Decision Tree Analysis for Organic Products

Kritik Assignment 2 Topic A

Spears School of Business

Watson Graduate School

02/01/2025

Content

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

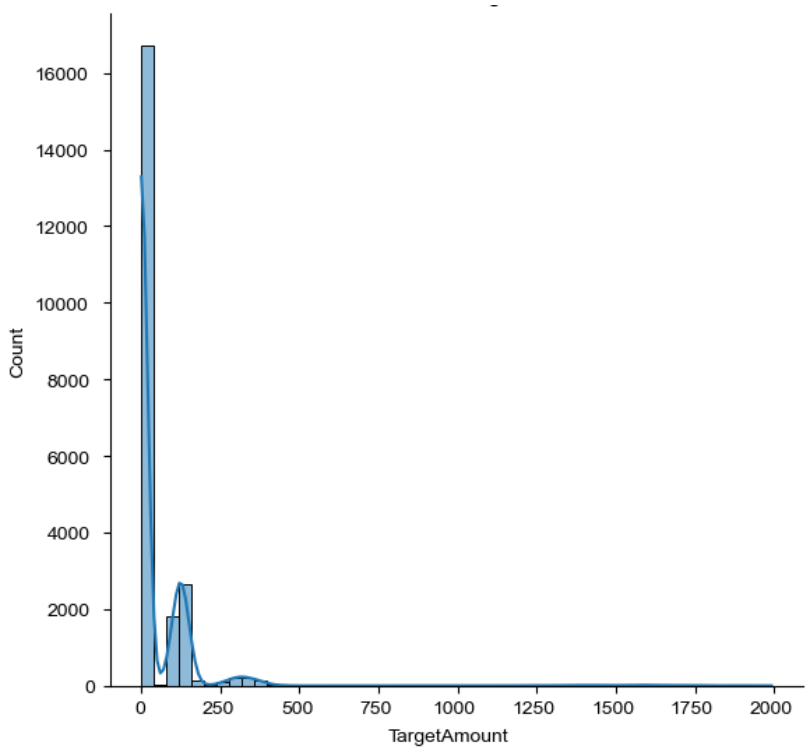
- **Time Spent on Promotions:** This column contains a continuous value that ranges from 0 to 39. Unfortunately, its unit is not defined but still it conveys the meaning that the larger the figure the more time the person has spent acquiring promotions. The column *PromTime* does not contain missing 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.



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.

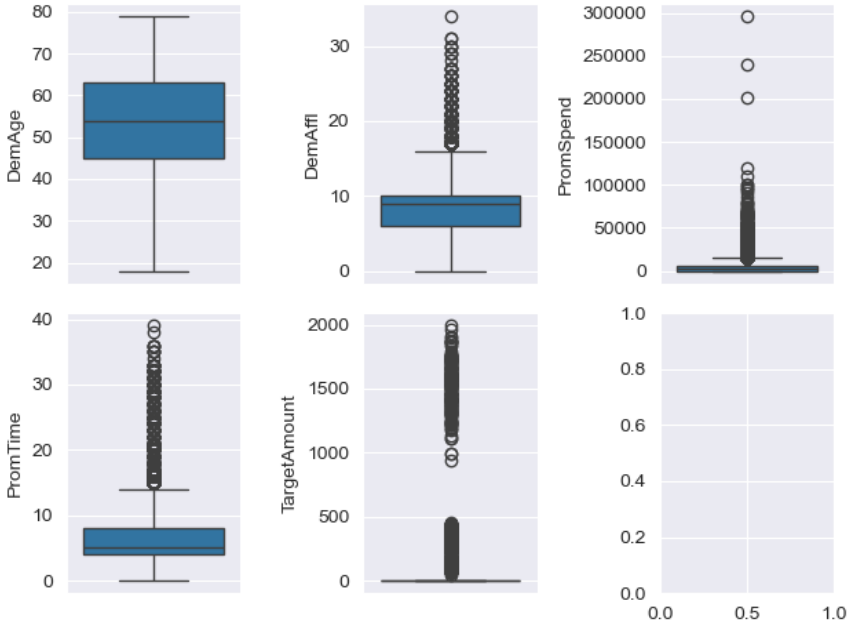
Data Preparation

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.

