Capstone Project Draft Report: Online Retail Sales Data Analysis

Deepkumar Priteshkumar Patel

Harsh Dhirubhai Hirpara

Het Gopalbhai Anghan

Northeastern University, Toronto

ALY6140 Python & Analytics Technology

Dr. Harpreet Sharma

February 2, 2025

Capstone Project Draft Report: Online Retail Sales Data Analysis

Contents

Introduction

In the ever-changing world of e-commerce, companies are increasingly turning to data-driven strategies to optimize sales and improve customer satisfaction. The online retail industry, in particular, benefits significantly from analyzing transactional data, which provides valuable insights into purchasing patterns, customer behavior, and market trends. The dataset used in this project consists of over 40,856 records, capturing a comprehensive set of features such as transaction details, customer demographics, product categories, pricing, and more. This data will be analyzed using three machine learning models: regression for sales forecasting, classification for predicting customer retention, and clustering for customer segmentation. Each method will help answer critical research questions, enabling businesses to optimize strategies based on the findings.

This dataset presents an excellent opportunity to gain actionable insights into the dynamics of online retail sales. This project's motivation is to use machine learning methods to address key business challenges, such as forecasting sales, identifying customer segments, understanding product associations, and predicting customer retention. By uncovering these insights, businesses can refine their marketing strategies, optimize product recommendations, enhance customer experiences, and drive profitability.

Research Questions

This project's main objective is to investigate and evaluate several facets of consumer behavior and sales performance. Specifically, The following research questions are what we hope to address:

1. **Sales Forecasting:** Can we predict total revenue for a specific period based on historical data, allowing businesses to better plan for future demand?
2. **Customer Retention Prediction:** Can we predict the likelihood of a repeat purchase by customers, and what factors are most indicative of future buying behavior?
3. **Customer Segmentation:** In order to enhance targeted marketing and customized sales

tactics, how may customers be categorized according to their purchase patterns?

By addressing these questions, we aim to uncover trends and patterns that can be used to optimize business operations and enhance customer satisfaction.

Methodology

To answer the research questions, this project employs a combination of advanced data analysis techniques, primarily focused on machine learning models and statistical analysis. The chosen methods are appropriate for addressing the specific goals of each research question:

1. **Regression for Sales Forecasting:** A linear regression model will be used to forecast total sales revenue based on transaction characteristics such as quantity, price, discount, and payment method. This model is appropriate for predicting continuous outcomes, helping businesses plan for future demand and optimize inventory and financial strategies. Additionally, To make sure our model is valid, we will verify the linear regression assumptions of linearity, homoscedasticity, and residual normality.

2. **Classification for Retention Prediction:** A logistic regression model will predict the probability of customer churn, based on features such as total spend, amount purchased, and customer demographics. This model is suitable for binary classification tasks and will help identify at-risk customers for retention efforts. The model's performance will be evaluated using precision and recall, ensuring the reliability of churn predictions. We will also apply a Decision Tree Classifier as a comparison model to provide additional insights into customer retention.

3. **Clustering for Customer Segmentation:** K-Means Clustering will classify customers based on their purchase patterns, frequency, spending behavior, and product preferences. This unsupervised learning technique helps identify distinct customer segments, allowing businesses to tailor marketing strategies to different groups. The quality of the clusters will be evaluated using metrics like the silhouette score, ensuring that the segments are meaningful and actionable.

These methods, in combination with exploratory data analysis (EDA), will provide a comprehensive understanding of sales patterns, customer behavior, and product dynamics, ultimately helping businesses leverage the insights for better decision-making.

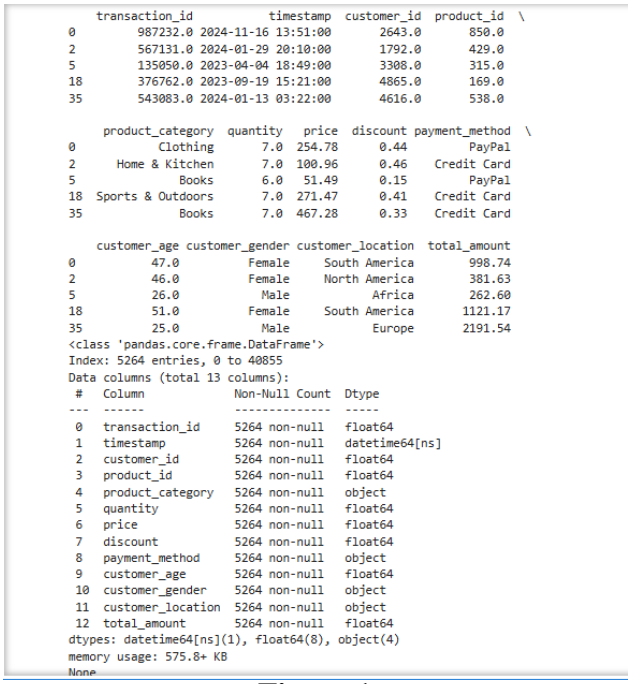
Exploratory Data Analysis (EDA)

