Customer Segmentation for Caravan Insurance

Ray Karpman
May 10, 2022

Marketing Caravan Insurance

- **Hypothetical Client:** insurance company offering variety of policies.
- **Goal:** help design cost-effective strategy to market caravan insurance.
- **Challenge:** caravan insurance is a niche product.
- **Strategy:** customer segmentation.
  - Identify groups of customers likely to purchase Caravan insurance.
  - Design marketing campaigns targeting key groups.

Customer Segmentation with machine learning

- Divide customers into clusters which share similar characteristics, compare rates of caravan insurance ownership between clusters.
- **Strategy:** Use k-means algorithm to find appropriate clustering.
  - Treats customer records as points in many-dimensional space, looks for clusters of close-together points.
- Why use machine learning?
  - Can reveal non-obvious patterns in data.
  - Algorithm may avoid human biases, preconceptions.

Evaluating a clustering

- Unsupervised learning problem. No hard-and-fast accuracy metric.
- Clusters should…
  - Reflect natural patterns in the data.
    - Measure using silhouette score.
  - Be balanced in size.
  - Have meaningfully different rates of caravan insurance ownerships.
  - Have a straightforward description in terms of customer attributes.

Finding and Describing Clusters

- **Step 1:** Use K-means algorithm to find clustering.
  - Check for differences in caravan insurance ownership between clusters.
- **Step 2:** Build decision trees that predict labels for clustered data.
- **Step 3:** Use decision trees to give concise description of each cluster.
  - First priority: simple, interpretable trees.

The Data

- Provided by Sentient Machine Research, retrieved from kaggle.com
- **Size:** 9822 records, 86 variables.
- **Records:** Dutch postal codes.
- **Features:** demographics, socioeconomic status, insurance product usage.
- **Target:** CARAVAN.
    - 1 if at least one caravan insurance policy in postal code, otherwise 0.

Variable names and encoding

- Nominal features.
  - Customer main type: 10 categories, coded as integers.
  - Custerer subtype: 41 categories, coded as integers.
- Ordinal features.
  - Number of houses 1-10
  - Average size of household 1-6
  - Average age, binned by decade.
  - Percent of postal code in demographic groups. (38 features)
    - Binned and coded as integers 0-9.
  - Total contribution to or number of certain insurance policies in postal code. (41 features)
    - Binned and coded as integers 0-9, bins of unequal size.

Data Cleaning

- Removed duplicate rows.
- Capped ordinal features to remove categorical outliers.
    - Ensured that highest, lowest values each had at least 85 categories.
- Removed 16 features which had only one unique value after capping outliers.
- Cleaned data had 8379 records, 69 features.

Feature Selection

- Why feature selection?
    - Data contains highly correlated features.
    - Clustering algorithm may not perform well with too many features.
- featurewiz package (Ram Seshadri, 2020).
    - Removes highly-correlated features, uses MIS with target to decide which features to drop.
    - Selects from list of uncorrelated variables by repeatedly fitting LightGBM model, choosing features important in model.
- featurewiz gave 15 key features.
    - Final model uses top 5 of these predictors by MIS score.

Key Predictors

Name	Meaning
PPERSAUT	Total contribution to car insurance policies
MOSHOOFD	Customer type (10 options)
MINKGEM	Average income
MOPLAAG	Percent with a low level of education
MAUT1	Percent owning one car