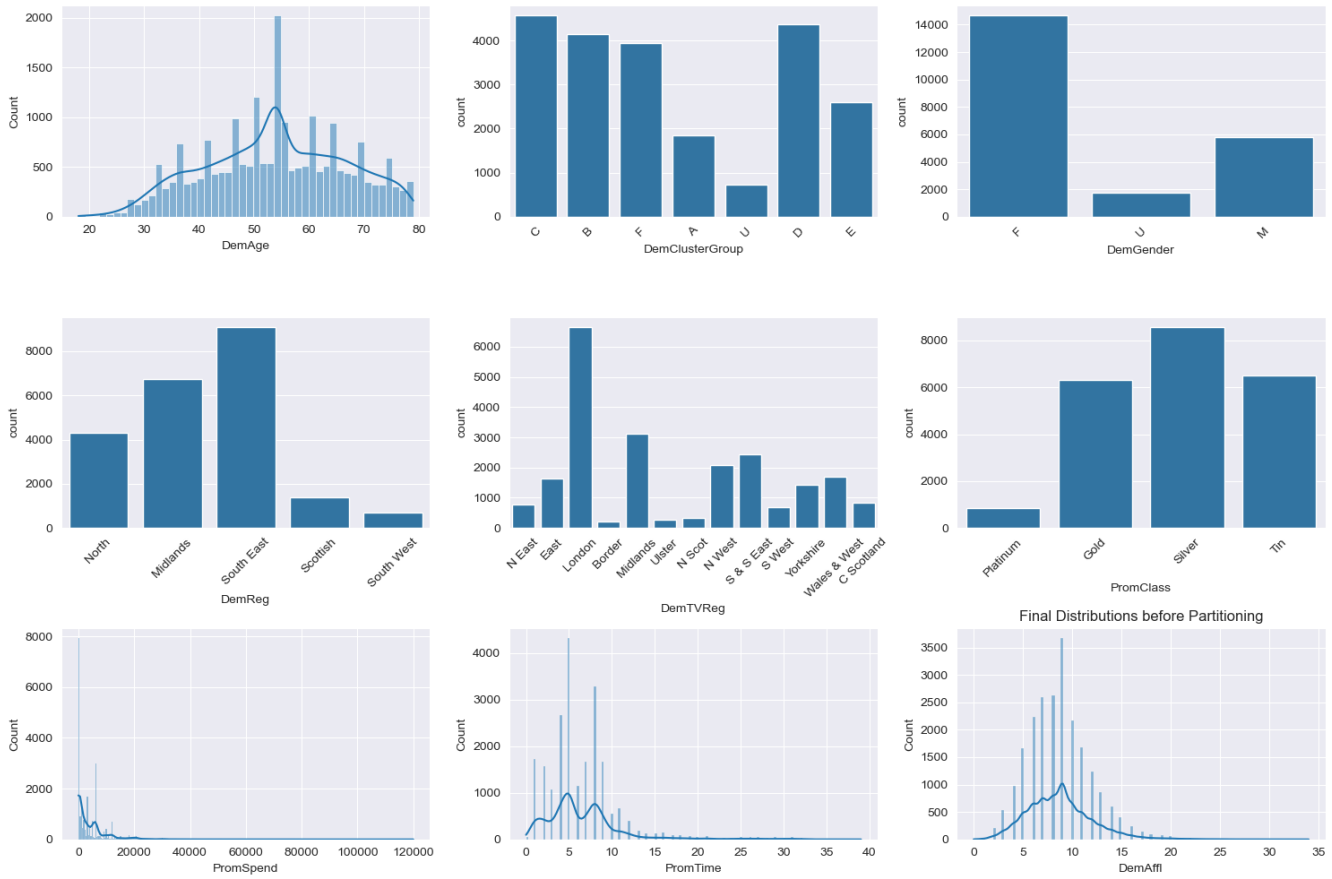
Figure 2. Scatter plots of each one of the continuous variables. Notice the 3 extreme values in PromSpend



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.

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.



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

11306	N_Scot	Scottish
6109	N_East	North
19339	N_West	North
21931	SS_East	South_East
21916	SS_East	South_East

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.

