Project 4

Assessment of Skills

1 – Data Transformation

The dataset provided contains details on customer sales over their lifetime, with 15 fields (including ID) and 500 records.

Date	Customer_ID	Age	Gender	Income	Spending_Score	Credit_Score	Loan_Amount	Previous_Defaults	Marketing_Spend	Purchase_Frequency	Seasonality	Sales	Customer_Churn	Defaulted
12/04/2024	1	56	Female	£142,418.00	7	391	£8,083.00	1	£15,376.00	3	Low	£32,526.00	0	0
21/02/2024	2	69	Male	£63,088.00	82	652	£34,328.00	2	£6,889.00	6	Low	£78,493.00	0	0
02/04/2024	3	46	Male	£136,868.00	91	662	£47,891.00	2	£6,054.00	29	Medium	£57,198.00	1	0
15/01/2024	4	32	Female		34	644	£25,103.00	2	£4,868.00	8	Medium	£48,395.00	0	0
16/04/2024	5	60	Male	£59,811.00	91	469	£44,891.00	1	£17,585.00	12	High	£29,031.00	1	0

Figure 1 – The first 5 rows of the unprocessed data loaded into Excel.

Data Cleaning Considerations:

- Blank values for: **Date**, **Income**, **Credit_Score**, and **Loan_Amount.**
  - Filled blanks using their **median** (rounded down).
- No duplicates were found.
- Using z-scores (**3. Std**) no outliers were found.

For the less obvious columns, I've defined them as the following:

- **Spending Score** - A scale value of **1-100** for customer spending, based on unknown calculations. Higher values mean higher and/or consistent spending.
- **Credit Score** - A range for a customer's creditworthiness. Likely **FICO Score 8**.
- **Loan Amount** - Total amount borrowed for credit purchases over customer lifetime.
- **Previous Defaults** - Count of customer past credit defaults.
- **Marketing Spend** – Total spent on customer acquisition or retention over lifetime.
- **Purchase Frequency** – Total orders over customer lifetime.
- **Seasonality** – How likely is the customer going to make a purchase during on-peak seasons?
- **Sales** – Total revenue over customer lifetime.
- **Customer Churn** – Where **0 is false and 1 is true**. Is the customer still active?
- **Defaulted** - Where **0 is false and 1 is true**. Has the customer currently defaulted their latest loan?

Aims & Objectives:

- Identify key sectors/target audience for targeted marketing.
- Determine key influencers for default risk.
- Derive three strategic action groups based on risk-value assessment.
- Find a threshold for 'risk of customer default' for flagging.

1.1 - Feature Engineering

Customer lifetime value (**CLV**) can be used for customer value segmentation, which is something we can calculate with our current data.

CLV = Customer Value * Lifetime

As we have a mix of churned and active customers, there will be two different calculations. We know how much value a churned customer has made via profits, whereas active customers will require predictive values to estimate both lifetime and their value.

Methodology:

- Churned Customers:
    - **CLV** = [**Sales**] - [**Marketing Spend**]

- Active Customers:
    - **Avg Sales** = {[**Sales**] - [**Marketing Spend**]} / [**Purchase Frequency**]

    - **Projected Purchases** = [**Purchase Frequency**] * [**Spending Factor**]

    - **CLV** = [**Avg Sales**] * [**Projected Purchases**] * [**Seasonality Factor**]

*Where Seasonality is **x** Seasonality Factor is **y**:*

x = High	**1.2**
x = Medium	**1.0**
x = Low	**0.8**

*Where Spending Score is **x** Spending Factor is **y**:*

x <= 33	**1.1**
x <= 66	**1.4**
x > 66	**1.7**

Note: Factors are used as heuristic values for predictions based on pre-existing data, they are based in between the pessimistic and optimistic ranges.