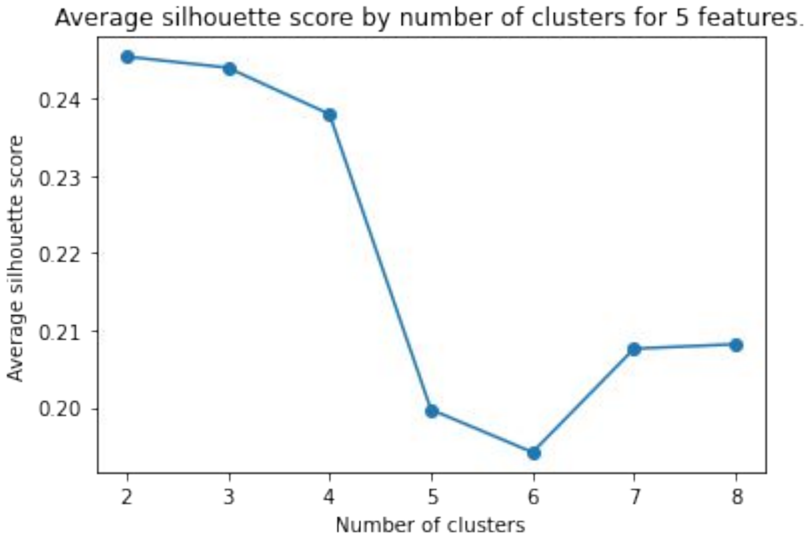
The K-means algorithms

- Start with k randomly chosen points (means)
- Assign each point in dataset to nearest mean.
- Calculate center (mean) of each resulting cluster.
- Repeat this process, with original k points replaced by cluster centers.
- Stop when cluster centers stabilize.

Finding an optimal clustering

- Used k-means algorithm to find natural groupings in data.
- Varied number of clusters, number of features used
  - Number of clusters: 2-8.
  - Number of features: top 5-10 by MIS with target score.
- Primary metric: average silhouette score.
  - Silhouette score measures how close a point is to others in own cluster, compared to other clusters.
  - Ranges from -1 (worst) to +1 (best).

Final Clustering



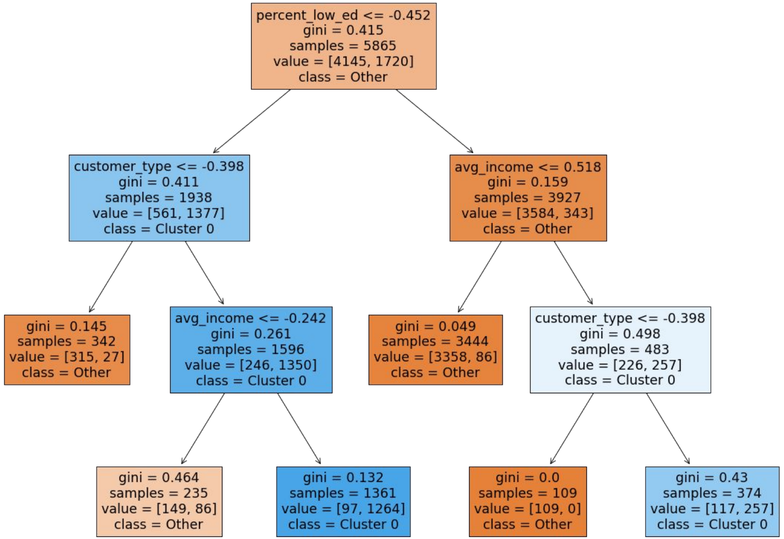
- Used 5 predictors, 3 clusters.
- Second-best silhouette score of any combination tried.
  - Just 2 clusters gives slightly higher score, but less informative clustering.

Caravan ownership by cluster

Cluster	Number of zip codes	% with a Caravan policy
Cluster 0	2408	10.71%
Cluster 1	3008	2.76%
Cluster 2	2963	7.66%

- Differences in rates of Caravan ownership were statistically significant for each pair of clusters, p-value < 0.001.