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.

```
Dimensions of Training Datasets (15553, 36) (15553,)
Dimensions of Validation Datasets (3333, 36) (3333,)
Dimensions of Test Datasets (3333, 36) (3333,)
```

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'

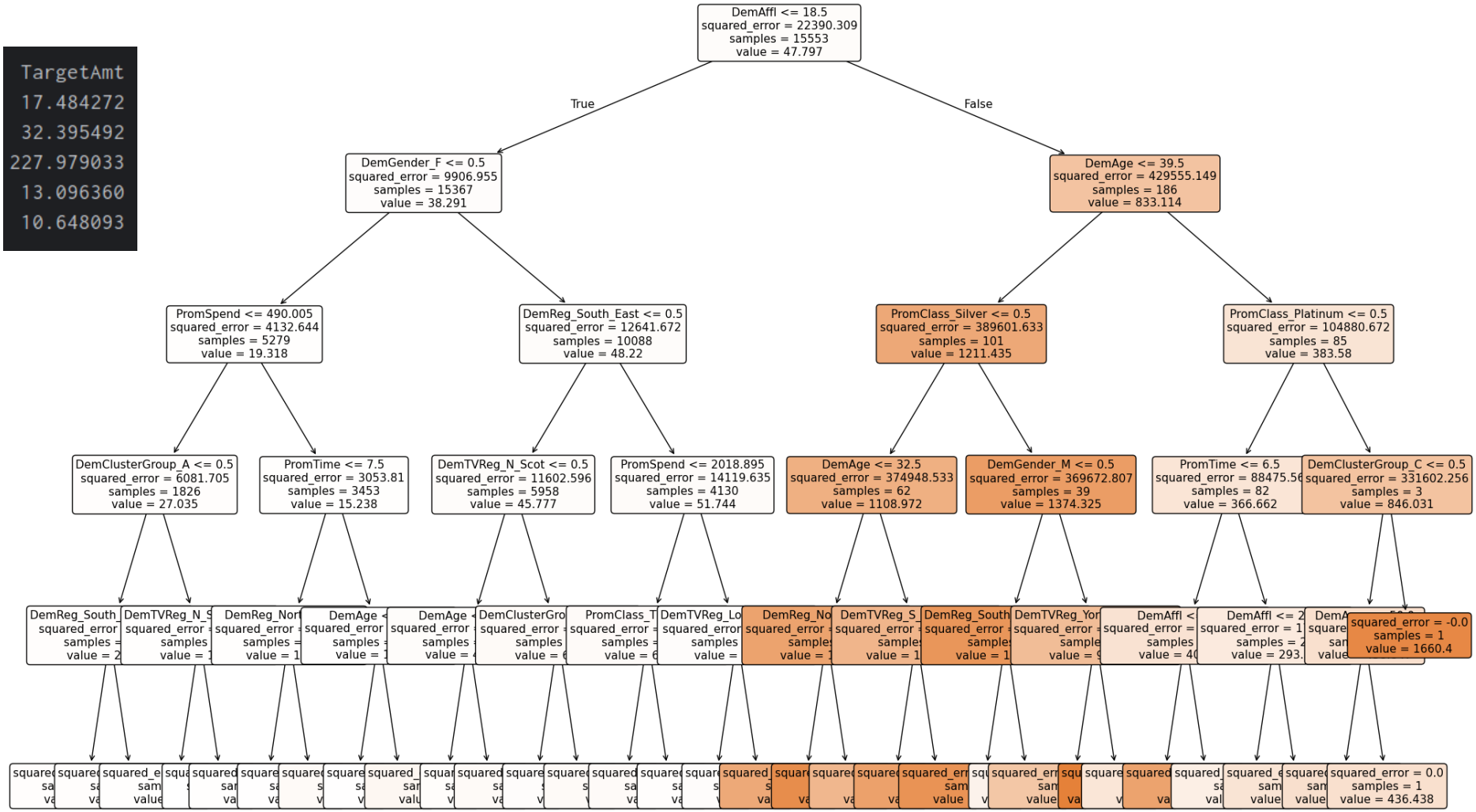
Screenshot 3. Outputs of the models 3 the one with the best performance.