Gender	Income	Spending_Score	Credit_Score	Loan_Amount	Previous_Defaults	Marketing_Spend	Purchase_Frequency	Seasonality	Sales	Customer_Churn	Defaulted	Profit	Loss	Spend_Factor	Seasonality_Factor	Avg_Sales	CLV
Female	£142,418.00	7	391	£8,083.00	1	£15,376.00	3	Low	£32,526.00	0	0	£17,150.00	0	1.1	0.8	£5,716.67	£15,092.00
Male	£63,088.00	82	652	£34,328.00	2	£6,889.00	6	Low	£78,493.00	0	0	£71,604.00	0	1.7	0.8	£11,934.00	£97,381.44
Male	£136,868.00	91	662	£47,891.00	2	£6,054.00	29	Medium	£57,198.00	1	0	£51,144.00	0	1.7	1.0	£1,763.59	£51,144.00
Female	£85,375.00	34	644	£25,103.00	2	£4,868.00	8	Medium	£48,395.00	0	0	£43,527.00	0	1.4	1.0	£5,440.88	£60,937.80
Male	£59,811.00	91	469	£44,891.00	1	£17,585.00	12	High	£29,031.00	1	0	£11,446.00	0	1.7	1.2	£953.83	£11,446.00
Male	£134,825.00	17	655	£15,754.00	1	£19,881.00	13	Low	£80,542.00	0	0	£60,661.00	0	1.1	0.8	£4,666.23	£53,381.68

Figure 2 – First 5 rows of the transformed dataset including engineered columns.

2 – Preliminary Analysis

A correlation matrix was derived from relevant quantitative values to identify any key linear relationships between the data. P-value masking was used to filter out any statistically insignificant relationships.

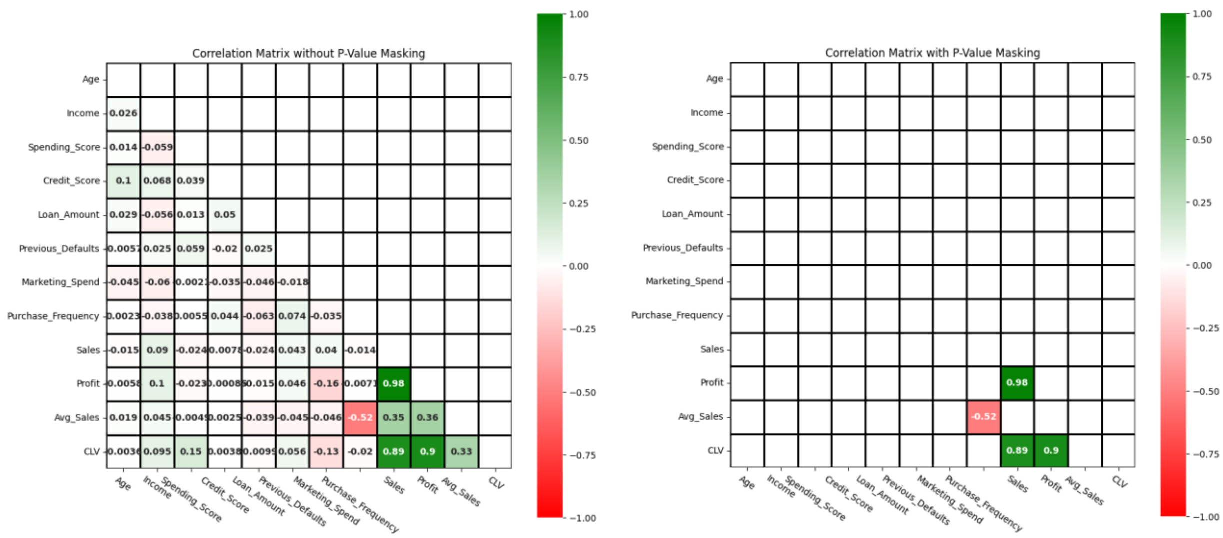

Figure 3 – Correlation strengths of linear relationships w/o & w/ p-value masking.

