Project 3

Advanced Techniques

1 – Data Transformation with Excel

The provided data contains a small dataset describing the lifecycle of recorded customers. A total of 8 fields (including ID) and 16 records are present, with no apparent duplicates when aggregated by **Customer_Name**, nor any unknown values.

Customer_ID	Customer_Name	Region	Total_Spend	Purchase_Frequency	Marketing_Spend	Seasonality_Index	Churned
101	John Doe	North	5000	12	2000	1.2	No
102	Jane Smith	South	3000	8	1500	1	Yes
103	Sam Brown	East	4500	10	1800	1.1	No
104	Linda Johnson	West	2500	5	1000	0.9	Yes
105	Michael Lee	North	7000	15	2500	1.3	No
106	Emily Davis	South	3200	7	1400	1	Yes
107	David Wilson	East	5300	14	2300	1.2	No
108	Susan White	West	2900	6	1100	0.8	Yes
109	Chris Martin	North	6000	13	2200	1.2	No
110	Anna Taylor	South	3100	8	1350	0.9	Yes
111	James Anderson	East	4700	11	1900	1.1	No
112	Patricia Thomas	West	2600	5	1050	0.8	Yes
113	Robert Jackson	North	5500	12	2100	1.2	No
114	Mary Harris	South	3300	9	1450	1	Yes
115	Daniel Clark	East	4900	11	2000	1.1	No
116	Barbara Lewis	West	2700	6	1150	0.9	Yes

Figure 1 – Unprocessed data in tabular format loaded in Excel.

The first three columns are straightforward; however, the subsequent columns have room for interpretation. I've defined them as the following:

- **Total_Spend** – The cumulative sales made from the customer over their lifecycle.
- **Purchase_Frequency** – The total count of orders the customer has made over their lifecycle.
- **Marketing_Spend** – The total cost the business has spent to acquire or retain the customer over their lifecycle.
- **Seasonality_Index** – The deviation of spending from the average amount over a time-series. As there is no time-series, this is most likely an average over the customer lifecycle.
- **Churned** – Is the customer still actively using services from the business?

Based on the available data, we can derive the '**Customer_Lifetime_Value**' (CLV) to categorically identify high and low value customers. With this, we can identify some patterns and conditions of variables to determine methods that can artificially induce high value customers for the business, or suggestions to increase customer value in general.

The data accounts for both **churned and active customers**, which is something to be considered when calculating CLV – therefore a heuristic value will be used to make value projections of active customers; to give a rough estimate of their CLV.

CLV Methodology -

- Churned Customers:
  - CLV = **[Total_Spend] – [Marketing_Spend]**
- Active Customers:
  - **Heuristic Multiplier = 1.5**
    - Conservative value between optimistic (**2.0**), and pessimistic (**1.1**).
  - Average_Purchase_Value = **[Total_Spend] * [Purchase_Frequency]**
  - Projected_Purchases = **[Purchase_Frequency] * Multiplier**
  - Projected_Spend = **[Marketing_Spend] * Multiplier**
  - CLV = **{[Average_Purchase_Value] * [Projected_Purchases] * [Seasonality_Index]} - [Projected_Spend]**

