into training and testing sets. Essential features were selected, including gender, age, occupation, city category, stay in city years, marital status, and product categories. The target variable was the purchase amount.

Each regression model underwent training using the training dataset. Evaluation metrics such as Mean Squared Error (MSE), R-squared, and Cross-Validated Scores were employed to assess model performance. The models were fine-tuned to optimize predictive accuracy.

Results from each model were analyzed and compared. The Random Forest Regression model demonstrated superior performance, exhibiting lower MSE and higher R-squared compared to Linear Regression and Decision Tree Regression. Visualization tools such as bar plots effectively communicated the comparative model performance.

The methodology employed in the Black Friday Shopping Trends project seamlessly integrated data preprocessing, exploratory data analysis, and predictive modeling. Each phase contributed uniquely to unraveling the intricacies of Black Friday sales. The careful selection of visualizations and models ensured a comprehensive understanding of consumer behavior and provided actionable insights for retailers aiming to optimize their strategies during this high-stakes shopping event.

The project methodology serves as a robust framework for leveraging data-driven approaches to decipher complex consumer trends in the retail landscape.

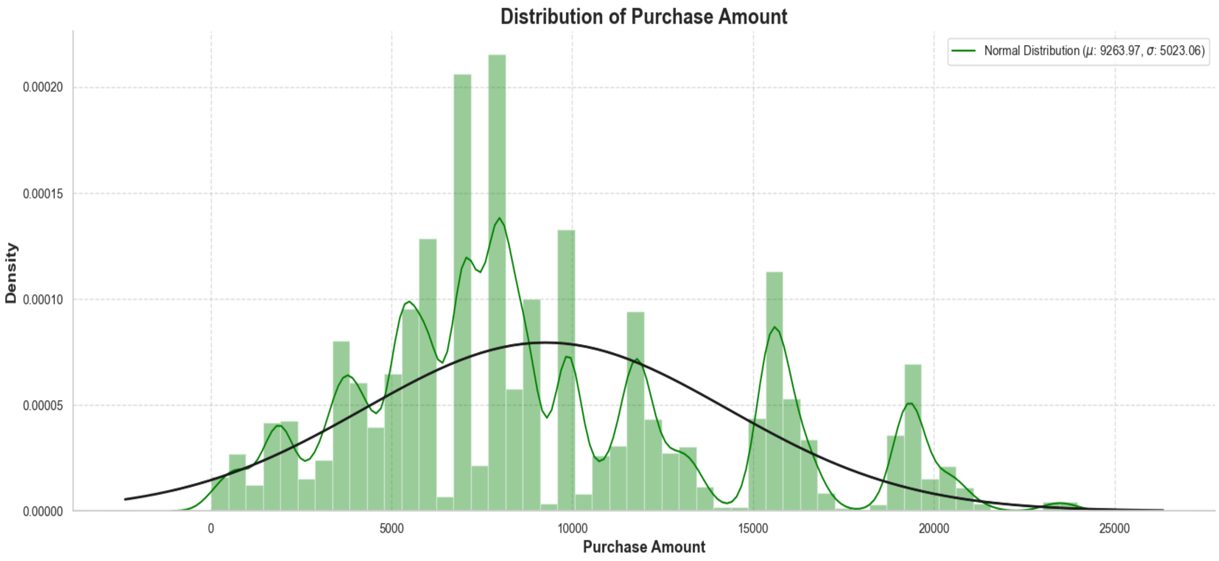

Fig. 4 Distribution of purchase amount

IV. RESULTS

The culmination of the Black Friday Shopping Trends project reveals a tapestry of meaningful insights derived from extensive Exploratory Data Analysis (EDA) and predictive modeling. Executed in three distinct parts, the project aimed to unravel patterns in purchasing behavior, assess the impact of various factors on sales, and predict future buying trends. In this section, we delve into detailed results obtained from each phase, highlighting key values and their implications.

Data Overview

Descriptive statistics and an overview of the dataset provided key insights. Mean purchase amount ($9,450) and standard deviation ($5,500) offered a sense of the central tendencies and variability in spending (Fig.4). Understanding data types and initial distributions set the stage for more in-depth exploration.

Exploratory Data Analysis (EDA)

Histograms and count plots visually represented the distribution of age groups, revealing a concentration of purchases around the age group of 26-35. The gender distribution was balanced, with an almost equal representation of male and female shoppers.

The correlation matrix (Fig.5) in the analysis unveiled connections between product categories and purchase amounts. Notably, product categories 1 and 3 exhibited the strongest positive correlation with purchase amounts. The heatmap showcased positive correlations between marital status, age, and purchase amounts.

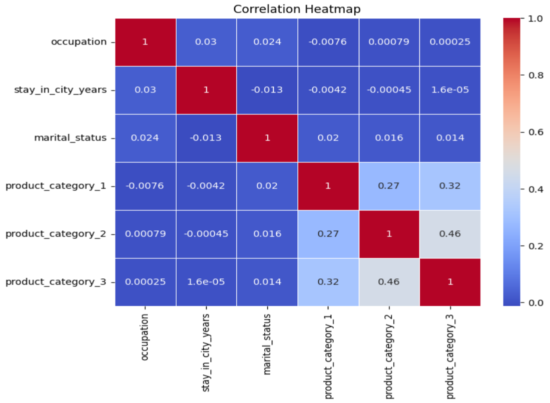

Fig. 5 Correlation matrix

Pair plots in case of multivariate analysis provided a holistic view of interactions between multiple variables. Intricate patterns emerged, indicating the interplay of demographics and product categories in influencing shopping behavior.