Key Findings:

- Most linear relationships are statistically insignificant.
- High profit is driven by high sales – Suggests profit metric is reliable.
- High CLV is driven by primarily by high financial contribution.
- High frequency customers spend less per order.
- Demographic factors aren't direct drivers – May require non-linear models.
- Risk factors aren't linearly related – Risk may require other models to measure.

3 – Target Audience & Key Domains

From the dataset, we can group customers by four categories:

- **Gender**.
  - Pre-categorized into male and female.
- **Age Group**
  - Grouped by intervals of 10, starting from 18 to 77.
- **Income Group**
  - Binned into four quartiles, to avoid category bloat and visual noise.
- **Seasonality**
  - Pre-categorized into low, medium, and high.

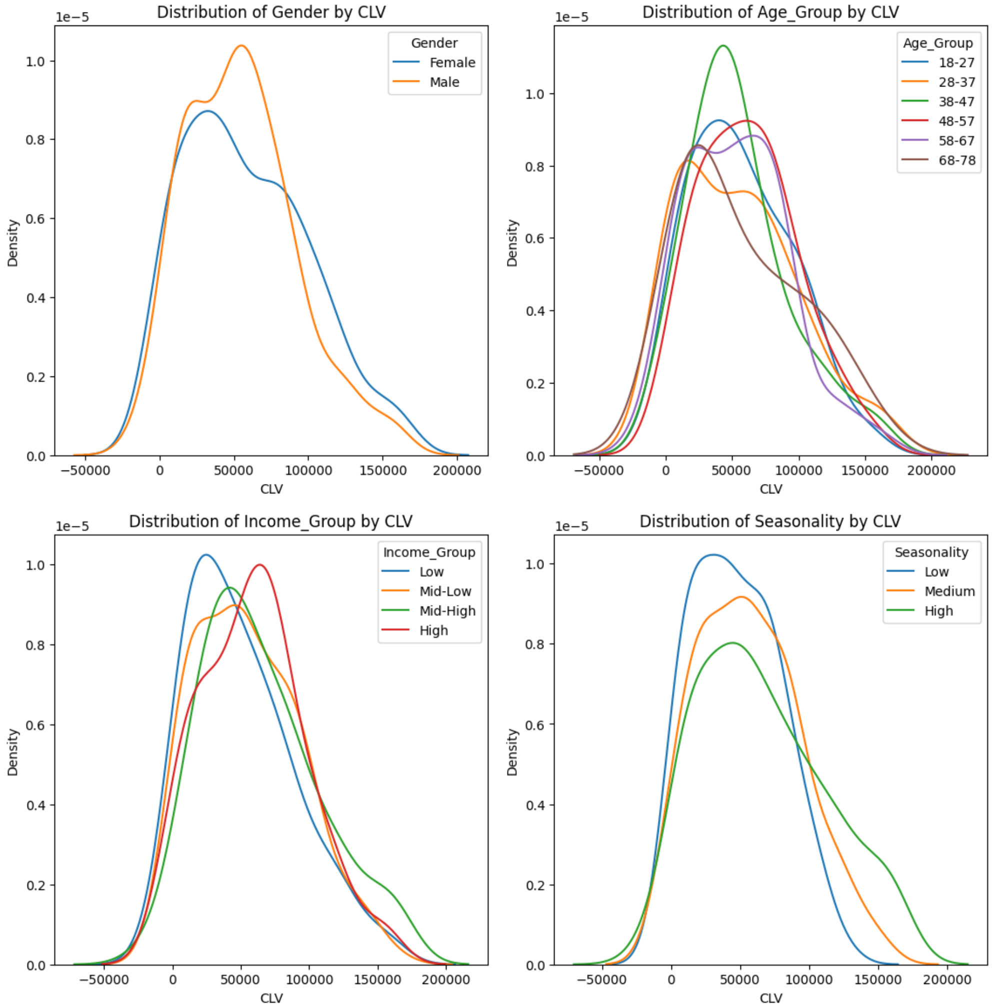

Figure 4 – KDE plots of grouping density by customer value.