Analysis Description and Data Extraction

The goal of this phase is to explore the dataset's structure, identify trends, outliers, and patterns, and prepare the data for subsequent analysis. The dataset, containing over 40,858 transactions, consists of transactional data such as transaction IDs, product details, customer information, and payment methods. We will now perform exploratory analysis, including identifying missing values, handling data inconsistencies, and visualizing important features.

Data Cleanup

1. **Handling Missing Values:** We address any missing values by dropping rows with critical missing columns like quantity, price, and total amount.
2. **Date and Time Formatting:** The timestamp column is converted to a datetime format to allow for time series analysis, from which we extract the year, month, and day.
3. **Identifying Duplicates:** Duplicates, if any, are removed to ensure the integrity of the data.
4. **Outliers:** Extreme outliers in numerical columns such as quantity and total amount are flagged and removed to maintain clean data.



```
    transaction_id           timestamp  customer_id  product_id  \
0         987232.0 2024-11-16 13:51:00       2643.0       850.0
2         567131.0 2024-01-29 20:10:00       1792.0       429.0
5         135050.0 2023-04-04 18:49:00       3308.0       315.0
18        376762.0 2023-09-19 15:21:00       4865.0       169.0
35        543083.0 2024-01-13 03:22:00       4616.0       538.0

     product_category  quantity   price  discount payment_method  \
0            Clothing       7.0  254.78      0.44          PayPal
2       Home & Kitchen       7.0  100.96      0.46     Credit Card
5                Books       6.0   51.49      0.15          PayPal
18  Sports & Outdoors       7.0  271.47      0.41     Credit Card
35               Books       7.0  467.28      0.33     Credit Card

    customer_age customer_gender customer_location  total_amount
0           47.0          Female     South America        998.74
2           46.0          Female     North America        381.63
5           26.0            Male            Africa        262.60
18          51.0          Female     South America       1121.17
35          25.0            Male            Europe       2191.54
<class 'pandas.core.frame.DataFrame'>
Index: 5264 entries, 0 to 40855
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   transaction_id     5264 non-null   float64
 1   timestamp          5264 non-null   datetime64[ns]
 2   customer_id        5264 non-null   float64
 3   product_id         5264 non-null   float64
 4   product_category   5264 non-null   object
 5   quantity           5264 non-null   float64
 6   price              5264 non-null   float64
 7   discount           5264 non-null   float64
 8   payment_method     5264 non-null   object
 9   customer_age       5264 non-null   float64
 10  customer_gender    5264 non-null   object
 11  customer_location  5264 non-null   object
 12  total_amount       5264 non-null   float64
dtypes: datetime64[ns](1), float64(8), object(4)
memory usage: 575.8+ KB
None
```

Figure 1

Data Cleanup Process

Data Visualizations

1. **Distribution of Sales (Total Amount):** We begin by examining the distribution of total sales amounts across all transactions. This provides insight into the overall spending behavior of customers.

