DSE5004 Segmentation and Profiling

Tracey Zicherman

2024-06-19

## Executive Summary

***SIDEKICK: TO DO - Do This Last***

## Methodology

In this project we performed a segmentation and profiling analysis using the following step-wise process:

1. ***Data Cleaning***: The data was cleaned and prepared for manipulation.
2. ***Feature Engineering***: New derivative features were created from preexisting features, including a binary feature identifying high retention customers, and some preexisting features were transformed.
3. ***Exploratory Data Analysis***: Extensive investigation of all features was conducted both univariate analysis (including visualizations of all single feature distributions) and bivariate analysis (including some pairwise distribution visualizations and regression modeling).
4. ***Feature Selection***: A custom subset of customer features was chosen for segmentation. The selection criteria employed were chosen to facilitate analysis, segmentation methods and profiling.
5. ***Supervised Segmentation***: Candidate customer segments were generated using a supervised method, with the binary feature HighRetention used as the target. The decision tree algorithm was employed to segment the feature space into 8 segments, corresponding to the leaves in the decision tree. The best fitting “pruned” tree was selected, for an optimal balance between relative error and complexity. The decision rules used by the tree algorithm were used on the data to create the corresponding segments.
6. ***Unsupervised Segmentation***: Candidate customer segments were generated using an unsupervised method, i.e. with no target feature in mind. The *k*-means algorithm was used to detect segments of customers similar to each other (“nearby”) in the numerical segment feature space. The elbow method was used to select a good number of segments, and the resulting segments used to create customer segments.
7. ***Segmentation Evaluation and Selection***: The two segmentation methods were evaluated, both individually, and with respect to each other, with several statistical metrics used to approximate each method’s ability to generate useful segment. The results of this evaluation process were used to select the better segmentation method, which was determined to be *k*-means.
8. ***Segment Profiling***: The segments generated by *k*-means were used to generating segment profiles. Specifically, the properties of these segments were investigated, specifically with respect to high and low retention customers, and corresponding characteristics identified and discussed.

### Preprocessing

The original dataset consisting of 5000 customers and 60 customer features. During the preprocessing phase, these features were subjected to standard checks, cleaned and prepared for modeling.

The following preprocessing steps are worthy of note:

1. Missing values of the internal company features DataLastMonth, DataOverTenure, EquipmentLastMonth, EquipmentOverTenure, VoiceOverTenure were set to 0. Missing values of other variables were imputed with the standard strategies of using the mode for categorical features and the mean for numerical features.
2. Features Internet and HomeOwner were recoded as ordinal integer features with values 1,…,5 and 0, 1, respectively. The integer ordinal was chosen for Internet to correct for inconsistencies in labelling, and intended to reflect tiers of internet services offered by the company.

### Feature Grouping

As a useful conceptual strategy, to aid the exploration and selection, given the large number of features, the following grouping of features into categories was used:

### Feature Engineering

New derivative features were then created, and preexisting feature transformed. Of particular interest was the preexisting feature PhoneCoTenure, which indicated the number of months a customer has been with the company. We used this feature to identify long-term vs short-term, i.e. high-vs-low retention customers by creating a new feature HighRetention, indicating which customers had a tenure greater than the 75% percentile, namely 59 months.

***SIDEKICK: TO DO - PhoneCoTenure distribution goes here***

The following five new derivative features were also created:

* TotalDebt = CreditDebt + OtherDebt: Total customer debt.
* AvgCardSpendMonth = CardSpendMonth/CardItemsMonthly: Average monthly credit card spending per item. Set to 0 if CardItemsMonthly == 0.
* AvgValuePerCar = CarValue/CarsOwned: Average value per car owned. Set to 0 if CarsOwned == 0.
* TechOwnership = OwnsFax + OwnsGameSystem + OwnsMobileDevice + OwnsPC: Number of technological items owned out of 4 possible (here the binary ownership features are 0/1 encoded)
* NumAddOns = Multiline + Pager + ThreeWayCalling + VM: Number of account add-ons out of 4 possible (again the binary add-on features are 0/1 encoded)

In addition, due to highly skewed distributions, the four features CardSpendMonth, DataOverTenure, VoiceOverTenure, HHIncome, CarValue. were transformed to standardized versions, i.e. they were transformed by subtracting the mean and dividing by the standard deviation). Although these features were not explicitly used during the segmentation process, they were effectively used, as all segmentation features were numeric and standardized.