Customer_ID	Customer_Name	Region	Total_Spend	Purchase_Frequency	Marketing_Spend	Seasonality_Index	Churned	Customer_Value	Average_Purchase_Value	Projected_Purchases	Projected_Spend	Customer_Lifetime_Value	Value_Category
101	John Doe	North	£5,000.00	12	£2,000.00	1.2	No	£3,000.00	£416.67	18.0	£3,000.00	£6,000.00	Mid
102	Jane Smith	South	£3,000.00	8	£1,500.00	1.0	Yes	£1,500.00	£375.00	8.0	£1,500.00	£1,500.00	Low
103	Sam Brown	East	£4,500.00	10	£1,800.00	1.1	No	£2,700.00	£450.00	15.0	£2,700.00	£4,725.00	Mid
104	Linda Johnson	West	£2,500.00	5	£1,000.00	0.9	Yes	£1,500.00	£500.00	5.0	£1,000.00	£1,500.00	Low
105	Michael Lee	North	£7,000.00	15	£2,500.00	1.3	No	£4,500.00	£466.67	22.5	£3,750.00	£9,900.00	High
106	Emily Davis	South	£3,200.00	7	£1,400.00	1.0	Yes	£1,800.00	£457.14	7.0	£1,400.00	£1,800.00	Low
107	David Wilson	East	£5,300.00	14	£2,300.00	1.2	No	£3,000.00	£378.57	21.0	£3,450.00	£6,090.00	High
108	Susan White	West	£2,900.00	6	£1,100.00	0.8	Yes	£1,800.00	£483.33	6.0	£1,100.00	£1,800.00	Low
109	Chris Martin	North	£6,000.00	13	£2,200.00	1.2	No	£3,800.00	£461.54	19.5	£3,300.00	£7,500.00	High
110	Anna Taylor	South	£3,100.00	8	£1,350.00	0.9	Yes	£1,750.00	£387.50	8.0	£1,350.00	£1,750.00	Low
111	James Anderson	East	£4,700.00	11	£1,900.00	1.1	No	£2,800.00	£427.27	16.5	£2,850.00	£4,905.00	Mid
112	Patricia Thomas	West	£2,600.00	5	£1,050.00	0.8	Yes	£1,550.00	£520.00	5.0	£1,050.00	£1,550.00	Low
113	Robert Jackson	North	£5,500.00	12	£2,100.00	1.2	No	£3,400.00	£458.33	18.0	£3,150.00	£6,750.00	High
114	Mary Harris	South	£3,300.00	9	£1,450.00	1.0	Yes	£1,850.00	£366.67	9.0	£1,450.00	£1,850.00	Low
115	Daniel Clark	East	£4,900.00	11	£2,000.00	1.1	No	£2,900.00	£445.45	16.5	£3,000.00	£5,085.00	Mid
116	Barbara Lewis	West	£2,700.00	6	£1,150.00	0.9	Yes	£1,550.00	£450.00	6.0	£1,150.00	£1,550.00	Low

Figure 2 – Transformed sales data with CLV and value categorization.

2 – Statistical Analysis

2.1 - Data Pre-processing

Z-scores were used for outlier detection, of three standard deviations – no values were detected. Non-derived categorical data such as Region and Churned are also already evenly distributed amongst their respective categories, so methods such as stratified sampling are not required for sample representation.

2.2 - Quantitative Relationships with Hypothesis Testing

Most of the data is quantitative; so, a correlation matrix can be insightful in visually identifying dependencies and determining what combinations of variables are/aren't suitable for pairing, when constructing insightful visuals or statistical models.

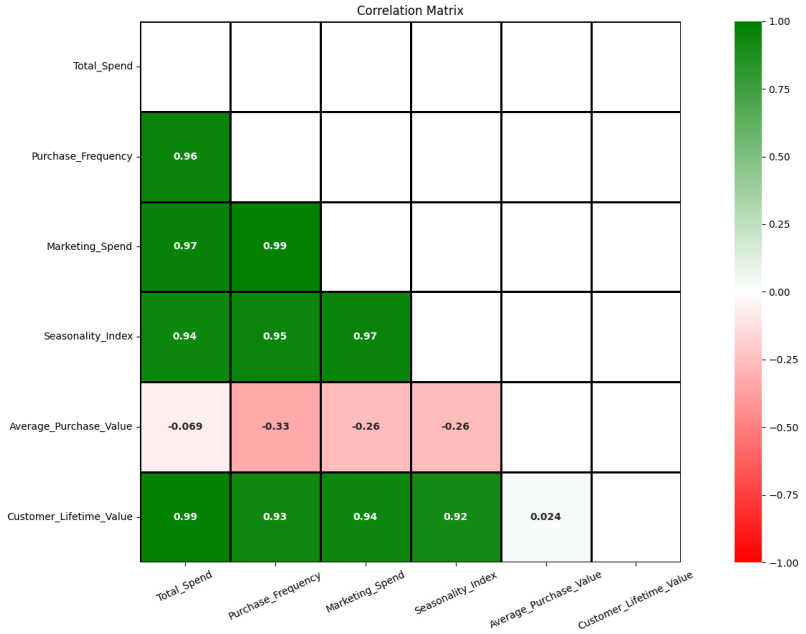

Figure 3 – Seaborn correlation matrix using p-value (< 0.05) masking.

P-values (pearsonr) was used as a form of hypothesis testing to determine whether the correlations between data are statistically significant. As no values were masked, the relationships are unlikely to have occurred by random chance. All relationships using **Average_Purchase_Value** have weak correlations, whereas **Marketing_Spend by Purchase_Frequency**, and **CLV by Total_Spend**, have the strongest positive correlations.

