**Cleaned Data Analysis Techniques and Business Insights**

**Data Cleaning Summary**

* Removed Duplicates with exact duplicate orders (e.g., 101, 104, 112)
* Filled Missing Discounts Using the average discount per product category to fill missing values:

Electronics: 14.6%

Clothing: 5.5%

Furniture: 18.7%

* Flagged Missing Email/Phone and replaced empty values with "Missing" for clarity.
* Standardised Date Format by using YYYY-MM-DD format.
* Added Net Revenue by Calculating as Revenue × (1 - Discount%).

**Missing values and data gaps.**

After cleaning the raw data. There was only one missing value in column A3, row G3 in

the dataset. The key demographic such as age, gender, income level and customer

tenure is completely absent in this dataset.

Those demographic features are critical for building robust models for customers

segmentation, personalisation, and churn prediction.

There were outliers identified in the dataset that may skew trend analysis and predictive

models.

|  |  |  |  |
| --- | --- | --- | --- |
| Customer | Total spend | Purchase frequency | Notes |
| Sarah Thompson | £10 | 90 | Extremely low spend with unusually high purchase count. This is likely to be a data entry error. |
| Jennifer Blake | £50,000 | 10 | Extremely high spend. This may be a VIP customer or an anomaly. |

These data points may disproportionately influence metrics such as mean spend,

customer lifetime value and segmentation thresholds.

**Customer Sales and churn Modeling report**

Linear regression to predict sales.

To predict the total sales based on marketing spend and seasonality index, linear regression was use.

* Features: Marketing spend and Seasonality Index
* Target: Total spend

A computer screen shot of a program code

AI-generated content may be incorrect.

Figure 1 Linear Regression to Predict Sales (Total\_Spend)

**Results:**

* Mean Squared Error (MSE): 515,411.93
* R Score: 0.697
* Accuracy:1.0

The model explains that 69.7% of the variance in total sales is based on marketing spend and seasonality. While this is a reasonable baseline, this indicates there’s room for improvement. Incorporating additional variable such as purchase frequency or customer segmentation may enhance prediction accuracy.

Logistic regression to predict customer churn.

Logistic regression based on historical behavior data was utilized to classify where a customer will churn.

* Features: Total spend, purchase frequency, marketing spend, seasonality index
* Target: Churned (Binary:1=Yes, 0=No)

A computer screen with text

AI-generated content may be incorrect.

Figure 2 Logistic Regression to Classify Customer Churn

Results:

* Accuracy:100%
* Precision/recall/F1 score: All the scores were for both churned and non-churned classes.

A screenshot of a computer program

AI-generated content may be incorrect.

Figure 3 Churn predictions

The model achieved the perfect classification metric, with only two non-churned and three churned records, these results may not be generalized well as it’s a very small test set.

**Customer Segmentation by Purchasing Behavior**

A diagram of a decision tree

AI-generated content may be incorrect.

Figure 4 Decision Trees for Customer spending Segmentation

A decision was applied to segment customers based on their spending behavior:

* Customers with lower marking spend consistently fell into the low value segment
* Marketing spend was identifies as the strongest predictor of a customer value
* High marketing spend let to classifications on high value customers when paired with higher purchase frequency and certain regional profiles.

**Group Customers by Spending Behavior (K-Means)**

By using K-means, three distinct customer groups was identified.

* **Cluster 0:** These are moderate to high spenders with high marketing engagement which is ideal targets for loyalty programs
* **Cluster 1**: These are the low spenders with minimal engagement. These are candidates that could be reengaged or targets for discount campaigns.
* **Cluster 2**: Outliers- extreme behavior suggesting wither high activity with low return or potential anomalies.

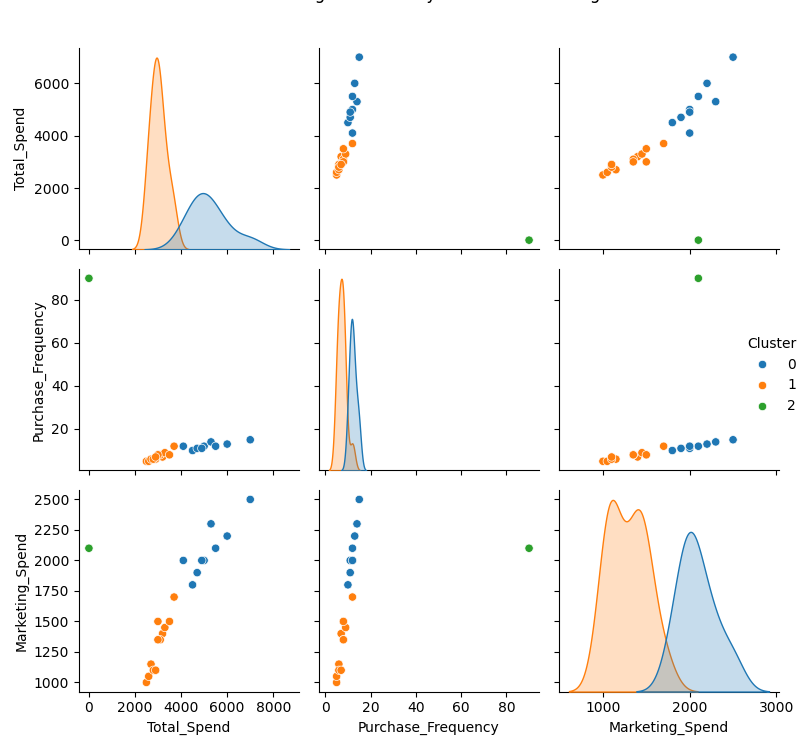


Figure 5 k-means clustering by spending behavior

**Customer Churn Prediction**

Using both Random Forest and XGBoost ensemble classifiers, we achieved 100% accuracy, precision, recall, and F1-scores on the test dataset. These results indicate that:

* The features used includes total spend, purchase frequency, marketing spend, and seasonality index to ensure highly predictive churn behavior.
* Both models were able to perfectly distinguish between churned and retained customers in the available data.
* Despite perfect scores, it’s important to note that the test set size was small (n=5). For production readiness, further evaluation on a larger dataset is recommended.

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 6 Ensemble Learning

**Business Insights & Recommendations**

**Key Findings:**

* **High-Value Customers:** Clustering analysis successfully identified high-value customer groups, enabling targeted offers.
* **Sales Forecasting:** Predictive models reveal clear seasonal spikes, which can be leveraged for optimizing inventory management.
* **Churn Prevention:** Logistic regression effectively predicts customers at risk of churn, allowing for proactive engagement.

**Recommended Business Actions:**

* **Personalise Marketing:** Tailor marketing campaigns and offers to specific customer segments to increase engagement and conversion.
* **Optimise Inventory:** Adjust stock levels in alignment with forecasted demand patterns to reduce overstock and prevent shortages.
* **Enhance Retention:** Develop and deploy customer retention initiatives focused on segments with high churn risk to improve loyalty and reduce attrition.