Key Findings:

- **By Gender**
  - Males - Broad base of dependable value.
  - Females – Strategic high-value individuals (larger high CLV tail).
- **By Age Group**
  - Older age groups have higher potential value (especially 68-77).
  - Younger age groups skewed towards lower value but with less spread.
  - '38–47' is a narrow high-peak normal distribution – Stable, mid-value.
- **By Income Group**
  - All have widespread – Income group has high variability in value, likely more factors at play.
  - All below High are skewed towards the left, lower average value.
  - Mid-High has largest spread, suggests high-income customers mixed in.
  - High generates the most value, but low tail suggests untapped potential.

- **By Seasonality**
    - All three categories peak at 50,000, but with varying densities.
    - Low & Mid have high peaks, suggesting less on-peak season dependency for the business.
    - High seasonality customers have high potential profit margins (right tail).

4 – Analysis on Risk

As we plan to segment customers based on their risk-value assessment, we need to derive a '**Risk**' value from our pre-existing data. A **risk score** was determined to be the most appropriate, as it can be used to flag customer defaults. This can be done similarly to feature engineering CLV with the following formula:

Risk Score = [Likelihood] * [Impact]

Using this formula, we can extract the following features as potential components with logical deduction.

Feature	Component of
Credit Score	Likelihood
Previous Defaults	Likelihood
Income	Likelihood
Loan Amount	Likelihood & Impact
Defaulted	Impact
Loss	Impact

Knowing these features, we now want to consider their respective weights for the formula, the reason for this is accuracy. As unlike CLV, the risk score was proposed for **flagging customers** who are at risk of defaulting.

As we know from *figure 3*, these features don't have any statistically significant relationship for linear regression, so it's worth considering other analytical options. The initial intent was to use factor loadings from factor analysis.

	KMO Score:
Credit Score	0.4804
Previous Defaults	0.4962
Income	0.4689
Loan Amount	0.4671
Defaulted	0.4852
Loss	0.4810

KMO Scores <= 0.6 are terrible.

Therefore, acceptable values:
KMO Score > 0.6.

Therefore, we deem these features unsuitable for Factor Analysis.

With this in mind, **RandomForest** was the next best choice is it categorically splits data based on hierarchical factors. Defaulted would serve as the classifier, and Loss was dropped from the features; as it too is considered a classifier.

4.1 - Data Pre-processing

As a supervised learning model, we want to consider any potential bias that could affect the model.

Actions Taken:

- Data was grouped based on classification of **Defaulted**.
- Based on the **minimum** grouping, a **random sample** was taken to ensure that the data set was **balanced**. This led to a **95:95** sample split of an original 500.
- Downscaled dataset was split into a **training** and **test** set by a **3:1** ratio.

