# Predictive Analytics for Customer Behavior in E-commerce
# Milestone 3 Report

OG Data Wizards

December 7, 2024

## Team Members


- Venkata Sai Sandeep Nirujogi - S564888
- Harichaithanya Kotapati - S567067
- Teja Kumar Muppala - S5565960
- Riyaz Hussain Shaik - S567058


## Abstract

The objective of this project is to leverage predictive analytics to forecast customer behavior in the e-commerce sector using the Online Retail dataset. By applying machine learning models and big data tools, we aim to enhance customer segmentation, forecast sales trends, and improve decision-making processes. This document details our implementation, results, conclusions, and references.

# 1. Implementation Steps

## 1.1 Imports and Data Preparation

In the Imports and Data Preparation step, we set up everything we needed to work with large data using PySpark. First, we created a Spark session, which helps process big data quickly and efficiently. Then, we loaded the Online Retail dataset from a CSV file, making sure it included the headers. Finally, we checked the data structure to ensure everything was correctly loaded and ready for the next steps.

```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("OnlineRetailAnalysis") \
    .getOrCreate()


file_path = "OnlineRetail.csv"
df = spark.read.option("header", "true").csv(file_path)
```

Figure 1: Implementing Data set import and prep

## 1.2 Data Cleaning and Feature Engineering

In Data Cleaning and Feature Engineering, we focused on getting the data ready for analysis by fixing any issues. We handled missing values to make sure the dataset was complete and accurate. Then, we transformed raw data into meaningful features, like calculating Customer Lifetime Value (CLV) to better understand customer worth. This step was essential for improving the accuracy of our predictive models and making the data more useful for segmentation and forecasting.

```python
from pyspark.sql.functions import col

# Drop rows with null values in essential columns
df_cleaned = df.dropna(subset=["CustomerID", "InvoiceNo", "Quantity", "UnitPrice"])

# Convert data types
df_cleaned = df_cleaned.withColumn("Quantity", col("Quantity").cast("integer"))
df_cleaned = df_cleaned.withColumn("UnitPrice", col("UnitPrice").cast("float"))

# Add Total Amount Column
df_cleaned = df_cleaned.withColumn("TotalAmount", col("Quantity") * col("UnitPrice"))
```

Figure 2: Data cleaning and handling null values.

# 2. Goals and Results

## 2.1 Customer Segmentation by Total Spending

The goal was to group customers based on their total spending to better understand their value to the business. Using PySpark, we calculated the total amount each customer spent and categorized them into three segments:

- **High Value:** Customers who spent $1,000 or more.
- **Medium Value:** Customers who spent between $500 and $999.
- **Low Value:** Customers who spent less than $500.

For example, customer 15555 spent approximately $4,758, placing them in the High Value segment, while customer 16250, who spent around $389, was classified as Low Value. This segmentation helps identify valuable customers for targeted marketing and retention strategies, while also highlighting potential growth opportunities among medium and low-value customers.

