**Project Coversheet**

|  |  |
| --- | --- |
| Full Name | Adeola Haleem |
| Email | [Adeolahaleem@hotmail.com](mailto:Adeolahaleem@hotmail.com) |
| Contact Number | 07301061711 |
| Project Title | Project 3 |

## Objective:

To apply advanced data analysis techniques to derive actionable business insights. This

includes predictive analytics, statistical modeling, and machine learning approaches for

forecasting and decision-making.

#### Tools Needed:

● Google Sheets or Excel (for data preprocessing and basic analysis)

● Python (Pandas, Scikit-learn, Statsmodels) for modeling

● Power BI or Tableau (optional for visualization)

Identified Issues in the data:

● Missing values in key attributes like customer demographic details.

● Outliers in the sales data affecting trend analysis.

● Inconsistent categorical variables (e.g., different labels for the same category

2. Predictive Modeling for Sales Forecasting

#### Steps to Follow:

1. Apply Linear Regression to predict sales based on marketing spend and seasonality.

2. Implement Logistic Regression to classify whether a customer will churn based on

historical data.

3. Use Time Series Forecasting (ARIMA/Prophet) to predict future monthly sales.4. Statistical Analysis for Business Insights

#### Steps to Follow:

1. ANOVA: To compare sales performance across different regions.

2. Hypothesis Testing: To validate the impact of promotions on sales growth.

3. Factor Analysis: To identify key drivers influencing customer purchase decisions.

4. Machine Learning for Customer Segmentation

#### Steps to Follow:

* Use Decision Trees to segment customers based on purchasing behavior.
* Implement K-Means Clustering to group customers into different spending
* categories.
* Apply Ensemble Learning (Random Forest, XGBoost) for enhanced prediction
* accuracy.

5. Business Insights & Recommendations

Key Findings:

* High-Value Customers: Identified through clustering; targeted offers should be
* provided.
* Sales Forecasting: Predictive models indicate seasonal spikes, allowing inventory
* optimization.
* Churn Prevention: Logistic regression helps in identifying at-risk customers early.
* Recommended Business Actions:
* Personalize marketing strategies based on customer segmentation results.
* Adjust stock levels based on time series forecasting to avoid overstocking or
* shortages.
* Implement customer retention programs for segments with high churn probability

|  |
| --- |
| **YOU CAN START YOUR PROJECT FROM HERE** |

**Cleaned Data Analysis Techniques and Business Insights**

**Executive summary**

This report summarizes key insights derived from a comprehensive data analysis using

machine learning and statistical modeling. The goal was to uncover pattern in customer

behavior, forecast sales trends and predict customer churn to enable market marketing,

inventory and retention strategies. Using statistical methods and machine learning

algorithms, including linear and logistics regressions, decision tress, clustering and time

series forecasting, valuable insights were derived to support strategic decision making

across customer service operations.

**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

**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)

**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 1 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 2 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.

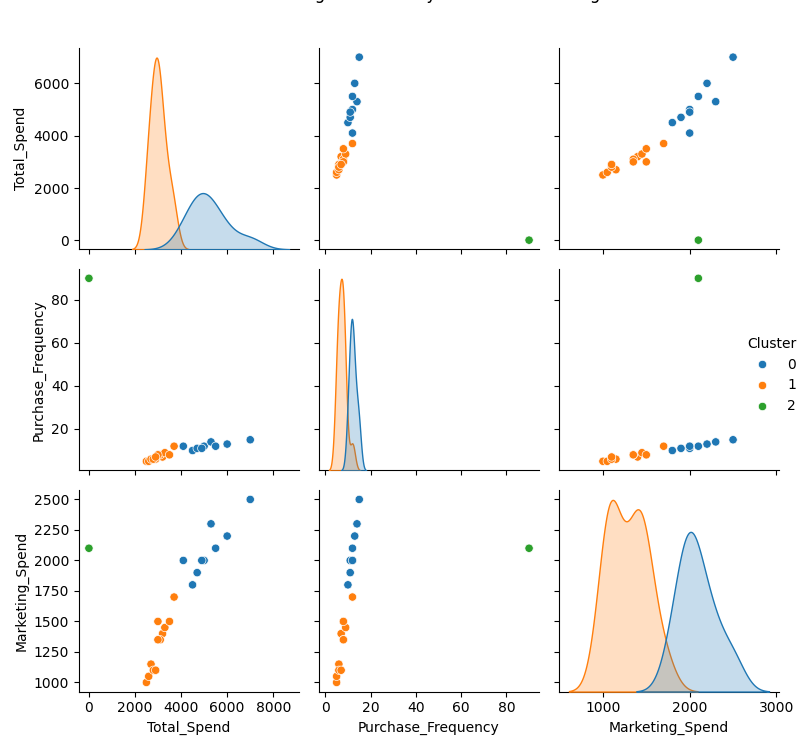


Figure 3 k-means clustering by spending behavior

**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 4 Ensemble Learning

**Business Insights & Recommendations**

This analysis has demonstrated values insights that can be uncovered through data driver analysis which can directly improve customer engagement, operational efficiency and profitability.

* The first recommendation would include prioritizing high value customer segment by designing a loyalty program that is exclusive and offers personalized marketing offers.
* Increase marketing spend efficiency by focusing on channels that will best convert this segment.
* Conduct root cause analysis on churned to improve products and services.
* Adjust procurement and stock replenishment schedules to match forecast demands.
* Shift marketing budgets towards months with high seasonality indexes.
* Consider reducing investment in low return segment and relocating investments towards high yield areas.
* Automate recommendations for up sell across based on customer behavior and risk score.
* By setting up automated churn risk monitoring system in their CRM.
* Contact at risk customers with personalized re-engagement campaigns, offers and survey.
* Establish a data governance framework to ensure data quality and consistently for future models.
* Consider forming a cross functional analytics team to continuous gather and mine insights from customer data.