2.2.1 - KMO Test

Factor analysis can be conducted based on the findings of *figure 3*, however, an additional test can be conducted to validate this statement. A Kaiser-Meyer-Olkin test was chosen, as the most straightforward method.

Feature	**KMO Score**
Total Spend	0.8572
Purchase Frequency	0.6902
Marketing Spend	0.7331
Seasonality Index	0.8100
Average Purchase Value	0.1360
Customer Lifetime Value	0.6933

Where: **KMO Score > 0.5** is acceptable.

Average_Purchase_Value should; be removed based on both studies.

All other values are kept as features.

2.2.2 - Eigenvalues & Factor Loadings

Using a FactorAnalyzer, eigenvalues are derived from the features, where only the first factor returns an eigenvalue (**4.83**) greater than the Kaiser Criterion Threshold (**>= 1**). Knowing this, we can take the factor loadings, ignoring all factors asides from the first.

Features	Factor 1
Total Spend	0.58
Purchase Frequency	0.66
Marketing Spend	0.81
Seasonality Index	0.78
Customer Lifetime Value	0.81

Where: **Factor Load > 0.5** is acceptable.

All features can collectively define **customer value** to the business; therefore, can be confidently used for segmentation.

2.3 - Statistical Modelling of CLV by Regions

Using a one-way ANOVA test, we can determine if there are any statistically significant differences between CLV of each region. The resulting value (rounded by 6. dp) was:

0.000001 < 0.05 (P-Value)

Rejecting the null hypothesis, we can infer that the CLV of one or more of the four regions greatly differ from the others. With this in mind, we can plot a box plot as a post-hoc test to visually deduce the differences of distributions between the various regions.

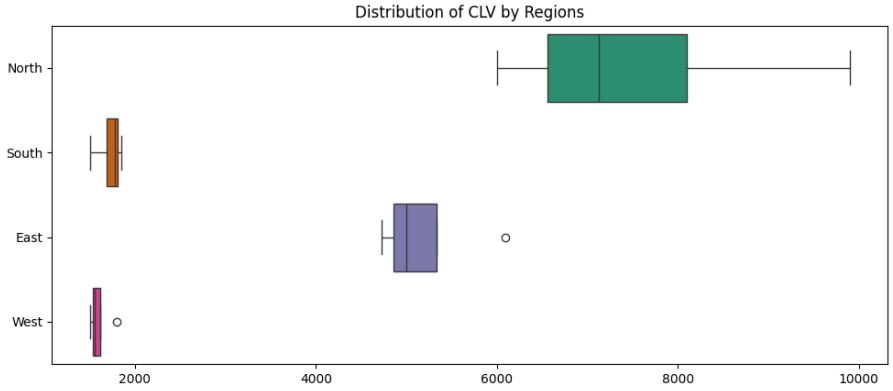

Figure 4 – CLV distributions of the four regions using a box plot.

Both North and East, have distinctively greater CLV values than the South and West. Possibilities regarding the CLV of South and West are; poor marketing and high churn rates – the root cause should be investigated, with solutions to offset the problem. Conversely customer retention of North and East should be further capitalized, as both regions clearly generate the largest bulk of value for the business.

3 – Customer Segmentation

With the features identified from *section 2.2*, we can segment the customers into various groups based on the value they bring, as well as targeted approaches in marketing.

3.1 - KMeans Clustering with Normalized PCA Values

KMeans clustering is an unsupervised approach of machine learning for grouping data. With this we can plot out the customer segments; and visually identify patterns in the data based on central nodes and the spread of data.

- Dataset was normalized (StandardScaler) to reduce bias from a small dataset.
- **P**rincipal **C**omponent **A**nalysis condenses 5 features into x and y axis values.