4.2 - Random Forest Model

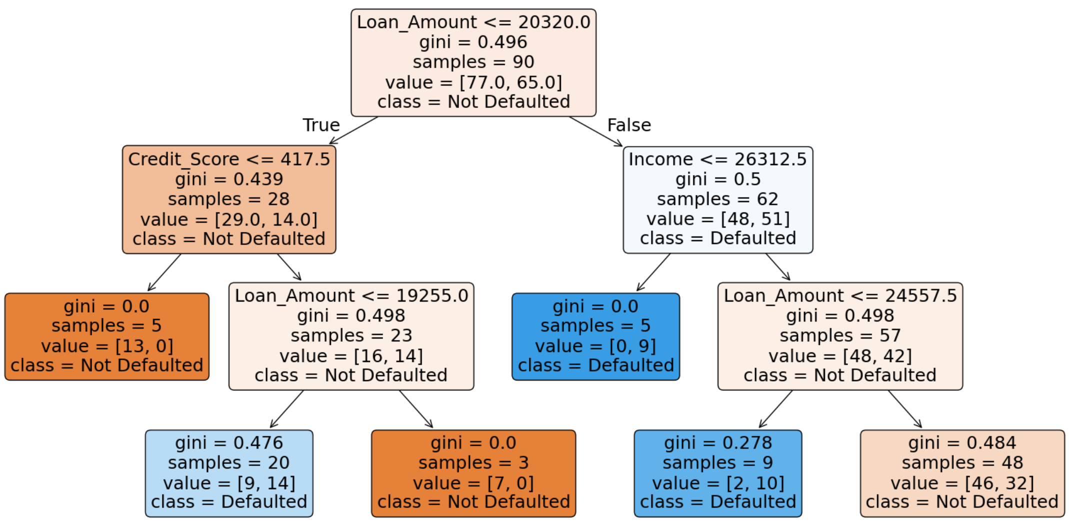

Figure 5 – 3-Depth RandomForest classification of Defaulted to Non-Defaulted.

Note: Originally had no depth restriction but was changed to 3-depth for readability. The effect on accuracy was negligible, so this was kept.

Initial Findings:

- Loan Amount has the greatest impact on classification.
- Credit Score & Income likely have similar weightings based on second decision node.
- Previous Defaults has little or no influence on categorizing Defaulted.

4.3 - Testing the Model

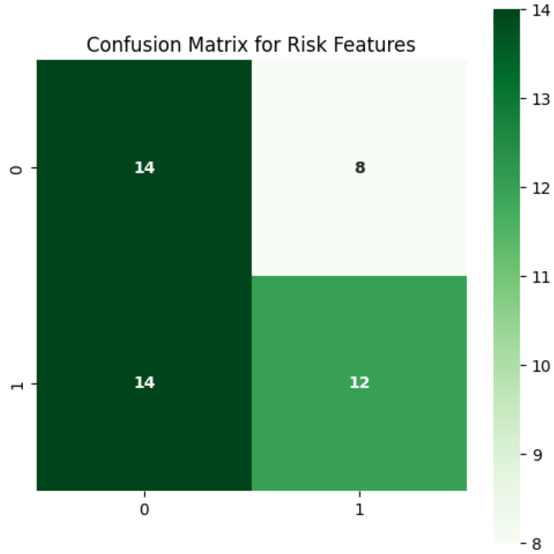

Figure 6 – Testing the RandomForest model with a Confusion Matrix.

Findings:

- Greater emphasis on predicting negatives (28 non-defaulted on left side).
- Missed classified a lot of defaulted customers.
- Poor accuracy based on the visual – Likely underfitted, or poor features for classification.

Additional testing was conducted to validate the model.

Accuracy Score	0.5417
F1 Score (Non-Defaulted)	0.5600
F1 Score (Defaulted)	0.5217

The model clearly isn't the most accurate at classifying either category. Even with hyperparameter tuning, and normalized data, this fact doesn't change.
However, it can still serve as a point of reference.

4.4 - Risk Score Formula

From the forest model we can derive these feature importances.

Loan Amount	0.3430
Income	0.3067
Credit Score	0.3170
Previous Defaults	0.0333

Loan Amount, Income, and Credit Score have around **equal** importance of classification.

Previous Defaults is negligible.

Considerations:

- Previous Defaults has a range of **0 – 2**.
- Previous Defaults is <u>tangible evidence</u> of risk and therefore should be **weighted** despite the low importance.
- Income & Loan Amounts are currency data, therefore have large ranges.
    - Should be **normalized**.
- Loan Amount, Income, & Credit Score have around equal importance.
    - Should share the **same weights**.
- Credit Score is based on FICO Score 8 based on the data.
    - Can be normalized by max (**850**).

Risk Score Formula:

Risk Score = {(850 – [Credit Score])/100} + (2 * [Previous Defaults]) + [Loan Amount Scaled] + [Income Scaled]

6 – Customer Segmentation

With CLV and risk score we can now consider segmenting the customer into strategy groups. KMeans clustering will be used.