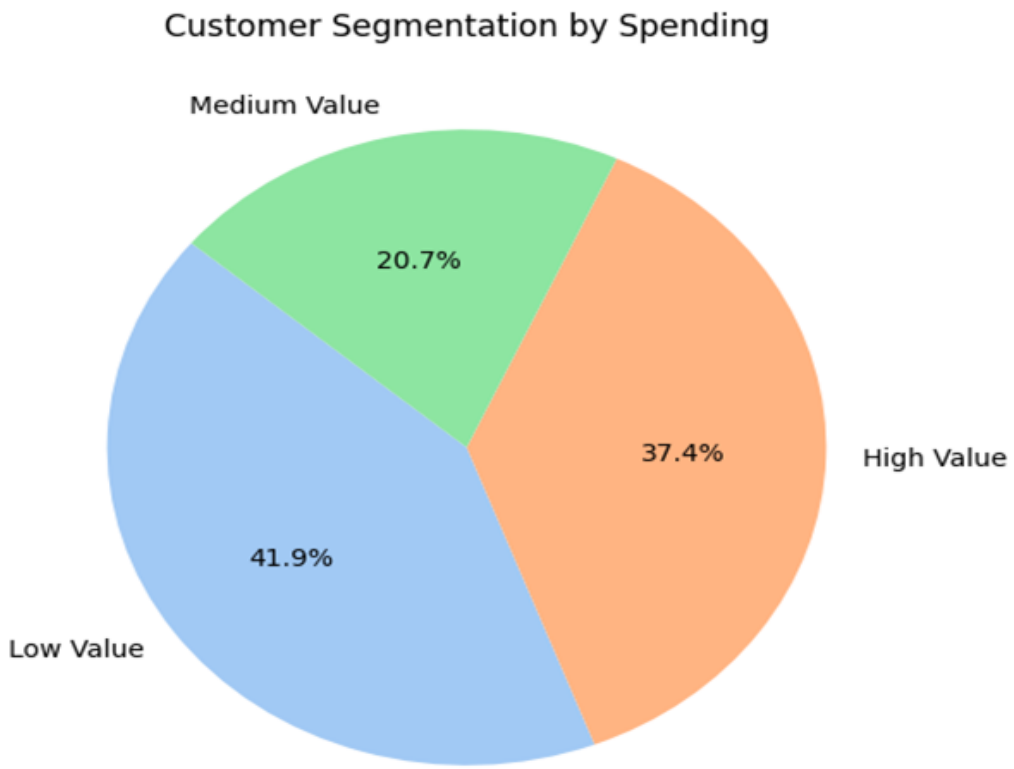


Figure 3: Customer Segmentation by Total Spending

## 2.2 Aggregate Total Sales by Product Category

The goal is to calculate the total sales for each product category, helping to identify which categories contribute the most to overall sales. By grouping the data based on product categories and summing up the sales amounts for each category, we can better understand trends and performance. This insight is valuable for making decisions on inventory management, promotional strategies, and resource allocation, as it highlights which categories are performing well and which might need more attention.

```
+--------------------+------------------+
|         Description|        TotalSales|
+--------------------+------------------+
|REGENCY CAKESTAND...|132870.39842033386|
|WHITE HANGING HEA...| 93823.85021972656|
|JUMBO BAG RED RET...| 83236.75914800167|
|       PARTY BUNTING| 67687.52895641327|
|             POSTAGE| 66710.23997282982|
|ASSORTED COLOUR B...| 56499.22161626816|
|  RABBIT NIGHT LIGHT| 51137.79873275757|
|       CHILLI LIGHTS| 45936.80920600891|
|PAPER CHAIN KIT 5...| 41500.48054909706|
|PICNIC BASKET WIC...|           39619.5|
|BLACK RECORD COVE...|39009.380932569504|
|JUMBO BAG PINK PO...| 36473.00962162018|
|      SPOTTY BUNTING| 35056.43924474716|
```

Figure 4: Aggregate Total Sales by Product Category

## 2.3 Top 10 Products with Highest Return Rates

This goal identifies the products with the highest return rates, which could signal quality issues or customer dissatisfaction. By analyzing the ratio of returns to sales, we can pinpoint the top 10 products with the most returns and take action to improve customer experience and reduce losses.

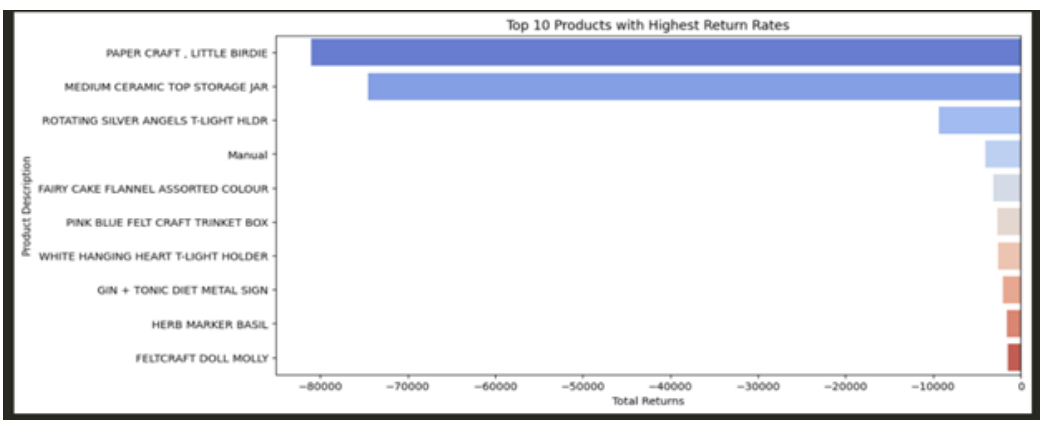


Figure 5: Top 10 Products with Highest Return Rates

## 2.4 Analyze Product Returns (Negative Quantities)

This goal focuses on understanding patterns in product returns. By analyzing which products are returned most often and the reasons behind the returns, we can identify potential quality issues, mismatches in customer expectations, or problems in the product description.