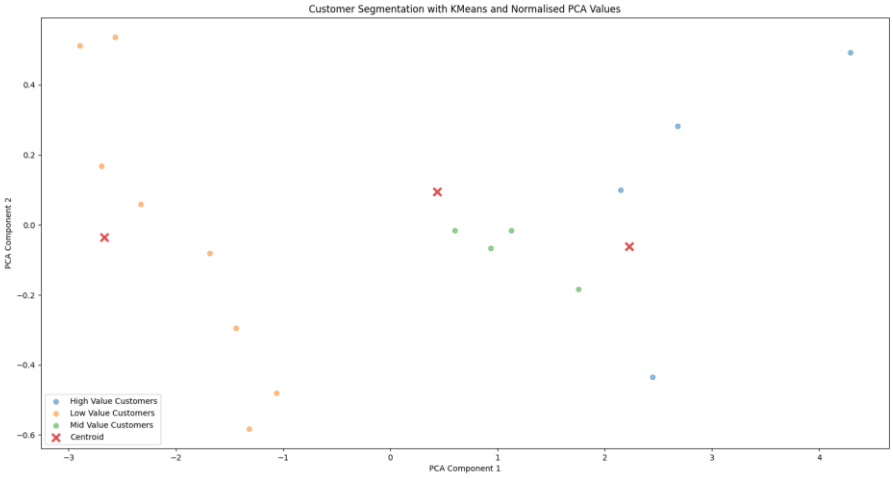

Figure 5 – Customer Segmentation into 3 categories, with Centroid using KMeans.

PCA components are summarized below.

	CLV	Total Spend	Purchase Frequency	Seasonality Index	Marketing Spend
PC1	0.4580	0.4576	0.4415	0.4296	0.4486
PC2	-0.5815	-0.3204	-0.0662	0.6505	0.3627

- Customers are segmented more so by the horizontal axis, suggesting that seasonality index and marketing spend alone are not enough as classifiers.

- **High**, **Mid**, and **Low** value customers can be derived from CLV, and total spend as primary features (similar weights), with the other three as secondary.

- The boundary between mid and high value customers is vague, implying there is a potential of conversion between the two segments.

3.2 - Ensemble Learning with Decision Trees and Testing

Ensemble learning can be used to further test the segmentation of customers (for improved accuracy), by feeding the categorical values into a decision tree as training and test data. The training data will be used to create the model of the decision tree.

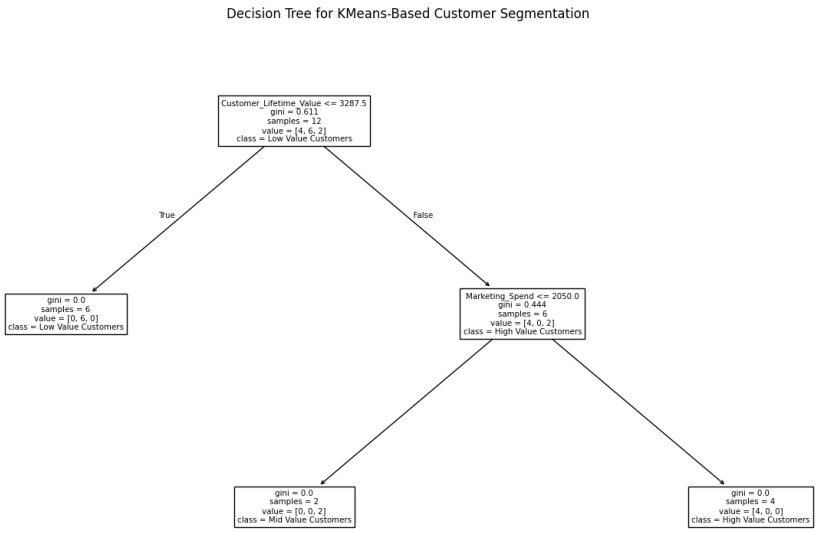

Figure 6 – 2-depth decision tree for high, mid, and low value customers.

Findings:

- Low CLV can cleanly separate low value customers.
- High marketing spend can cleanly separate high from mid value customers.
- High gini values suggests difficulty in categorization (5 features).
- Leaf nodes are all pure either suggesting incredible accuracy or overfitting.

With the test data, we can create a confusion matrix and classification report to examine the last finding; by determining the accuracy of the model. The outcome was **2 true positives and negatives** each, with **100% precision**. This is obviously too idealistic; therefore, we can deduce that the model has likely overfitted due to the small sample.

The primary issue is the small sample for both training and validation, which cannot be increased at the present; therefore, further ensemble learning is unlikely to have a significant impact.

4 – Predictive Analysis

As there is no timeframe data, time-series analysis such as forecasts are out of the question. Instead, we Churn is available which can be utilized to create logistic regression graphs to identify critical breakpoint for customer retention. Training data will be used to model the three regression lines, with testing data used for determining model accuracy.