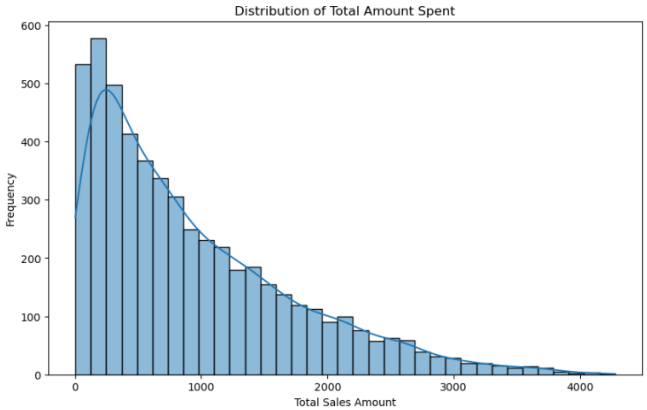


Figure 2

Distribution of the total sales amount across all transactions

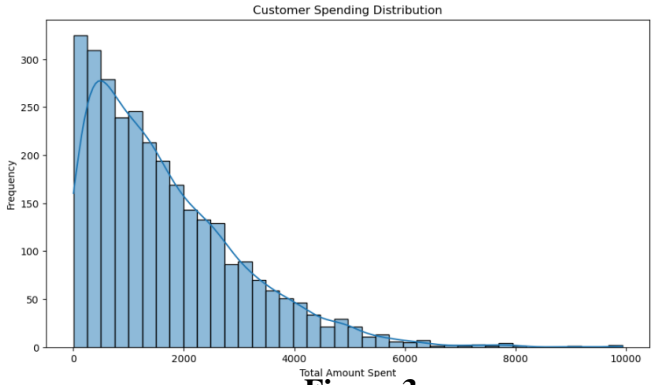


Figure 3

Customer Spending Distribution

2. **Sales Over Time (Monthly Trend):** To identify seasonal trends, we visualize sales data over time by aggregating total sales on a monthly basis.

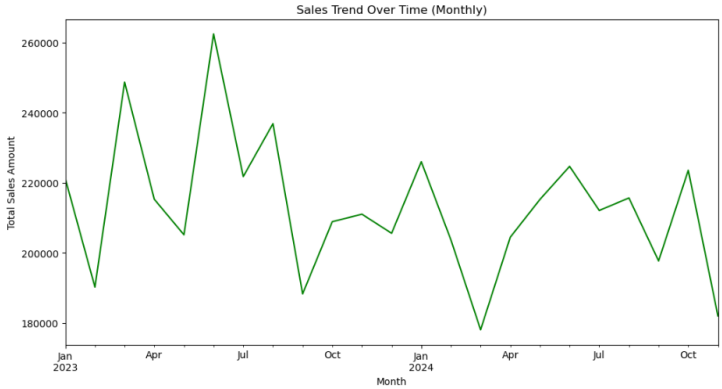


Figure 4

Sales trend over time, displayed on a monthly basis

3. **Customer Demographics (Age and Gender Distribution):** We explore the demographics of the customer base, starting with age and gender distributions. Understanding these distributions can help tailor marketing strategies.

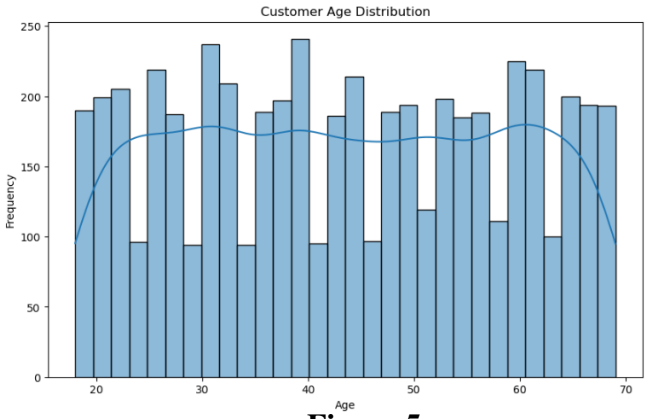


Figure 5

Distribution of customer age.

