**Leveraging Sales Analysis and Predictive Modeling: Unveiling Insights for Enhanced Online Retail Performance**

College of Professional Studies, Northeastern University

Master of Professional Studies in Analytics, ALY6040

Prof Kasun Samarasinghe,

May 30, 2023

Group 4 –

* Shah Dhairyav Jatin
* Likhitha Varakala
* Manmitha Pantangi

**Introduction**

In today's digital age, online shopping has become increasingly popular, attracting a massive portion of the population with internet access. As a result, large retailers are constantly seeking innovative strategies to maximize their profits in this evolving landscape. One powerful technique employed by these retailers is sales analysis, which involves delving into customer purchasing behaviors and patterns. By understanding these patterns, retailers can gain valuable insights that enable them to make informed decisions and enhance their sales performance.

In this report, we focus on a comprehensive dataset that encompasses online sales in the United States. The dataset comprises a wide range of products, including various merchandise and electronic items, sold across different states. With the goal of predicting future sales and increasing profitability, our analysis incorporates exploratory data analysis (EDA) techniques to examine the dataset's characteristics and identify relevant patterns. Through this report, we seek to shed light on the valuable insights that can be extracted from sales data in the context of online retail. By understanding customer preferences and anticipating their needs, retailers can seize opportunities to increase sales, enhance customer satisfaction, and thrive in the competitive online marketplace. Our analysis will provide actionable recommendations and insights that can empower retailers to make data-driven decisions and stay ahead of the curve.

**Data Preprocessing**

During the data preprocessing stage, we conducted a thorough examination of the dataset to ensure its quality and integrity. To ensure consistency and ease of use, we modified the column names in the dataset. We used regular expressions to remove spaces, dots, and symbols from the column names. Additionally, we replaced spaces in the column name "Phone No." with an underscore, making it consistent with the other column names. This formatting step ensured that the column names were clean and adhered to a standardized naming convention. In order to perform meaningful analysis on the dataset, we needed to convert certain columns to their appropriate data types. Firstly, we converted the 'order\_date' column to the datetime data type using the pd.to\_datetime() function. This allowed us to easily manipulate and extract information from the dates. Secondly, we converted the 'year' column to the year data type using the same function, enabling us to analyze the dataset based on specific years.

In the feature engineering process, we extracted the month information from the 'month' column and created a new column called 'full\_name'. The month extraction was performed by applying the extract\_month() function to the 'month' column, using datetime.strptime() to parse the date string and strftime() to format it as the month abbreviation. This resulted in a new column that contained the month information.

Fortunately, no missing data or duplicate records were identified, indicating the dataset's prominent level of completeness and uniqueness. This allowed us to proceed with the analysis confidently, knowing that the dataset provided a reliable representation of the online sales transactions. The absence of missing data eliminates the need for imputation techniques, ensuring that our analysis is based on complete and accurate information. Additionally, the absence of duplicate records ensures that each observation in the dataset is unique, preventing any biases or distortions in our analysis.

**EDA**

The dataset encompasses various dimensions of the online shopping experience, including order details, item information, customer demographics, and location specifics. With the help of visualizations such as histograms, scatter plots, and bar charts, we will examine the dataset to identify patterns, trends, and potential relationships between variables.

A screenshot of a computer

Description automatically generated

The dataset provides insights into online shopping sales and customer behavior. The dataset consists of 286,392 records, with key variables such as item\_id, qty\_ordered, price, value, discount\_amount, total, cust\_id, year, ref\_num, age, Zip, and Discount\_Percent.

The descriptive statistics analysis reveals valuable insights into the sales data:

* item\_id: The 'item\_id' column represents the unique identifier for each item. The average number of unique items sold is approximately 741,665, with a standard deviation of 95,746. This indicates that a wide range of items were sold.
* qty\_ordered: The 'qty\_ordered' column denotes the quantity of each item ordered. On average, customers ordered around 3 items per order, with a standard deviation of 4.57. The minimum order quantity is 1, while the maximum order quantity reaches up to 501, suggesting some extreme values or potential outliers.
* price: The 'price' column reflects the individual item's price. The average price is approximately $851.39, with a standard deviation of $1,741.75. The minimum price is $0, while the maximum price is $101,262.59. The distribution of prices appears to have a wide range, indicating variation in the pricing of items.
* value: The 'value' column represents the total value of each item ordered, calculated as the quantity ordered multiplied by the price. The average value is approximately $885.88, with a standard deviation of $2,073.25. The minimum value is $0, while the maximum value is $101,262.59.
* discount\_amount and Discount\_Percent: These columns provide information about the discounts applied to each order. The average discount amount is approximately $70.04, with a standard deviation of $256.88. The average discount percentage is 6.07%. This suggests that, on average, customers receive a moderate discount on their purchases.
* total: The 'total' column represents the total amount paid for each order, including any applicable discounts. The average total is approximately $815.84, with a standard deviation of $1,983.58. The minimum total is $0, while the maximum total is $101,262.59.

The below graph displays the number of occurrences for each payment method, providing a visual representation of the popularity and usage of different payment options among customers. According to the graph, the most common payment method is COD (Cash on Delivery), with a count of 102,916. It can be a popular option due to its convenience and the trust it provides for customers. With COD, customers have the flexibility to pay in cash at the time of delivery, eliminating the need for online transactions or the use of credit or debit cards. This convenience is particularly beneficial for customers who may not have access to digital payment methods or prefer the simplicity of cash transactions. Easypay follows closely behind with 69,679 occurrences. Payaxis and Easypay Voucher are the next most popular options, with counts of 31,049 and 29,763, respectively. Bank Alfalah and Easypay Mobile Account are also frequently used, with counts of A graph of payment method

Description automatically generated with low confidence23,057 and 11,536, respectively.

A picture containing text, line, plot, diagram

Description automatically generatedOther payment methods such as Jazzwallet, Jazzvoucher, Customer Credit, APG, and MCB Lite have lower frequencies but still contribute to the overall payment landscape. Cash at Doorstep and Finance Settlement have the lowest counts, indicating less prevalent usage.

Online sales exhibit a fluctuating pattern over time, with peaks and valleys indicating variations in sales volume. There is a noticeable spike in sales around December, potentially driven by the holiday shopping season, while the sales remain lower in January and February.

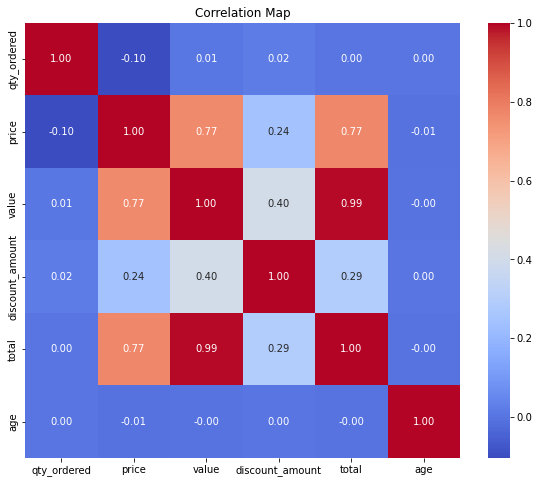
A picture containing text, screenshot, diagram, plot

Description automatically generated

The bar graph displays the frequency of distinct categories in the dataset. "Mobiles & Tablets" is the most prevalent category with 61,761 occurrences, indicating a strong demand for mobile devices. "Men's Fashion" follows closely with 40,713 occurrences, suggesting a significant interest in fashion among male consumers. Other notable categories include "Appliances" with 33,034 occurrences and "Women's Fashion" with 28,334 occurrences. Understanding these category frequencies can assist businesses in targeting their marketing and inventory strategies to align with consumer preferences and maximize sales potential.

A picture containing text, diagram, line, screenshot

Description automatically generated

The boxplot visualizes the distribution of total sales across various categories, with the x-axis representing the categories and the y-axis representing the sum of total sales. Notably, the 'Health & Sports' category stands out with an extreme outlier of 100k sales, indicating a significantly higher sales figure compared to other categories. The 'Mobiles & Tablets' category also exhibits outliers ranging from 40k to 80k, suggesting that there are a few instances of exceptionally high sales within this category. The boxplot confirms that the 'Mobiles & Tablets' category has the highest overall sales, followed by the 'Entertainment' and 'Computing' categories, both having extreme sales values of 50k. These insights offer a clear understanding of the distribution of sales and the dominance of the 'Mobiles & Tablets' category in terms of revenue.