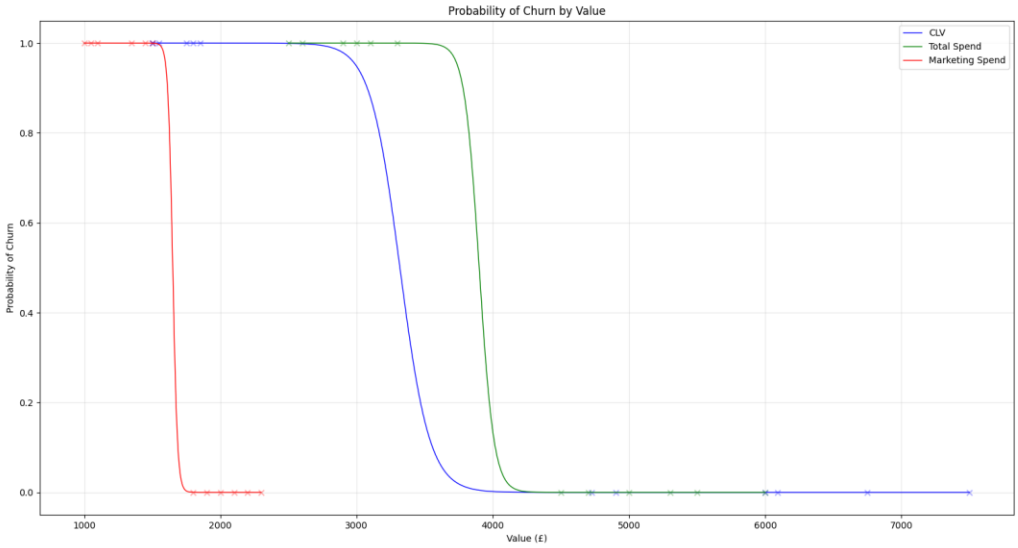

Figure 7 – Multiple logistic regression for CLV, total spend, and marketing spend.

Findings:

- Negative correlations: **Higher value = Lower churn probability**.
- Marketing Spend visually shows the steepest curve – Suggests it has the highest return on interest at threshold points **(£1500 to £1700)**.
- CLV has the smoothest curve *suggesting* the weakest influence on churn – Points are the most spread out indicating a wider range needed for conversion **(£3200 to £3800)**.
- Total spend steepness is between the two, suggesting gradual churn influence **(£3800 to £4200)**.

It's worth noting that the values were **not normalized**, as it would be more difficult for a business to visually interpret the results. This can result in curves appearing steeper due to points being closely clustered together – so logistic regression coefficient was used to identify steepness.

Features	**β_1**
Customer Lifetime Value	1.7318
Total Spend	1.7227
Marketing Spend	1.6733
Purchase Frequency	1.6173
Seasonality Index	1.6492

CLV has the highest β_1 value, suggesting it to be greatest influencer to churn probability.

Marketing spend conversely has the least of the three.

Running a confusion matrix, we get **2 true positive and negative values** each, once again suggesting that this model may be overfitted. So, the predictions shouldn't be taken at face value; and instead, should be used as a reference.

5 – Summary

Customers can be divided into three distinctive categories:

- **High value** – Stable customers who generate the most value for the business.
- **Mid value** – Semi-frequent customers with potential to be high value.
- **Low value** – Low engagement customers with high risk of churning.

Low Value Customers:

- Customer retention is the highest priority – Use **proactive marketing** for customers **flagged** with low CLV values as it is the best indicator for churn.
- **Re-engagement campaigns** can incentivize spending or at least extend their lifetime value.
- Increase/introduce **subscription models** or **split payment** methods to increase purchase frequency and extend lifetime value, whilst reducing dependency of high value customers.

Mid Value Customers:

- Have potential to convert to high value customers – Use **loyalty programs** with regular offers to maintain engagement, **subscription models** can also be used.
- Seasonality index begins to have a greater influence – **Seasonal promotions** have a greater impact from mid to high value customers.

High Value Customers:

- Requires the least pro-active attention – Use exclusive **VIP benefits** in loyalty programs to foster brand loyalty – E.g. Free deliveries, or trial products.

Other Considerations:

- **Increase stock** according to the number of mid to high value customers, especially **during active seasons**.
- Marketing has greatly diminished returns for customer retention around **£1600**, **optimize the marketing budget** around this.