

Decision Tree Analysis for Organic Products

Kritik Assignment 2 Topic A

**Spears School of Business**

**Watson Graduate School**

**02/01/2025**

Content

[Business Understanding 1](#_Toc190049833)

[Data Understanding 2](#_Toc190049834)

[Geographic Predictors 2](#_Toc190049835)

[Demographic Predictors 2](#_Toc190049836)

[Promotional Predictors 3](#_Toc190049837)

[Target Variable 4](#_Toc190049838)

[Data Preparation 5](#_Toc190049839)

[Data Cleaning 5](#_Toc190049840)

[Encoding and Partitioning of the Data 7](#_Toc190049841)

[Modeling 8](#_Toc190049842)

[Evaluation 8](#_Toc190049843)

[Scoring 10](#_Toc190049844)

[Reporting 11](#_Toc190049845)

[Annexes 12](#_Toc190049846)

[Annex 1 12](#_Toc190049847)

[Annex 2 12](#_Toc190049848)

[Annex 3 13](#_Toc190049849)

[Annex 4 13](#_Toc190049850)

[Annex 5 14](#_Toc190049851)

[Annex 6 14](#_Toc190049852)

[Annex 7 15](#_Toc190049853)

[Annex 8 15](#_Toc190049854)

[Annex 9 16](#_Toc190049855)

[Annex 10 16](#_Toc190049856)

[Annex 11 17](#_Toc190049857)

[Annex 12 17](#_Toc190049858)

# Business Understanding

**Importance of Customer Loyalty to Organic Products**

Customer loyalty to organic products is a critical element for supermarkets aiming to capture a stable and predictable segment of the market. Loyal customers are more likely to make repeat purchases, ensuring consistent revenue. Organic products typically have higher profit margins compared to non-organic products, and loyal customers are willing to pay a premium for these items.

Satisfied loyal customers often become brand advocates, promoting the store to friends and family, which can attract new customers. Understanding the purchasing habits of loyal customers helps supermarkets better manage their inventory, reducing waste and optimizing stock levels. Loyalty indicates trust and satisfaction with the quality and sourcing of the products, strengthening the overall customer relationship.

**Predicting Purchase Amount vs. Purchase Intent**

Predicting the actual amount spent on organic products, rather than just whether a customer bought them or not, can provide more profound insights. Knowing how much a customer spends can help in tailoring marketing campaigns and offers specifically to individual spending patterns. Accurate predictions of spending amounts can enhance sales forecasting, helping supermarkets to plan their supply chain and inventory more efficiently.

Detailed spending data allows for better customer segmentation, identifying high-value customers who contribute more to the bottom line. Understanding spending amounts can inform decisions about product placement and promotional strategies, ensuring that high-revenue-generating products receive optimal visibility. It also allows for better allocation of resources towards products and services that yield the highest returns, ensuring efficient use of marketing budgets and in-store efforts.

In summary, while knowing whether customers buy organic products is useful, predicting the actual spending amount provides a more comprehensive view of customer behavior, enabling supermarkets to make more informed and strategic business decisions.

# Data Understanding

The dataset contains 22223 rows, and 13 variables that can be used for segmenting the target market. However, for the scope and purpose of this project only 10 variables will be chosen. The dependent is the *TargetAmt* and its predictors are divided into three main categories: demographic, geographic and behavioral variables, explained below. The unique customer identifier ‘ID’ is not included in the decision tree but still, we use it to compare the rows to find possible duplicates.

## Geographic Predictors

* **Region**: Represented by a nominal variable in the column *DemReg.* This indicator encapsulates the region where the customer lives. In the UK there are basically 10 regions used for various administrative purposes, including statistical analysis and regional planning.

The variable has no missing values, its mode is ‘South East’ with a frequency of 9099 which leads to understand that most of the customers sampled were from London City and its surroundings (see annex 1 for more information).

* **TV Region:** Closely related to the variable explained above, the column *DemRegTV* identifies the location of the customer in terms of the television broadcasting regions. Those are areas defined by the British Broadcasting Corporation (BBC) for advertising purposes. That way, they ensure that local advertisers can effectively reach their target audiences. The column *DemRegTV* is a nominal variable with 13 possible values, with no missing ones. The mode is ‘London’, which makes us understand there could be a very high collinearity with *DemReg* since the names of the geographical and TV regions might be the same in numerous instances (see annex 2 for more information).