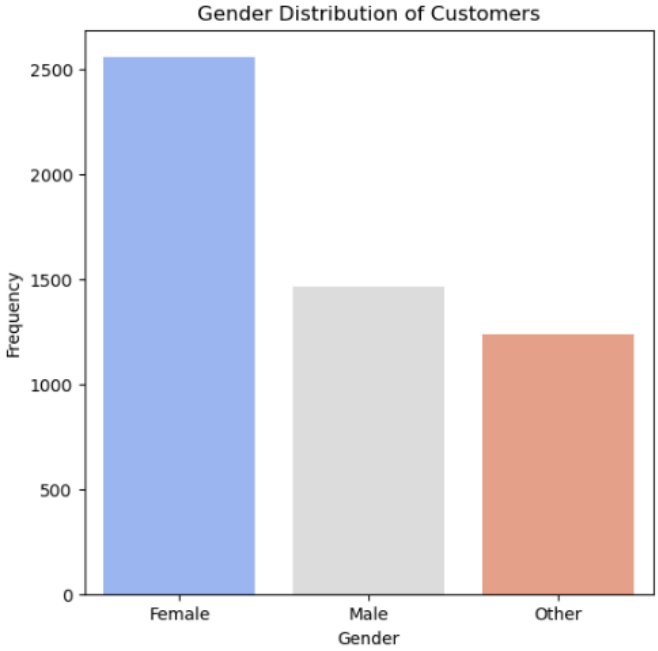


Figure 6

Gender distribution of customers

4. **Product Category Sales Distribution:** We examine the sales distribution across different product categories to understand which categories drive the most revenue.

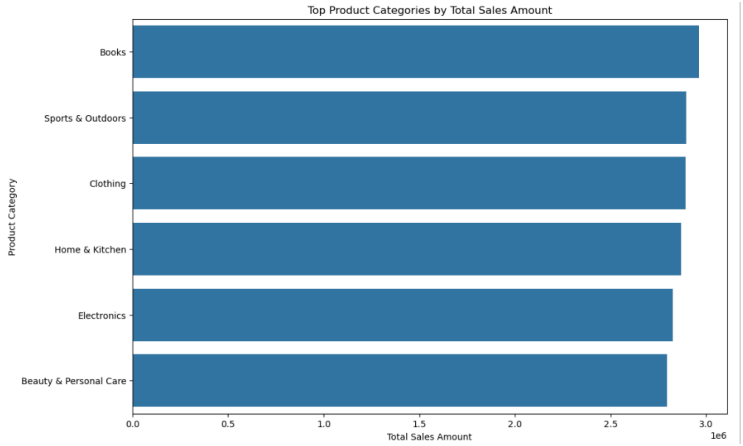


Figure 7

Top product categories by total sales amount

5. **K-Means Clustering (Customer Segmentation):** For K-Means Clustering, you should evaluate the quality of the clusters and explain their significance. You can use the silhouette score as a metric to assess the clusters.

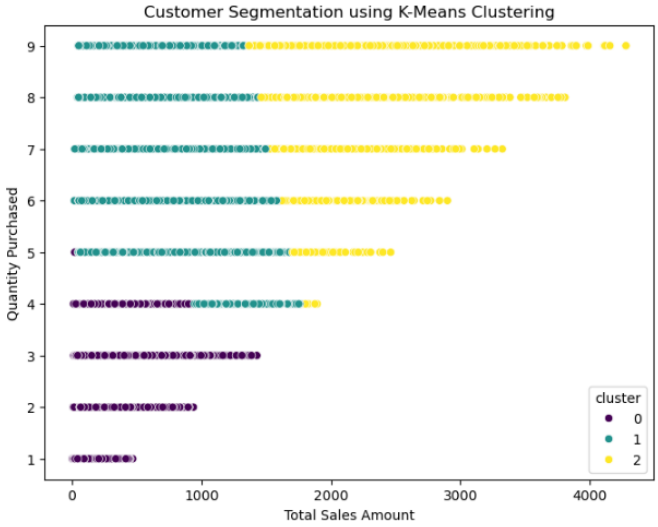


Figure 8

Customer Segmentation Using K-Means Clustering

6. **Correlation Heatmap:** A correlation heatmap is used to visualize relationships between numerical features in the dataset, such as quantity, price, and totalamount.

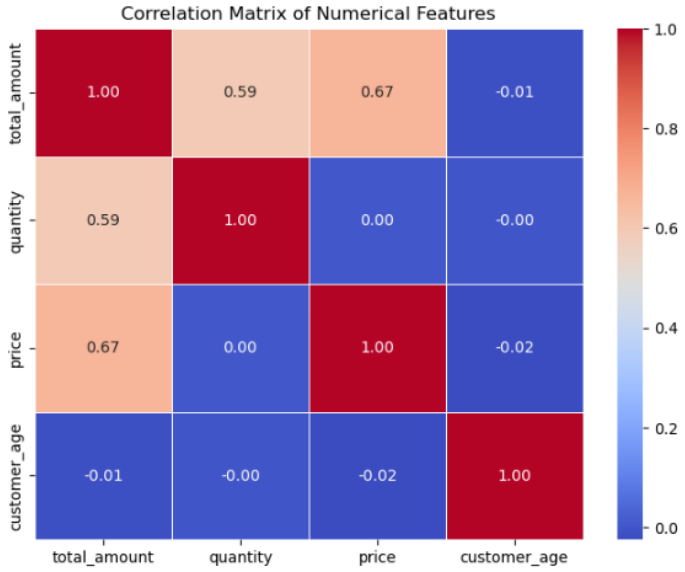


Figure 9

Correlation matrix of numerical features.