No.	Tree Type	Adjustment	Dataset	MSE	R^2	MAE
1	Unpruned	None	Test	19867.1	0.139	52.924
1	Unpruned	None	Validation	20555.7	-0.11	55.025
2	Pruned	Estimated Pruning Parameters	Test	11612.9	0.497	54.778
2	Pruned	Estimated Pruning Parameters	Validation	10566.4	0.429	54.4
3	Unpruned	RFE (Top 14)	Test	23972	-0.039	54.77
3	Unpruned	RFE (Top 14)	Test	24190.5	-0.306	58.574
4	Pruned	RFE + Pre-pruned	Test	12734.4	0.448	54.542
4	Pruned	RFE + Pre-pruned	Validation	11659.1	0.37	53.385

Scoring

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

TargetAmt
17.484272
32.395492
227.979033
13.096360
10.648093



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

```
Column name: DemReg
Data type: object
Missing values: 0
Unique values: 5
count          22223
unique             5
top       South East
freq            9099
Name: DemReg, dtype: object
Value -> Quantity : South East -> 9099
Value -> Quantity : Midlands -> 6741
Value -> Quantity : North -> 4324
Value -> Quantity : Scottish -> 1368
Value -> Quantity : South West -> 691
```

Annex 2

Main descriptive indicators of the variable DemRegTV

```
Column name: DemTVReg
Data type: object
Missing values: 0
Unique values: 13
count      22223
unique        13
top       London
freq        6654
Name: DemTVReg, dtype: object
Value -> Quantity : London -> 6654
Value -> Quantity : Midlands -> 3123
Value -> Quantity : S & S East -> 2445
Value -> Quantity : N West -> 2096
Value -> Quantity : Wales & West -> 1703
Value -> Quantity : East -> 1649
Value -> Quantity : Yorkshire -> 1443
Value -> Quantity : C Scotland -> 836
Value -> Quantity : N East -> 785
Value -> Quantity : S West -> 691
Value -> Quantity : N Scot -> 329
Value -> Quantity : Ulster -> 266
Value -> Quantity : Border -> 203
```

Annex 3

Main descriptive indicators of the variable DemAffl

```
Column name: DemAffl
Data type: int64
Missing values: 0
Unique values: 33
Quantiles:
0.15     5.0
0.25     6.0
0.50     9.0
0.75    10.0
0.90    13.0
Mean: 8.7259595914143
Median: 9.0
Minimum: 0
Maximum: 34
Standard Deviation: 3.33
```

Annex 4

Main descriptive indicators of the variable DemAge

```
Column name: DemAge
Data type: int64
Missing values: 0
Unique values: 62
Quantiles:
0.15    39.0
0.25    45.0
0.50    54.0
0.75    63.0
0.90    71.0
Mean: 53.81091661791837
Median: 54.0
Minimum: 18
Maximum: 79
Standard Deviation: 12.75
```

Annex 5

Main descriptive indicators of the variable DemClusterGroup

```
Column name: DemClusterGroup
Data type: object
Missing values: 674
Unique values: 7
count     21549
unique        7
top           C
freq       4566
Name: DemClusterGroup, dtype: object
Value -> Quantity : C -> 4566
Value -> Quantity : D -> 4378
Value -> Quantity : B -> 4144
Value -> Quantity : F -> 3949
Value -> Quantity : E -> 2608
Value -> Quantity : A -> 1850
Value -> Quantity : U -> 54
```

Annex 6

Main descriptive indicators of the variable DemGender

```
Column name: DemGender
Data type: object
Missing values: 0
Unique values: 3
count     22223
unique        3
top           F
freq      14661
Name: DemGender, dtype: object
Value -> Quantity : F -> 14661
Value -> Quantity : M -> 5815
Value -> Quantity : U -> 1747
```

Annex 7

Main descriptive indicators of the variable PromTime

```
Column name: PromTime
Data type: int64
Missing values: 0
Unique values: 39
Quantiles:
0.15     2.0
0.25     4.0
0.50     5.0
0.75     8.0
0.90    11.0
Mean: 6.570175043873465
Median: 5.0
Minimum: 0
Maximum: 39
Standard Deviation: 4.627829
Mode: 0    5
Name: PromTime, dtype: int64
```

Annex 8

Main descriptive indicators of the variable PromSpend

```
Column name: PromSpend
Data type: float64
Missing values: 0
Unique values: 2615
Quantiles:
0.15        0.01
0.25        0.01
0.50     2000.00
0.75     6000.00
0.90    12000.00
Mean: 4420.590040948567
Median: 2000.0
Minimum: 0.009999999999999998
Maximum: 296313.85
Standard Deviation: 7559.04752230292
Mode: 0    0.01
Name: PromSpend, dtype: float64
```