The findings from the correlation analysis revealed significant insights into the sales dataset. Initially a substantial positive correlation of 0.77 between the product's price and the order's total value was found. This suggests that expensive items have a significant influence on the total sales value. It implies that a key factor in determining the effectiveness of sales is pricing strategy. In addition, a substantial positive correlation of 0.99 between the total order value and the total sales was discovered. This suggests that the order's value is a highly accurate indicator of the overall number of sales produced. The entire sales increase in direct proportion to the increase in order value. This result highlights the significance of precisely assessing the order value when predicting sales performance.

**Prediction Models**

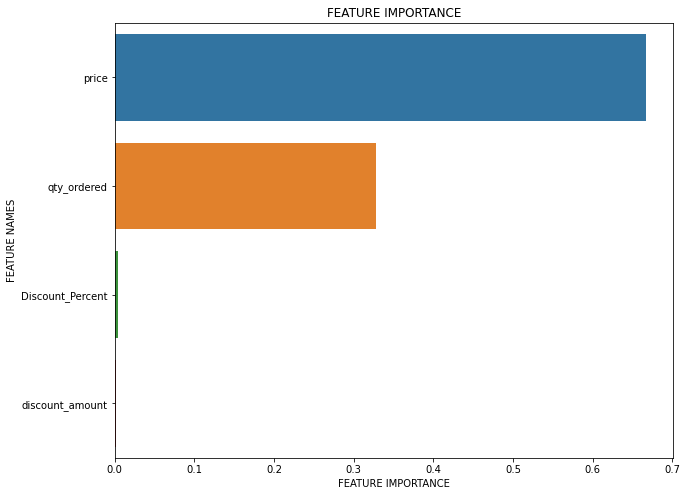
Further, we present the findings of our analysis on various regression models for predicting sales. We compared three models: Random Forest, XGBoost, and Decision Tree Regressor. The aim of our analysis was to determine which model performs best in terms of mean squared error (MSE) and root mean squared error (RMSE) metrics. Additionally, we examined the feature importance for each model to identify the key factors influencing sales.

Decision Tree Regressor

We initially applied the Decision Tree Regressor to the training data and made predictions on the test data. The model's performance was evaluated based on the MSE, RMSE, and R-squared metrics. Additionally, we conducted a feature importance analysis to determine the relative influence of each input variable on sales.

* **Decision Tree Regressor with Hyperparameter Tuning**

To enhance the performance of the Decision Tree Regressor, we conducted hyperparameter tuning using a random search with cross-validation. This enabled us to identify the optimal combination of hyperparameters that yielded the best results. We evaluated the performance of the tuned model based on MSE, RMSE, and R-squared.

The model achieved a mean squared error of 6977.14, a root mean squared error of 83.53, and an R-squared value of 0.9982. Feature importance analysis revealed that the most influential feature was price (66.7%), followed by quantity ordered (32.8%), discount amount (0.11%), and discount percentage (0.4%).

**After** **Hyperparameter Tuning**

The best model achieved a mean squared error of 6869.72, a root mean squared error of 82.88, and an R-squared value of 0.9982. The optimal hyperparameters were determined as follows:

min\_samples\_split=2, min\_samples\_leaf=1, max\_features='auto', and ax\_depth=None.

* **Random Forest**

Next, we trained a Random Forest regressor with 100 estimators using the training data. This ensemble model allowed us to obtain more accurate predictions by aggregating the results from multiple decision trees. The model's performance was assessed based on MSE, RMSE, and R-squared. Additionally, we conducted a feature importance analysis to identify the variables that significantly influenced sales.

A blue and orange rectangle

Description automatically generated with low confidenceThe model yielded a mean squared error of 3646.10, a root mean squared error of 60.38, and an R-squared value of 0.9991. Feature importance analysis indicated that price had the highest importance (66.58%), followed by quantity ordered (32.95%), discount amount (0.15%), and discount percentage (0.33%).

* **XGBoost**

For our final model, we utilized the XGBoost regressor. This advanced gradient boosting algorithm is known for its strong predictive performance. We fitted the training data to the XGBoost model and made predictions on the test data. The model's performance was evaluated based on MSE, RMSE, and R-squared. Additionally, we conducted a feature importance analysis to identify the key variables affecting sales.