6.1 - Identifying K-clusters

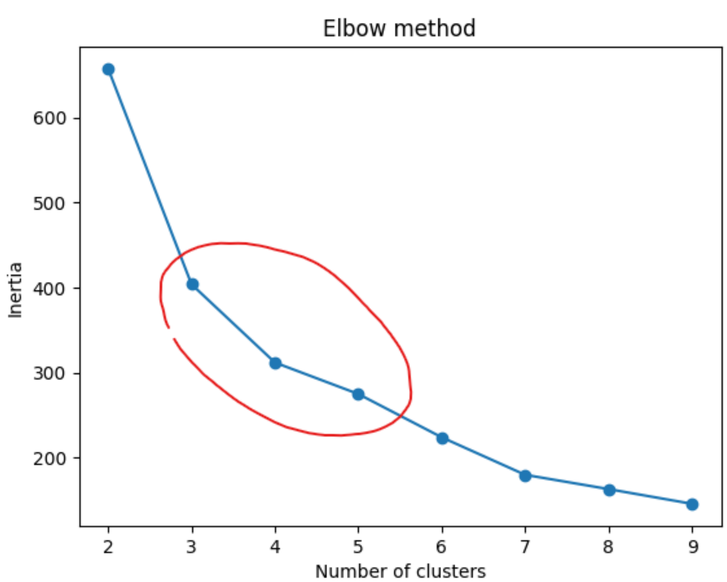

Figure 7 – Identify suitable number of clusters using Elbow Method.

We can identify that the range **k=3 to 5**, is the most suitable when clustering for CLV and risk score. To add further clarity, silhouette scores were used to identify the best k.

K=3	0.3736	**K=3** gives the best silhouette score.
K=4	0.3600	
K=5	0.3217	Therefore, use 3 clusters.

5.2 - KMeans Clustering

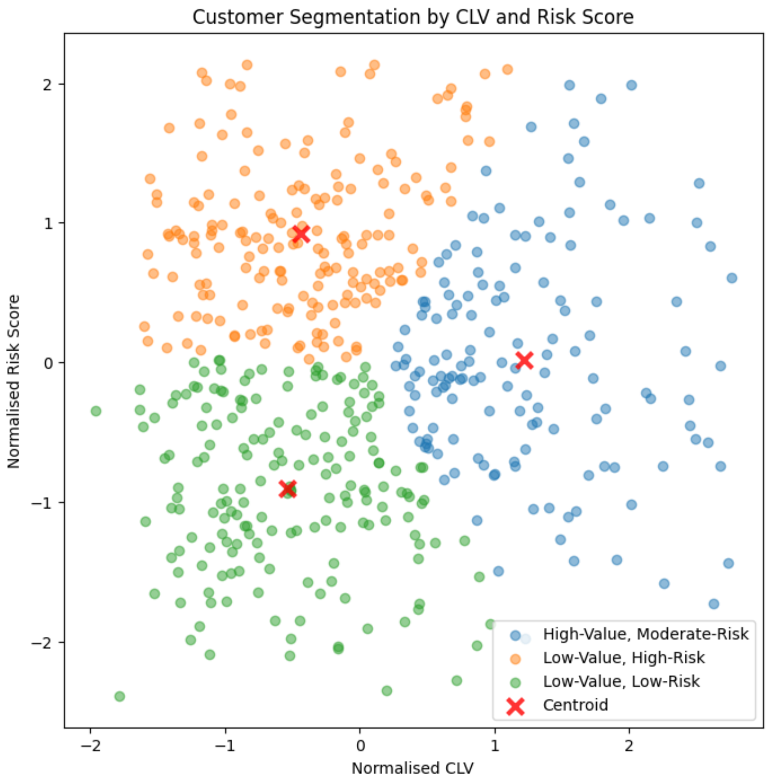

Figure 8 – Customer segmentation with centroids by value and risk.

Key Findings:

- Segments closely border one another.
  - Room for segment conversion between the three.
- Clear separation by CLV (left to right side).
  - High-CLV, Moderate-Risk
    - The most profitable segment, with the most spenders
- Low-Value customers can be further segmented by risk.
  - Low-CLV, Low-Risk
    - Most stable customer segment.
  - Low-CLV, High-Risk
    - The risk prone customer segment.