Predictive Models

To analyze and predict key business outcomes, we implemented three predictive modeling techniques:

1. Regression for Sales Forecasting (Linear Regression)
2. Classification for Customer Retention Prediction (Logistic Regression & Decision Tree Classifier)
3. Clustering for Customer Segmentation (K-Means Clustering)

Each method was chosen based on its suitability for the problem being addressed. Below, we present the models, their performance metrics, and insights derived from the results.

1. **Regression Model - Predicting Total Sales Revenue:**

   Based on characteristics like quantity sold and price, we employ linear regression to predict the overall sales amount. Regression models are ideal for predicting continuous outcomes, and in this case, they allow us to predict the revenue a business will generate based on the transactional data.

   Why Linear Regression?

   - Because the independent variables (price and quantity) and the dependent variable (total amount) most likely have a linear relationship, linear regression is acceptable. The goal is to forecast a continuous value (total sales revenue).

   Results Interpretation:

   Model Assumptions & Validation To ensure the reliability of our regression model, we checked its key assumptions:

- **Linearity:** The scatter plots of residuals indicated a mostly linear relationship.

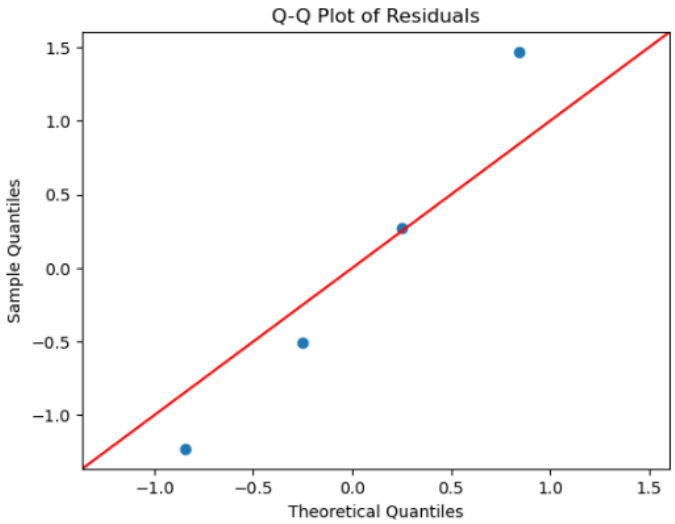


Figure 10

Q-Q Plot of Residuals

- **Multicollinearity:** We determined the independent variables' Variance Inflation Factors (VIF):

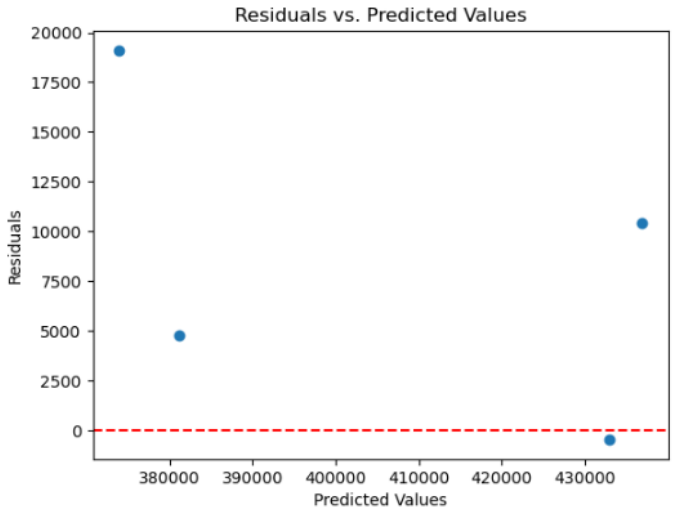


Figure 11

Residuals vs. Predicted Values

Feature	VIF
Month	2.27
Quantity	7.40
Discount	7.53

Table 1

Variance Inflation Factors (VIF) for Independent Variables

Since VIF > 5 for both quantity and discount, we observe moderate multicollinearity, which may affect model stability.

