Predictive Analytics

Prajakta Pohare

1. Cluster Analysis

1.1 Methodology

Customers were divided into two groups using the K-Means clustering technique according to the following criteria:

- Income: The income levels of the customers.
- Recency: How many days have passed since the last purchase?
- NumWebPurchases: The quantity of goods bought via the website.
- NumCatalogPurchases: The total number of catalog-based purchases.
- NumStorePurchases: The total number of in-store purchases.

1.2 Data Preprocessing

1. Handling Missing Data: We excluded rows that contained missing values.
2. Scaling Features: To guarantee comparability across scales, the chosen variables were normalized using z-score standardization.

1.3 Clustering Results

The dataset was divided into two clusters using K-Means clustering with k=2 clusters and 25 random initializations. Each cluster's centroids were computed to provide a summary of its properties:

Cluster 1:

- High income.
- Reduced recency (clients recently made purchases).
- Increased sales in catalogs and online.
- In-store purchases range from moderate to high.

Cluster 2:

- Lower income.
- Higher recency (less recent purchases).
- Fewer web and catalog purchases.
- Lower in-store purchases.

1.4 Evaluation of Clustering

The interpretability of clusters was confirmed by comparing their behavioral characteristics, despite the fact that K-Means clustering by itself does not offer a clear indicator of accuracy. The clusters showed significant income and purchase behavior-based segmentation.

Each cluster's summary was exported for additional analysis.

2. Predictive Analytics

2.1 Methodology

Within Cluster 1, a classification tree was created to forecast consumer reactions. Response, which indicates whether a client reacted favorably or unfavorably to marketing activities, was the response variable utilized for prediction.

2.2 Data Preparation

I. Subsets of training (70%) and testing (30%) data were extracted from Cluster 1.

II. The following independent variables were included in the predictive model:

- Recency
- Income
- NumWebPurchases
- NumCatalogPurchases
- NumStorePurchases

2.3 Model Development

Using the rpart package, a classification tree was created. To increase the homogeneity of responses in each node, the tree recursively divided the data according to the independent variables.

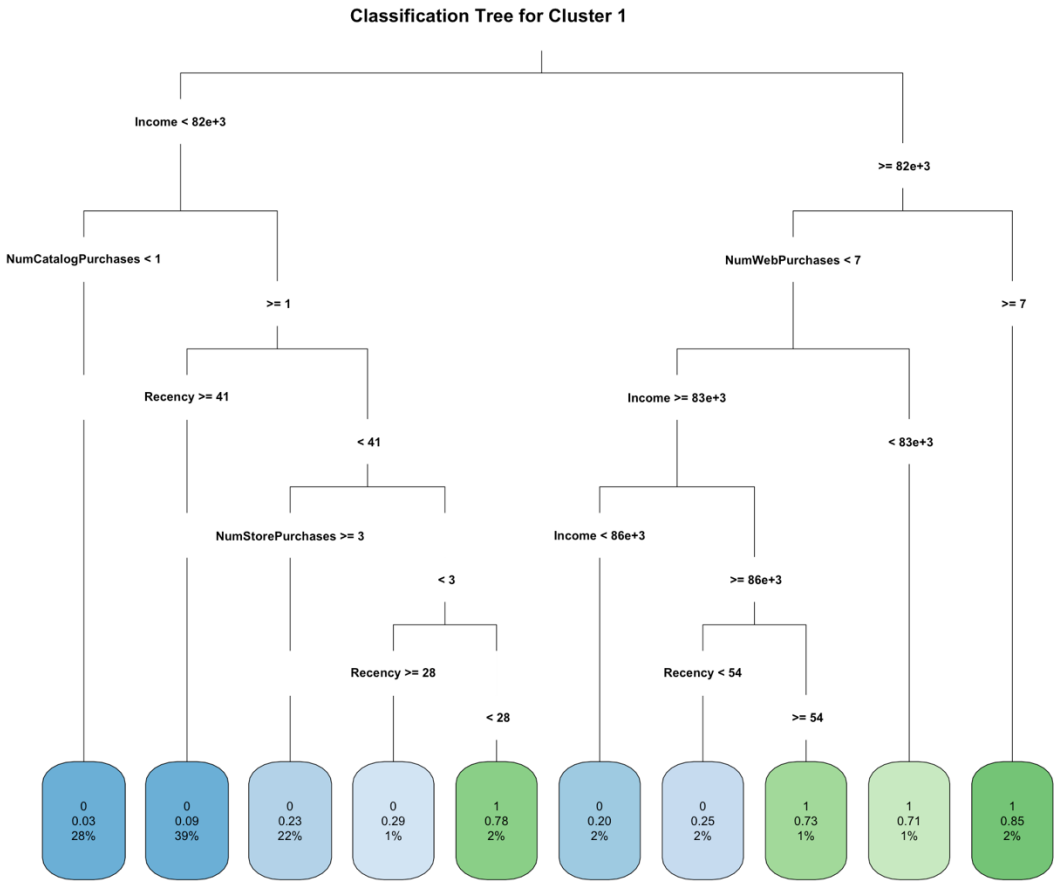
2.4 Significant Variables

The following factors were identified by the tree as being the most important for forecasting client reactions:

- Income: Response patterns varied among customers with higher incomes.
- NumCatalogPurchases: Positive responses were more likely to come from customers who had made more catalog purchases.
- Recency: Respondents were more likely to be customers who had recently made purchases.
- NumStorePurchases. This variable also contributed to differentiating responses.

2.5 Results and Interpretation

The classification tree revealed that:



- Positive responses were less likely to come from customers with lower incomes and fewer catalog purchases.
- Consumers with more recent purchases and greater incomes tended to react favorably.
- Particularly when recency was low, the likelihood of a response rose as the number of retail transactions grew.

2.6 Model Evaluation

The following metrics and a confusion matrix were used to evaluate the model's performance:

- Accuracy: 86.57%
- Sensitivity (Recall): 30.61%
- Precision: 57.69%
- F1 Score: 40%

These findings show that although the model did well overall, it had poor sensitivity, or the capacity to accurately identify positive responses. When the model predicts a favorable reaction, the precision number indicates that it is dependable.

3. Conclusion

Customers were successfully divided into two significant clusters by the analysis, and within one cluster, a predictive model for consumer responses was created. Important revelations include:

- Cluster Characteristics: Income levels, recent purchases, and channel-specific purchasing patterns were used to establish the clusters.
- Significant Variables: The factors that had the biggest effects in forecasting consumer reactions were income, recent purchases, and catalog purchases.
- Predictive Performance: The classification tree's moderate sensitivity and good accuracy suggest room for improvement.

Future research might investigate other clustering strategies, such hierarchical clustering, and improve the predictive model's sensitivity by utilizing ensemble methods like boosting or random forests.