6 – Risk Score Flagging

Since we have a risk score formula, we can use this to flag customers who are likely to default using a threshold value, as opposed to quartile bin classifications.

6.1 - Statistical Method

From the clusters in the KMeans model, we can determine the cluster with the highest default rate (**Low-CLV, High-Risk**). We can then take the average risk score (**3.662**) from this segment to serve as a frame of reference for the threshold.

We can then test this threshold using a confusion matrix, and f1 scores.

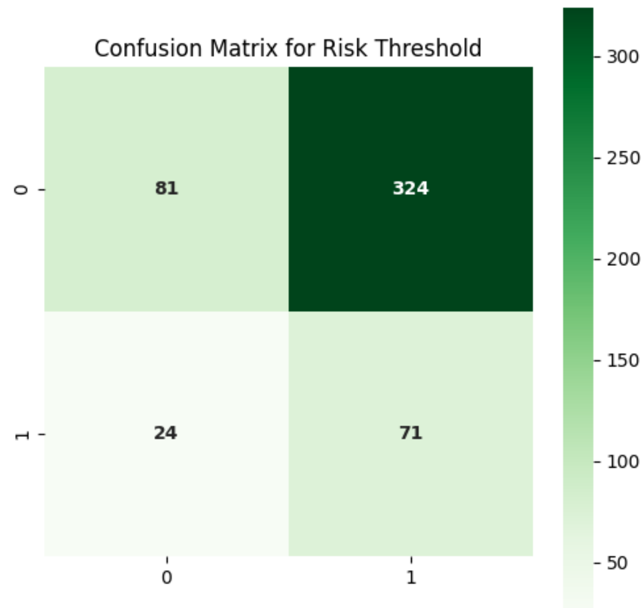

Figure 9 – Testing default classification of risk score threshold w/ confusion matrix.

We can tell there are a lot of misclassifications for non-defaulted as defaulted. This suggests that this threshold may be too low, therefore flagging a lot of non-defaulted customers as potential risk customers.

F1 Score (Non-Defaulted)	0.3176	Poor accuracy for both classifications.
F1 Score (Defaulted)	0.2898	

6.2 - Decision Tree Method

Like we did with the Random Forest, we will normalize the dataset using a balanced training/testing set for defaulted values. Using the risk score as the only feature in a 1-depth forest we get the following.

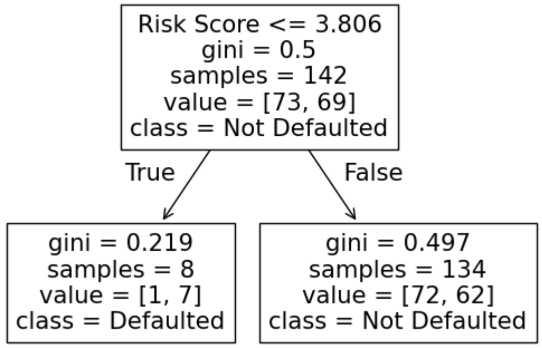

Figure 10 – 1-depth decision tree for identifying a risk score threshold for defaulted.

Once again, we want to test this threshold to determine the accuracy of this threshold.

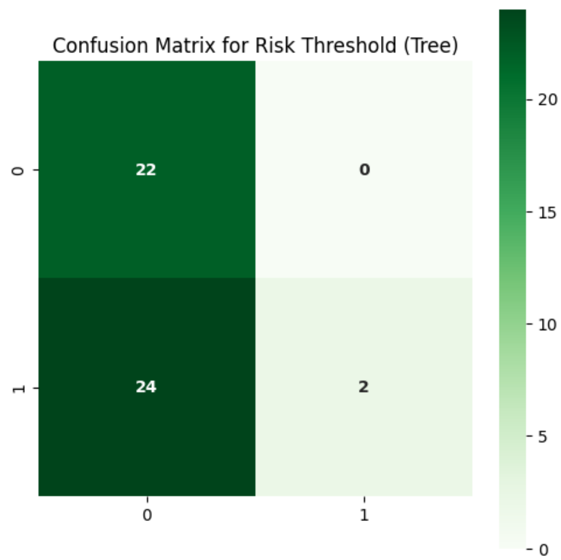

Figure 11 – Testing risk score threshold (tree) for classifying defaulted customers.