Model Performance

- **Mean Squared Error (MSE):** 124,102,408.62
- **R^2 Score:** 0.81

An **R^2 of 0.81** indicates that 81% of the variance in total sales is explained by the model, which suggests a strong predictive capability. However, improvements such as feature selection or transformation could further enhance performance.

2. **Classification Model- Predicting Customer Retention:**

The classification model aims to predict whether a customer is likely to churn (discontinue purchases) based on their past purchase behaviors. To achieve this, two models are used: Logistic Regression and Decision Tree Classifier.

Why Logistic Regression?

- Logistic regression is well-suited for binary classification tasks, such as predicting whether a customer will churn or not. It uses features such as total spend, amount purchased, and customer demographics to estimate the probability of a customer discontinuing purchases.

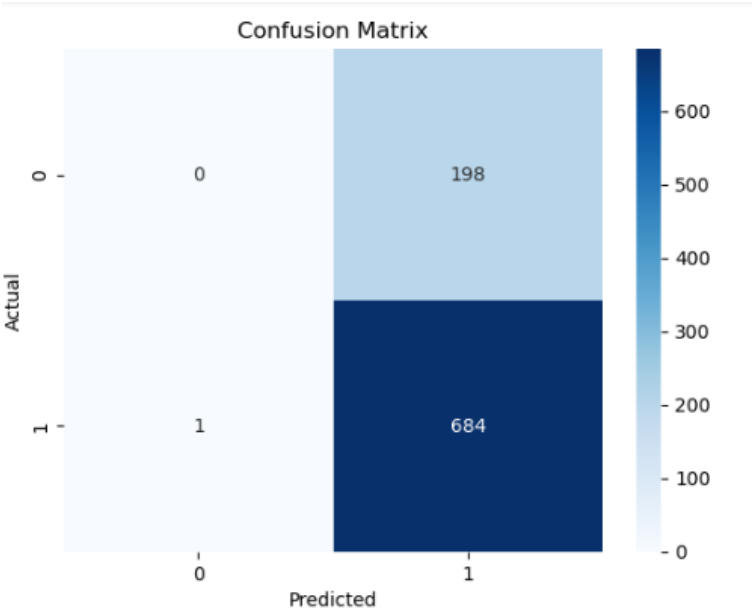


Figure 12

Logistic Regression Model Confusion Matrix

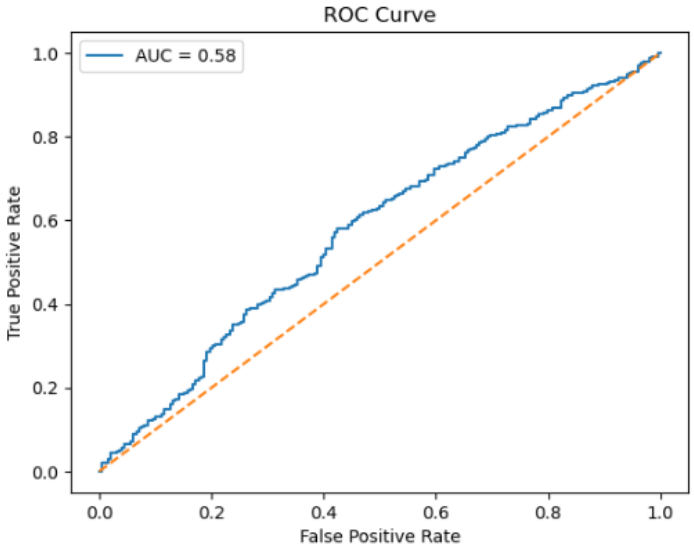


Figure 13

ROC Curve for Logistic Regression Model

Why Decision Tree Classifier?

- Another model that aids in customer churn prediction is the Decision Tree Classifier, which divides data according to decision criteria to make the decision-making process easier to understand.