Annex 9

Main descriptive indicators of the variable PromClass

```
Column name: PromClass
Data type: object
Missing values: 0
Unique values: 4
count      22223
unique         4
top       Silver
freq        8572
Name: PromClass, dtype: object
Value -> Quantity : Silver -> 8572
Value -> Quantity : Tin -> 6487
Value -> Quantity : Gold -> 6324
Value -> Quantity : Platinum -> 840
```

Annex 10

Main descriptive indicators of the variable TargetAmount

```
Column name: TargetAmount
Data type: float64
Missing values: 0
Unique values: 5371
Quantiles:
0.15      0.00000
0.25      0.00000
0.50      0.00000
0.75      0.00000
0.90    133.80044
Mean: 47.426329009942876
Median: 0.0
Minimum: 0.0
Maximum: 1992.5590105565261
Standard Deviation: 148.0307548913
Mode: 0    0.0
Name: TargetAmount, dtype: float64
--------------------------
```

Annex 11

Dimension of the dataset of predictors. `(22220, 36)`

List of variables created after generating dummy columns.

```
DemAffl
DemAge
PromSpend
PromTime
DemGender_F
DemGender_M
DemGender_U
DemReg_Midlands
DemReg_North
DemReg_Scottish
DemReg_South East
DemReg_South West
```

```
DemTVReg_Border
DemTVReg_C Scotland
DemTVReg_East
DemTVReg_London
DemTVReg_Midlands
DemTVReg_N East
DemTVReg_N Scot
DemTVReg_N West
DemTVReg_S & S East
DemTVReg_S West
DemTVReg_Ulster
DemTVReg_Wales & West
```

```
DemTVReg_Yorkshire
PromClass_Gold
PromClass_Platinum
PromClass_Silver
PromClass_Tin
DemClusterGroup_A
DemClusterGroup_B
DemClusterGroup_C
DemClusterGroup_D
DemClusterGroup_E
DemClusterGroup_F
DemClusterGroup_U
```

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.

No.	Tree Type	Adjustment	Dataset	MSE	R^2	MAE
1	Unpruned	None	Test	19867.1	0.139	52.924
1	Unpruned	None	Validation	20555.7	-0.11	55.025
2	Pruned	Estimated Pruning Parameters	Test	11612.9	0.497	54.778
2	Pruned	Estimated Pruning Parameters	Validation	10566.4	0.429	54.4
3	Unpruned	RFE (Top 20)	Test	20755.1	0.101	52.218
3	Unpruned	RFE (Top 20)	Test	20217.3	-0.092	54.279
4	Pruned	RFE + Pre-pruned	Test	19405.5	0.159	58.891
4	Pruned	RFE + Pre-pruned	Validation	16368.8	0.116	56.858

No.	Tree Type	Adjustment	Dataset	MSE	R^2	MAE
1	Unpruned	None	Test	19867.1	0.139	52.924
1	Unpruned	None	Validation	20555.7	-0.11	55.025
2	Pruned	Estimated Pruning Parameters	Test	11612.9	0.497	54.778
2	Pruned	Estimated Pruning Parameters	Validation	10566.4	0.429	54.4
3	Unpruned	RFE (Top 17)	Test	22247.5	0.036	54.468
3	Unpruned	RFE (Top 17)	Test	23187.7	-0.252	57.31
4	Pruned	RFE + Pre-pruned	Test	17770.1	0.23	58.118
4	Pruned	RFE + Pre-pruned	Validation	14647.8	0.209	55.717

No.	Tree Type	Adjustment	Dataset	MSE	R^2	MAE
1	Unpruned	None	Test	19867.1	0.139	52.924
1	Unpruned	None	Validation	20555.7	-0.11	55.025
2	Pruned	Estimated Pruning Parameters	Test	11612.9	0.497	54.778
2	Pruned	Estimated Pruning Parameters	Validation	10566.4	0.429	54.4
3	Unpruned	RFE (Top 14)	Test	23972	-0.039	54.77
3	Unpruned	RFE (Top 14)	Test	24190.5	-0.306	58.574
4	Pruned	RFE + Pre-pruned	Test	12734.4	0.448	54.542
4	Pruned	RFE + Pre-pruned	Validation	11659.1	0.37	53.385