A blue and orange rectangle

Description automatically generated with low confidenceThe XGBoost model demonstrated a mean squared error of 1800.61, a root mean squared error of 42.43, and an R-squared value of 0.9995. Feature importance analysis revealed that price was the most important feature (49.29%), followed closely by quantity ordered (48.78%), discount amount (0.66%), and discount percentage (1.27%).

XGBoost demonstrated greater performance in terms of both mean squared error (MSE) and root mean square error (RMSE), according to a comparison analysis of the three regression models. Its lowest MSE (1800.61) and RMSE (42.43), which show greater predicting accuracy, were attained. Additionally, XGBoost had the best R-squared value (0.9995), indicating that it was able to account for a sizeable chunk of the variance in the sales data. Price and quantity ordered were consistently recognized by the feature importance analysis as having the greatest influence on sales across all three models. This emphasizes how important client demand and pricing strategies are in determining sales performance. On the other hand, all models largely ignored the effects of discount amount and discount percent.

**Conclusion**

In conclusion, our comprehensive analysis of the online sales dataset and the comparison of regression models for sales prediction have provided valuable insights into customer behavior, sales patterns, and predictive accuracy in the online retail industry. The absence of missing or duplicate data in our analysis ensures the reliability of our findings and allows us to make confident decisions based on the results. We observed fluctuations in sales volume over time, with peak periods coinciding with the holiday season, highlighting the significance of strategic marketing during these times. Furthermore, our examination of the product categories revealed the dominant presence of "Mobiles & Tablets" in terms of both frequency and total sales. This finding emphasizes the strong demand for these products among online shoppers. Retailers can capitalize on this insight to optimize their inventory management and align their marketing efforts to meet customer preferences. Moreover, our evaluation of payment methods indicated the popularity of options such as COD and Easypay. This finding underscores the importance for retailers to offer diverse payment options to cater to the varied preferences of their customers, thereby enhancing the overall customer experience. Additionally, our comparative analysis of regression models showcased the superior performance of the XGBoost model in terms of predictive accuracy. The XGBoost model consistently achieved the lowest mean squared error, root mean squared error, and the highest R-squared value among the evaluated models. This indicates that the XGBoost model provides the most accurate predictions for sales, enabling businesses to make informed decisions based on reliable forecasts.

Recommendations:

Based on our interpretations, we provide the following recommendations for retailers:

1. Pricing Strategy: Our analysis revealed a significant positive correlation between product price and the total value of sales. Retailers should carefully assess their pricing strategies to ensure they align with customer expectations and market trends. Consider conducting market research and competitor analysis to determine optimal pricing points that maximize profitability while remaining competitive.
2. Inventory Management: The category analysis highlighted the dominance of "Mobiles & Tablets" in terms of both frequency and total sales. Retailers should capitalize on this demand by ensuring an ample supply of popular products and monitoring inventory levels closely. Aligning inventory with customer preferences can help maximize sales potential and avoid stockouts or excess inventory.
3. Marketing and Promotions: Utilize insights from the sales patterns over time to strategically plan marketing and promotional activities. Leverage peak periods, such as the holiday season, to launch targeted campaigns and attract a larger customer base. Consider offering special discounts or promotions during off-peak periods to drive sales and maintain customer engagement throughout the year.
4. Payment Options: The analysis of payment methods revealed the popularity of options like COD and Easypay. To cater to diverse customer preferences, retailers should offer a variety of secure and convenient payment options. Consider integrating digital payment gateways and exploring partnerships with popular payment platforms to expand the range of available payment methods.
5. Predictive Analytics: Implement predictive analytics models, such as the XGBoost model, to forecast future sales accurately. Continuously update and refine the models using real-time data to improve their accuracy over time. Leverage these forecasts to optimize inventory planning, resource allocation, and overall business strategy.

**References**

* 1. Brownlee.J (Sept 2014) *Discover Feature Engineering, How to Engineer Features and How to Get Good at It* <https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/>
  2. D'Agostino (Jul 2022)  *Exploratory Data Analysis in Python — A Step-by-Step Process* <https://towardsdatascience.com/exploratory-data-analysis-in-python-a-step-by-step-process-d0dfa6bf94ee>
  3. *N.d. Online Sales in USA* <https://www.kaggle.com/datasets/ytgangster/online-sales-in-usa>