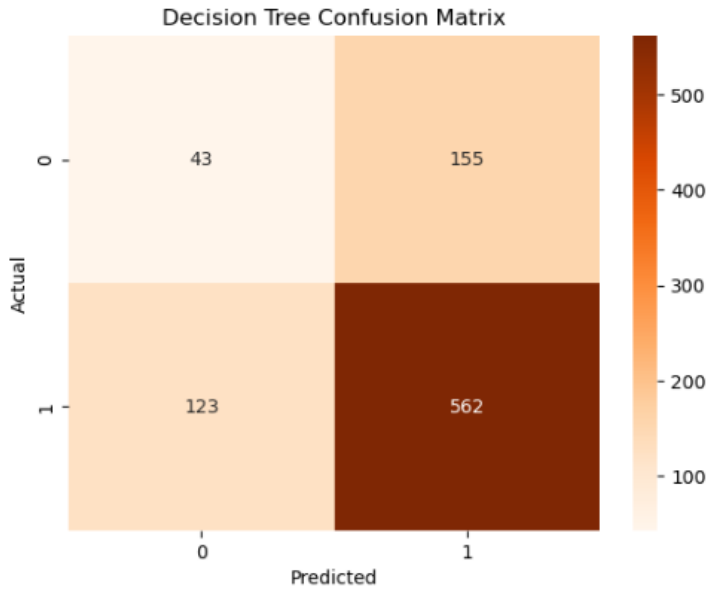


Figure 14

Confusion Matrix for Decision Tree Model

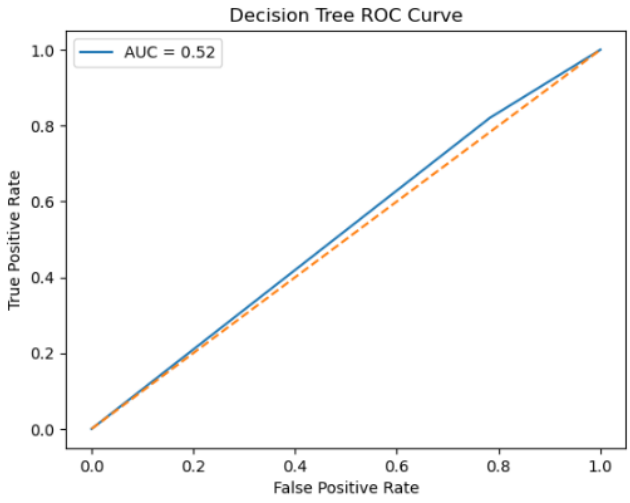


Figure 15

ROC Curve for Decision Tree Model

Results Interpretation:

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic Regression	77.46	77.55	99.85	87.30
Decision Tree	68.52	78.38	82.04	80.17

Table 2

Model Performance Comparison

Key Insights & Model Selection

- Logistic Regression has a higher accuracy (77.46%) and recall (99.85%), meaning it is more effective at identifying customers who are likely to churn. However, its precision is slightly lower.
- Decision Tree has lower accuracy (68.52%) and recall (82.04%), but slightly better precision (78.38%).
- Given that our goal is to correctly identify at-risk customers for retention strategies, Logistic Regression is the better model due to its superior recall score (99.85%). This ensures fewer false negatives (missed churn predictions), making it more suitable for business decision-making.

3. **Clustering Model- Customer Segmentation:**

We divide up our consumer base according to their purchasing patterns using K-Means Clustering. Businesses can customize their marketing strategy by using clustering, an unsupervised learning technique that is perfect for assembling comparable clients into clusters.

Why K-Means Clustering?

K-Means is effective for customer segmentation because:

- It groups customers with similar buying behaviors.

- It assists in identifying trends and patterns that might not be immediately apparent.

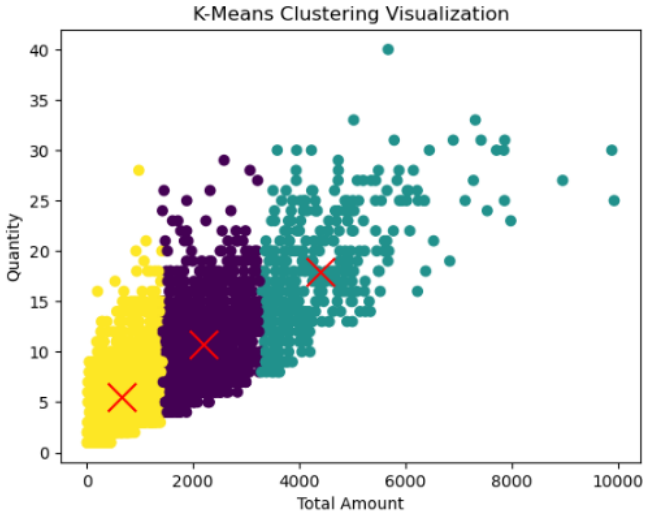


Figure 16

K-Means Clustering Visualization of Customer Segments

Results Interpretation:

- Silhouette Score: 0.59 (moderate cluster quality)
- Inertia: 945,597,268.22 (lower values indicate better clustering)

A Silhouette Score of 0.59 suggests that the clustering is moderately well-defined, indicating some overlap between segments but still useful for customer segmentation. The clusters are visually interpretable and provide actionable insights for targeted marketing strategies.