## Demographic Predictors

* **Affluence:** This continuous variable (*DemAffl*) identifies the socio-economic status of the individual. It ranges from 0 to 34, has 33 unique values and no missing ones. 90 percent of the whole sample is equal or under the 13th level of affluence, with a mean of 8.7, a median of 9 and a standard deviation of 3.3. So, these findings indicate that indeed there are some extreme values that fall very far from the quartiles 2nd and 3rd like the very maximum of 34th.
* **Age:** The age of the customer is represented by the continuous variable *DemAge.* Its values go from 18 to 79 and there are no missing ones. Interestingly, the mean and median are very close with 53.8 and 54 years old respectively, with leads to understanding the aging of the British population. This column does not seem to have outliers (see annex 4 for more information).
* **Demographic Cluster:** There is no further information about the meaning of the nominal values assigned to this column (*DemClusterGroup*), whose values are in the set {A, B, C, D, E, F}. Consequently, the tuples with value ‘U’ that can be interpreted as ‘*Undefined*’. This is very useful because there are 674 missing values that could be imputed to ‘U’ as well. (see annex 5 for more information).
* **Gender:** When it comes to the sex of the individuals in the sample, the indicator uses nominal values: F, M, U. There are no missing values (see annex 6 for more information).

## Promotional Predictors

* **Time Spent on Promotions:** This column contains a continuous value that ranges from 0 to 39. Unfortunately, its unit is not defined but still it conveys the meaning that the larger the figure the more time the person has spent acquiring promotions. The column *PromTime* does not contain missing values but it seems to have outliers since 90% of the values are under 12 units with a standard deviation of only 4.62 and a mean of 6.57 (see annex 7 for more information).

**Amount Spent on Promotions:** This indicator is also continuous and although its unit is not mentioned in the Data Dictionary, it can be inferred to be in Sterling Pounds. The variable named *PromSpend* ranges from £0 to £296313 with a median of £2000 and mean of £4420.59. These findings indicate the potential existence of outliers that needs to be addressed (see annex 8 for more information).

* **Promotion Classification:** The nominal variable *PromClass* defines the types of promotions the person has access to in the supermarket. Its values are Silver, Tin, Gold, and Platinum. There are no missing values or outliers detected (see annex 9 for more information).

## Target Variable

* **Target Amount to Spend:** *TargetAmt* is continuous because it represents a figure in Sterling Pounds. Its values vary from £0.0 to £1992.55 and its distribution is very skewed to the right because it has a maximum very, far from the mean of £47.42. Besides, 90% of all values are under £133.80. (see annex 10 for more information).

Figure 1 Histogram of the target variable Target Amount. Notice the sharp skewness to the right which indicates the possible presence of outliers.

A graph of a distribution of target amount

Description automatically generated

All in all, there are 10 variables that will be used in the decision tree. From them 9 are the predictors which are divided into: 5 nominal, and 4 continuous. All the continuous variables seem to have outliers which will need to be confirmed and consequently processed. Then, concerning the missing values, only the nominal predictor variable DemClusterGroup displays 674 values that can be imputed instead of deleted.

# Data Preparation

### Data Cleaning

The Decision Tree is not significantly affected by outliers in the input space, but only locally by outliers in the output variable. Therefore, because our primary goal is prediction accuracy, removing outliers in this case is not necessary, as decision trees are generally robust to them.

As you can see from figure 2, only the variable *PromSpend* has actually 3 very extreme values that represent anomalies in the distribution. For that reason, it was decided to remove only those top three readings.

Figure 2. Scatter plots of each one of the continuous variables. Notice the 3 extreme values in PromSpend

A group of graphs showing different sizes and numbers

Description automatically generated with medium confidence

Still, although there are other outliers in *PromTime*, *TargetAmount* and *DemAffl* they are preserved because their clustering could indicate unusual spending patterns worth of further research. In fact, the skewness in those three columns follow the same direction which make us understand this is not due to bad readings but indeed there was a sample of people who outstand according to their purchase behavior. However, as was mentioned above, only *PromSpend* displays very extreme values that are not clustered and noticeably break with the distribution of the variable (See figure 3).

Figure 3. Pairplot of all continuous variables grouped by DemGender. Notice the 3 extreme values in all PromSpend plottings that stand out.

A screenshot of a graph

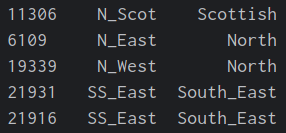
AI-generated content may be incorrect.

On the other hand, when it comes to the missing values, only the variable *DemClusterGroup* have 374 missing values that can be easily imputed to the ‘U’ value. After the deletion of the three extreme outliers in *PromSpend* and the imputation of the missing values in *DemClusterGroup* these are the resulting statistics and visualizations of the 9 predictors and the dependent variable.