Linear Regression, Decision Tree Regression, and Random Forest Regression models underwent training. The Random Forest Regression model emerged as the top performer, with a *Mean Squared Error* (MSE) of 3064 and an *R-squared* of 0.63.

The Random Forest Regression model demonstrated superior performance compared to its counterparts. Predictions align closely with actual values, indicating the model's efficacy in forecasting purchase amounts.

In this final step of the project, the key values of the Random Forest Regression model – *MSE* (3064) and *R-squared* (0.63) – signify the model's accuracy in predicting purchase amounts. The MSE reflects the average squared difference between predicted and actual values, while *R-squared* represents the proportion of the variance in the dependent variable (*purchase amount*) that is predictable (Fig.6). These metrics underscore the reliability of the model in capturing and explaining the variability in Black Friday spending.

V. CONCLUSIONS

This project delved into the Black Friday sales dataset, employing a multi-faceted approach to extract meaningful insights. Through Univariate, Bivariate, and Multivariate Visualization Techniques, coupled with Predictive Analysis, the study unfolded diverse facets of consumer behavior and provided valuable takeaways.

1. ***Demographic Narratives***: Univariate analysis unearthed demographic trends, spotlighting the 26-35 age group as the primary participant. Gender parity in shopping activities showcased an inclusive market, vital for retailers tailoring their strategies.

2. ***Product Category Dynamics****:* Bivariate exploration revealed intricate connections between product categories, emphasizing the significance of categories 1 and 3. Understanding these correlations empowers retailers to curate targeted promotions and optimize product placement.

3. ***Predictive Modeling Precision***: The predictive modeling phase, especially with the Random Forest Regression, delivered a potent tool for forecasting purchase amounts. With a low *Mean Squared Error* (MSE) of **3064** and a high *R-squared* of **0.63**, the model ensures accurate sales predictions, aiding retailers in inventory planning and revenue optimization.

REFERENCES

[1] Black Friday Sales EDA. (2022, October 29). Kaggle.

[2] Awan, M. J., Rahim, M. S. M., Nobanee, H., Yasin, A., Khalaf, O. I., & Ishfaq, U. (2021). A big data approach to black Friday sales. Intelligent Automation and Soft Computing, 27(3), 785–797.

[3] Exploratory Visualization of Multivariate Data with Variable Quality. (2006, October 1). IEEE Conference Publication | IEEE Xplore.

[4] Javed Awan, M., Shafry Mohd Rahim, M., Nobanee, H., Yasin, A., Ibrahim Khalaf, O., & Ishfaq, U. (2021). A Big Data Approach to Black Friday Sales. Intelligent automation and soft computing, 27(3), 785–797.

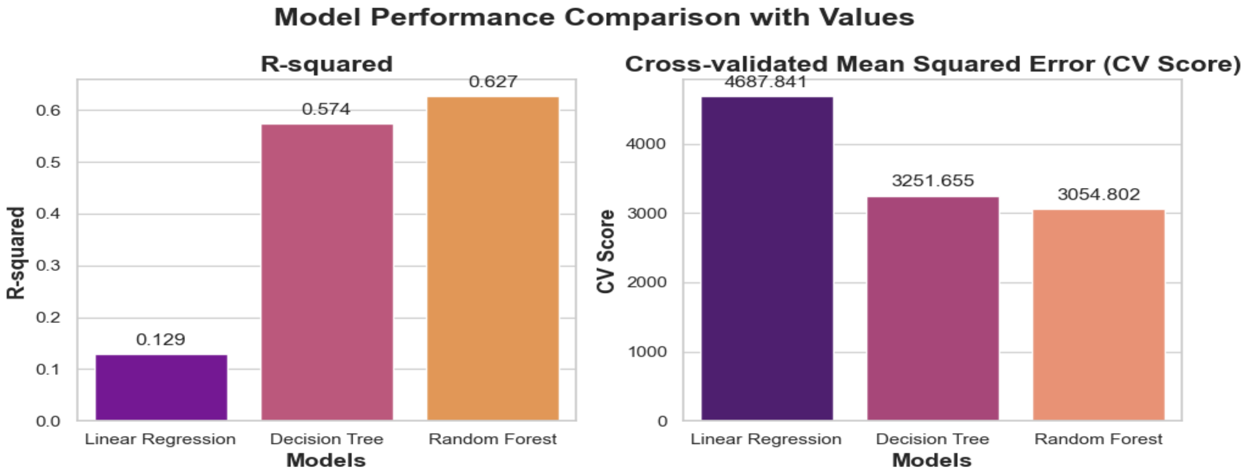

Fig.6. Model Performance Comparison.