| Description | TotalReturns |
|---|---|
| PAPER CRAFT , LIT... | -80995 |
| MEDIUM CERAMIC TO... | -74494 |
| ROTATING SILVER A... | -9367 |
| Manual | -3995 |
| FAIRY CAKE FLANNE... | -3150 |
| PINK BLUE FELT CR... | -2617 |
| WHITE HANGING HEA... | -2578 |
| GIN + TONIC DIET ... | -2030 |
| HERB MARKER BASIL | -1527 |
| FELTCRAFT DOLL MOLLY | -1447 |

Figure 6: Product Returns Analysis (Negative Quantities)

## 2.5 Plot Total Sales by Product

This goal involves visualizing the total sales for each product to identify which items are the most popular and generate the highest revenue. By plotting total sales, we can easily spot trends, compare product performance, and make informed decisions about inventory management, pricing strategies, and marketing efforts. This visualization helps the business focus on top-performing products and optimize sales strategies accordingly.

```
#Plot Total Sales by Product:
import matplotlib.pyplot as plt

plt.figure(figsize=(9, 5))
plt.bar(total_sales_pd["Description"][:10], total_sales_pd["TotalSales"][:10])
plt.title("Top 10 Products by Total Sales")
plt.xlabel("Product Description")
plt.ylabel("Total Sales")
plt.xticks(rotation=45)
plt.show()
```

Figure 7: Total Sales by Product

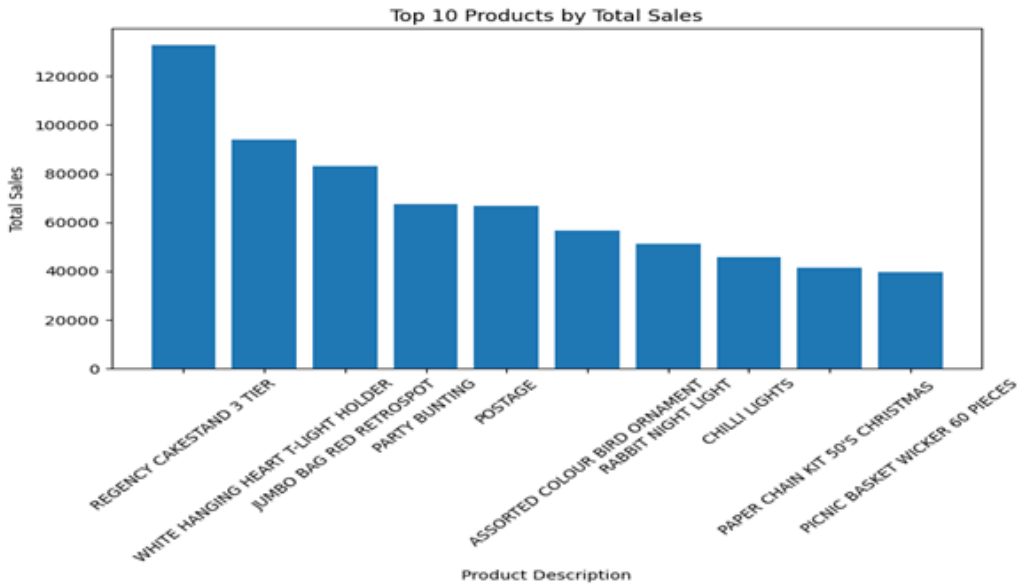


Figure 8: Additional Product Sales Visualization

# 3. Conclusion

In this project, we successfully analyzed customer behavior and sales trends using predictive analytics. By segmenting customers based on spending, we identified high-value customers for targeted marketing. We also explored product performance, identifying top-selling items and those with high return rates, helping improve product offerings. Analyzing product returns highlighted areas for improvement in quality or customer expectations. Overall, the insights gained from this project can help optimize sales strategies, improve customer satisfaction, and guide better decision-making for the business.

# 4. References


- UCI Machine Learning Repository - Online Retail Dataset: `https://archive.ics.uci.edu/ml/datasets/online+retail`
- Apache Spark Documentation: `https://spark.apache.org/docs/latest/`

- Seaborn Documentation: `https://seaborn.pydata.org/`
- Jupyter Book: `https://docs.jupyter.org/`
- Matplotlib: `https://matplotlib.org/`

**GitHub Repo Link:** BigData Project.