A group of blue and white graphs

AI-generated content may be incorrect.

One final step was the substitution of the spaces in the string values in the nominal columns. It was necessary because when they are converted to dummy the new column name will contain white spaces which is not recommendable. Thus all spaces were changed to underscores.



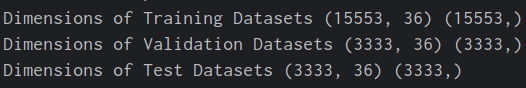
Screenshot 1. Sample of the resulting values of columns DemTVReg and DemReg after substituting the spaces for underscores.

## Encoding and Partitioning of the Data

As part of the data preparation process it was necessary to encode the nominal variables to dummy, binary variables to use them in the Decision Tree Regression because our output is a continuous value. For that reason, the dimension of the dataset increased to 36 columns (see annex 11).

The data was partitioned and used 70% for training and 15% for testing and 15% for validation. Also, it was ensured to use 12345 as the seed for the partitioning. Something worth mentioning is the fact that during the partitioning there was no stratification because some classes had only one value.

Screenshot 2. Dimensions of the resulting datasets after partitioning.



# Modeling

To comply with the requirements of the assignment we ran 4 decision three regression versions as follows:

* Test 1: Decision tree regression with all parameters and unpruned.
* Test 2: For conducting the manually pruned decision tree regression some extra steps were performed. First, it was necessary to run a hyperparameter tuning through cross-validation using GridSearchCV. This way it was obtained the best combination of parameters for pre-pruning the tree. Hyperparameter tuning is the process of finding the best hyperparameters for a machine learning model to enhance its performance. While it doesn't directly prune the decision tree, it helps identify the ideal combination of hyperparameters like max\_depth, max\_features, criterion, and splitter. This indirectly manages the complexity of the decision tree and helps prevent overfitting, making it a type of post-pruning technique
* Test 3: Feature-Selected Decision Tree: For selecting the best features to run the decision tree regression it was applied the recursive feature elimination (RFE) for 15 features. Once the algorithm threw its selection of parameters the decision tree was modelled.
* Test 4: We added an extra decision tree that uses both the best parameters and the features selected on the previous steps.

# Evaluation

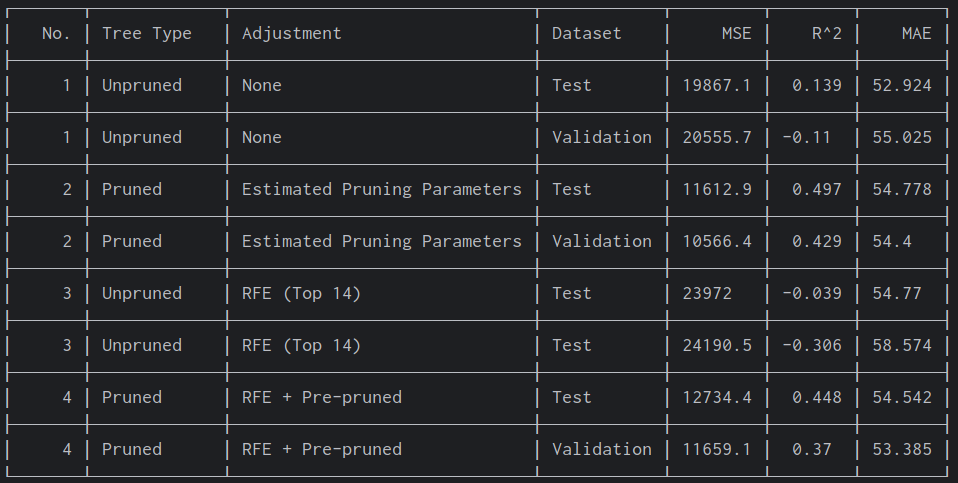
The modeling was run 3 times with the same sample size for the training, test and validation datasets. Each one of the occasions the number of features was changed from 20, to 17 and 14. This resulted in an improvement of the R^2 of the models meaning that the one with the lowest number of features turned to have the highest score (see annex 12).

By comparing models 1 & 2 and 3 & 4 suggests that pruning generally improves performance. In fact pruned trees (2 and 4) tend to have lower MSE and higher R^2 compared to their unpruned counterparts. Besides, when it comes to Recursive Feature Elimination (RFE), those models using RFE for feature selection (3 and 4) show mixed results. In some cases, RFE seems to improve performance (e.g., Model 4 vs. Model 2 on the Validation set), while in others, it doesn't show clear benefits or even worsens performance (e.g., Model 3 vs. Model 1).