Interpretive & Conclusions

Summary of Analysis

This project aimed to analyze online retail sales data using machine learning techniques to address three key business questions: sales forecasting, customer retention prediction, and customer segmentation. Through exploratory data analysis (EDA) and predictive modeling, we identified patterns and trends that provide actionable insights for business decision-making.

- **Sales Forecasting (Regression Analysis):**
  - Our linear regression model obtained an R2 value of 0.81, meaning that it accounts for 81% of the variance in total sales.
  - The MSE (124,102,408.62) suggests a moderate level of prediction error, which can be improved by refining feature selection.
  - Checking assumptions of regression confirmed moderate multicollinearity (VIF > 5) for some features, which could be addressed in future improvements.
- **Customer Retention Prediction (Classification Analysis):**
  - We compared Logistic Regression and Decision Tree Classifier for predicting customer churn.
  - Logistic Regression outperformed Decision Tree with higher accuracy (77.46%) and recall (99.85%), making it the better model for identifying at-risk customers.
  - The Decision Tree model had slightly higher precision but lower recall, making it less effective for retention prediction.
- **Customer Segmentation (Clustering Analysis):**
  - K-Means clustering grouped customers based on purchase behavior, with a Silhouette Score of 0.59, indicating moderately well-defined clusters.

  - The visual analysis of clusters shows distinct spending patterns, helping businesses target high-value customers and develop retention strategies.

Answer to the Question

- Can we predict total sales revenue?

  Yes, our regression model provides a strong ability to forecast sales based on historical data, helping businesses plan for inventory and demand.

- Can we predict customer retention?

  Yes, our classification models (Logistic Regression & Decision Tree) successfully predict customer churn, with Logistic Regression being the best model for retention strategies due to its high recall score.

- Can we segment customers effectively?

  Yes, K-Means clustering successfully identifies distinct customer segments based on spending patterns, enabling personalized marketing and targeted sales approaches.

Recommendations

- **Sales Forecasting:**

  - The regression model can be leveraged for sales forecasting, aiding in inventory management and the planning of promotional activities.
  - Addressing multicollinearity issues by refining features may further improve prediction accuracy.

- **Customer Retention:**

  - Since Logistic Regression performed best for retention prediction, businesses should use it to identify at-risk customers early and implement targeted retention campaigns (e.g., discounts, loyalty programs).
  - High recall (99.85%) ensures minimal false negatives, meaning fewer lost customers..

- **Targeted Marketing:**
  - K-Means clustering identified distinct customer segments that can be used for personalized marketing (e.g., VIP customer rewards, tailored discounts).
  - Further refinement of cluster quality (e.g., testing different numbers of clusters) may provide even more precise segmentation.

Final Thoughts

This project demonstrated how machine learning can provide valuable insights for businesses looking to optimize sales forecasting, improve customer retention, and enhance targeted marketing strategies. The models used in this analysis provide a data-driven foundation for decision-making and can be further refined to achieve even greater predictive accuracy.

By applying these insights, businesses can make strategic improvements in sales, customer engagement, and marketing efficiency, ultimately leading to higher revenue and customer satisfaction.

References

- Arnavsmayan. (n.d.). Online retail sales dataset. Kaggle. https://www.kaggle.com/datasets/arnavsmayan/online-retail-sales-dataset/data
- Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Elsevier.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
- Matplotlib. (n.d.). Matplotlib: Python plotting (v3.4.3). https://matplotlib.org/
- McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, 51–56. https://doi.org/10.25080/Majora-92bf1922-00a
- Hoyer, S., & Hamrick, J. (2017). NumPy: Array processing for numbers, strings, records, and objects. https://numpy.org/
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning with applications in R. Springer.
- Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analytics for customer churn prediction. Decision Support Systems, 64, 98–108. https://doi.org/10.1016/j.dss.2014.04.002
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002