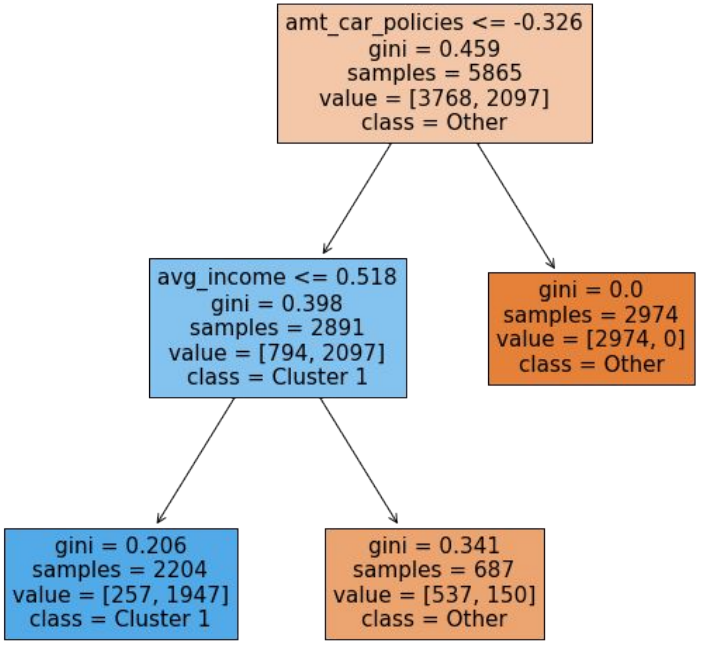
Describing Cluster 0



Cluster 0 Customers:

- Moderate to high education levels.
- From “customer types” more likely to own caravan insurance.
  - **Type 1:** Successful hedonists
  - **Type 2:** Driven growers
  - **Type 3:** Average family.
  - **Type 8:** Families with grownups.
  - **Type 9:** Conservative families.
- Higher average income.
- **Most likely** to own Caravan insurance.

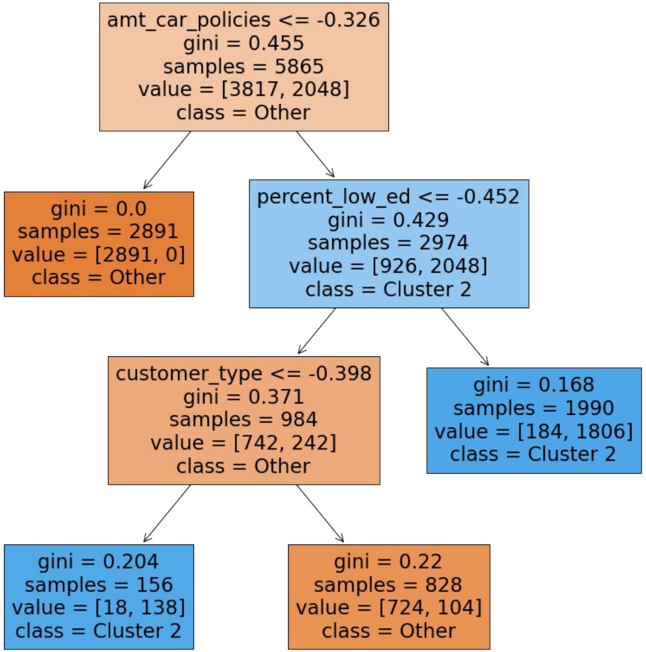
Describing Cluster 1



Cluster 1 Customers:

- Contribute less to car insurance policies
- Lower average income
- **Least likely** to own caravan insurance.

Describing Cluster 2



Cluster 2 Customers:

- Contribute more to car insurance policies
- Lower education levels
- May be from “customer types” less likely to own Caravan insurance.
- **Average likelihood** of owning Caravan insurance.

Summary and Recommendations

- **Primary focus:** Cluster 0.
  - Higher-income, better educated.
    - Advertise in venues that cater to this group.
  - Often family groups or well-off pleasure-seekers.
    - Consider ads showing happy families on vacation.
- **Secondary focus:** Cluster 2.
  - Contribute more to car insurance policies.
    - Consider advertising to current car policy holders.
  - Lower education levels.
    - May be successful blue-collar workers.

Limitations and further work

- Only using 5 predictors, may be missing important information.
  - Check for differences in other features between our clusters.
  - Explore clusterings with more predictors.
  - Try lower threshold for eliminating correlated features.
  - Consult domain expert on key features to retain.
- Clustering is relatively weak. (Silhouette score 0.24)
  - Try different clustering algorithms. (Ex: hierarchical clustering, mean-shift.)