Based on table displayed in the screenshot 3, Model 2 (Pruned with Estimated Pruning Parameters) appears to have the best overall performance, with the lowest MSE and highest R^2 on both Test and Validation datasets. For these reasons it was the choice for predicting the Target Amount. The parameters estimated by the model were:

* 'criterion': 'squared\_error'
* 'max\_depth': 5
* 'max\_features': 'log2'
* 'min\_samples\_split': 2
* 'splitter': 'best'

Screenshot 3. Outputs of the models 3 the one with the best performance.

.

# A screenshot of a cell phone AI-generated content may be incorrect.A diagram of a diagram AI-generated content may be incorrect.Scoring

Figure 4. Prediction of Target Amount with a Decision Tree Regression from the dataset provided for Scoring (5 new cases)

# Reporting

The decision tree regression model reveals key insights into customer spending behavior, enabling the company to craft targeted organic product promotions. Affluence level (DemAffl) is the most influential predictor, with higher-affluence individuals (>18.5) displaying significantly greater spending. Among these, younger customers (DemAge ≤ 32.5) in the North and South West regions exhibit the highest spending potential, reaching values above 1,200–1,500. This segment should be targeted with premium organic products, personalized offers, and exclusive membership benefits. Additionally, customers with a Silver or Platinum loyalty class tend to spend more, indicating that upgrading more customers into these tiers could boost organic product sales.

For lower-affluence customers (DemAffl ≤ 18.5), spending varies significantly by region and gender. Males with low promotional spending (<490.01) and residing in certain clusters (e.g., North Scotland TV region) show minimal engagement, with spending dropping close to 0. This segment may not be an immediate priority for organic promotions but could be gradually introduced through budget-friendly organic options. Conversely, females in the South East region, particularly under age 39.5, exhibit moderate spending potential (72–118). This group could be encouraged to shift towards organic products with educational campaigns and targeted discounts.

Promotional spending history (PromSpend) also plays a crucial role, particularly for high-value customers. Those with PromSpend > 2018.90 and residing in London TV regions spend significantly more (~38.42), making them strong candidates for digital ad campaigns and premium product promotions. However, even within high-affluence groups, there is variation. Older customers (DemAge > 39.5) exhibit slightly lower spending unless they have Platinum status or higher engagement time (PromTime > 6.5). To maximize engagement from this segment, time-sensitive discounts and high-value product bundles can be introduced.

Overall, the company should prioritize high-affluence, younger consumers in key regions, particularly those with existing promotional spending habits and loyalty status. Lower-affluence groups require different strategies—affordable organic product lines, regional promotions, and gradual upselling tactics. Aligning organic product marketing with these insights will help maximize revenue and improve conversion rates.

# Annexes

## Annex 1

Main descriptive indicators of the variable DemReg

A screenshot of a computer

Description automatically generated

## Annex 2

Main descriptive indicators of the variable DemRegTV

A screenshot of a computer

Description automatically generated

## Annex 3

Main descriptive indicators of the variable DemAffl

A screenshot of a computer

Description automatically generated

## Annex 4

Main descriptive indicators of the variable DemAge

**A screenshot of a computer

Description automatically generated**

## Annex 5

Main descriptive indicators of the variable DemClusterGroup

A screenshot of a computer

Description automatically generated

## Annex 6

Main descriptive indicators of the variable DemGender

A screenshot of a computer

Description automatically generated

## Annex 7

Main descriptive indicators of the variable PromTime

A screenshot of a computer

Description automatically generated

## Annex 8

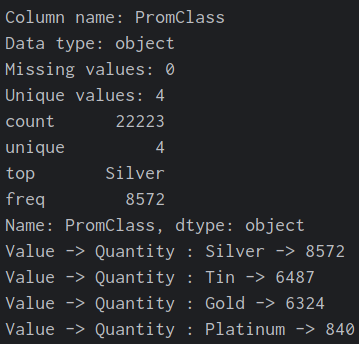
Main descriptive indicators of the variable PromSpend

**A screenshot of a computer

Description automatically generated**

## Annex 9

Main descriptive indicators of the variable PromClass



## Annex 10

Main descriptive indicators of the variable TargetAmount

A screenshot of a computer

Description automatically generated

## Annex 11

Dimension of the dataset of predictors. 

List of variables created after generating dummy columns.

A screenshot of a computer screen

AI-generated content may be incorrect. A screen shot of a computer

AI-generated content may be incorrect. A screen shot of a computer program

AI-generated content may be incorrect.

## Annex 12

Outputs of the three models tested. Notice the difference between the number of features: 20, 17 and 14, that also improve the performance.

A screen shot of a computer

AI-generated content may be incorrect.