Unlike the statistical model, the tree risk score threshold, is more suitable for classifying non-defaulted customers. But as a result, has misclassified more defaulted customers as non-defaulted.

F1 Score (Non-Defaulted)	0.6471	Poor accuracy for Defaulted classifications, with *okay* classifications for non-defaulted
F1 Score (Defaulted)	0.1429	

6.3 - Risk Score Threshold

We have two threshold values we can use: statistical, and tree model. Normally, we would just take the average of the two; however, from a business perspective, **it's better to misclassify non-defaulted** customers as defaulted customers, than the other way around. This is because it's easier to mitigate issues from customer flagging, than a customer defaulting. Therefore, we will add weights (0.75 & 0.25) to the thresholds in the calculation, prioritizing the KMeans model.

Calculation for Risk Score Threshold:

$$(\underline{0.75} * 3.662) + (\underline{0.25} * 3.806) = \mathbf{3.698}$$

Therefore, customers with a **risk score >= 3.698** should be flagged.

7 – Summary

Customers can be segmented into three groups by value and risk.

Customer Segment Strategies:

- High-Value, Moderate-Risk – Priority value segment.
    - **VIP benefits** for brand loyalty.
    - Add **tiers** to VIPs based on payment behavior – **Low risk customers deserve better treatment**. Builds an incentive around goodwill & trust.
    - **Increase credit-cap** based on successful repayments.
    - **Personalized service** for risk flagging – e.g. letters of credit, grace periods, or alternative financial options.
    - Consider **internal flagging** for risk customers, but no action unless it is a repeat offender.
- Low-Value, Low-Risk – Stable value segment.
    - **Cost-effective methods** such as social media/email marketing.
    - **Subscription models** for sustained value generation.
    - **Upselling methods** like bundled offers or value-added services such as first-day delivery.
    - **Retention methods** e.g. loyalty programs or re-engagement campaigns.
    - **Soft/passive interventions** for flagging such as notifications.
    - Let low-risk flags accumulate before acting – Treat **first offense as a warning**.
- Low-Value, High-Risk – Risk prone segment.
    - **Limit credit exposure** based on customer or region – **Terminate credit transactions** for repeat offenders.
    - **Automate payment reminders**.
    - **Real-time risk flagging**.
    - **Auto-limit credit limit**, and manual approval of larger orders.
    - **Automated reminders** with escalations if needed.
    - **Partial forgiveness** with exit clause, or financial alternatives.

General Considerations:

- Implement **pro-active risk mitigation** methods for flagged customers.
- Low customer seasonality, but high profit margins on-season.
    - Focus more on **off-season marketing** than on-season.
    - Experiment with **upselling during on-peak** seasons such as limited-time bundles or value-added services.
- Profit-margin increases with income in exchange for less customers.
    - **Volume-driven** methods for customers **below 50**% income groups.
    - **Value-driven** methods for the **upper 50**% income groups.
- Value increases with age.
    - Focus on raising **brand awareness** for those **below 37** years old.
    - Increase **quality of services** for those **above 37**, e.g. dedicated support section.

7.1 - Ethical Considerations

Key points:

- Customer information – ID, Age, Date, Gender, & Income.
    - Contains **sensitive information** and should be masked if this information is to be presented or stored.
    - Hash functions are the simplest solution.
    - Appropriate security measure must be taken, so that it's secure and encrypted.
    - Data retention policies.
- Data Collection & Sharing
    - If this information is to be shared, **consent** must be acquired from said customer/s.
    - Let customer know what their data is used for.
    - Data that does not have consent (or withdrawn from consent) must be immediately removed.