It made sense to treat missing values as zero. In particular, we choose to do this for DataLastMonth, DataOverTenure, EquipmentLastMonth, EquipmentOverTenure, VoiceOverTenure, since presumably the company has access to this information, and if it isn’t present, we assumed it can be treated as 0.

Finally for the remaining features, we use a standard imputation strategy, using the mode for categorical features and mean values for numerical features.

### EDA and Feature Selection

The preexisting and engineered features were used for exploratory data analysis, in which a subset of useful customer features was identified for use in segmentation and profiling.

Overall, the ability to provide useful and potentially novel insights through segmentation and profiling was the main criteria in feature selection. In particular, it was surmised that stakeholders and decision makers may be interested in identifying customer that may end up having a long tenure but that currently do not.

For that reason, certain variables with potentially useful information about internal customer behavior over a long tenure were omitted, namely DataOverTenure, EquipmentOverTenure, and VoiceOverTenure, that is, feature whose future values over a long tenure would be currently unknown.

Moreover, other time-dependent features such as Age, Employment were identified as potentially confounding, that is, features with strong associations to long tenure (and thus to the derivative retention feature), and were also omitted.

Finally, the specific interest in using *k*-means segmenting as an unsupervised method, due to its ability to detect unknown patterns, was the next most important criteria, and had a large impact on the choice of features. Primarily, it resulted in a choice of purely numeric features for the segmentation process. Then, using the segments thus constructed, categorical feature characteristics for high and low retention customers were identified and discussed.s

With these criteria in mind, the following hand-picked selection of fifteen customer demographic, behavioral and financial features was used in out our segmentation

# customized set segmentation targets and features  
seg\_target <- "HighRetention"  
seg\_feats <- c("CommuteTime", "HouseholdSize", "TownSize", "CardItemsMonthly",  
 "DebtToIncomeRatio", "HHIncome", "CarsOwned",   
 "TVWatchingHours", "Region", "TotalDebt",  
 "CardSpendMonth", "HHIncome", "CarValue", "TechOwnership", "NumAddOns")

### Segmentation Methods

#### Supervised Decision Tree Segmentation

The aim of supervised learning is in general to find meaningful associations between the features and the target. In this case, we employed a supervised-learning-based segmentation method in the hopes of finding a detectable useful pattern between the custom customer features selected for segmentation and the target customer feature HighRetention, thus capturing a meaningful association between segments and high and low value customer segments.

##### Decision Tree Diagram

A decision tree algorithm was used to discover statistically meaningful decision rules among the numerical segmentation features. The rules are produced by the algorithm by optimizing a (mathematically defined) criterion which essentially measures how well the rules classify the observations according to the target feature, in this case, the binary customer retention feature HighRetention.

The algorithm is so-named because the resulting decision rules can be seen as a partition of the feature space into segments, or alternatively, as a sequence of choices about how to place observations (in this case customers) into segments, and these rules can be easily visualized in a tree diagram.

Several decision trees were fit, and the optimal decision tree was chosen which balanced model complexity and accuracy. The inclusion of model complexity in this choice helps improve the expected ability of the fit to generalize to unseen data, that is, to future customers.

The decision rules obtained by the optimal decision tree were used to number all leaves in the tree diagram in order from left to right (note all nodes contain at least one observation). These are the segment labels, and we can assign observations to these segments based on the decision rules.

***SIDEKICK: TO DO - Decision Tree Diagram goes here***

Note that the decision rules corresponding to this decision tree only involve a small subset of the segmentation features. The aforementioned complexity minimization criteria employed, which potentially helps reduce generalization error (and thus, reflects a true relationship between features and response, rather than a spurious artifact of the given dataset), often results in such a “pruned tree”.

For this reason, the resulting decision tree segmentation was seen as perhaps less than ideal, given that potentially useful information contained in the other features wasn’t utilized.

##### Feature Importances

We include here a plot of segmentation features ranked by importance with respect to their use by the decisions trees from which the optimal tree was selected (we omit a more technical discussion of how these features are selected (that is, what precisely “important” means), and invite the curious reader to research the topic further).

***SIDEKICK: TO DO - Feature Importances Plot goes here***

Note that the most important features are those seen in the decision rules for the pruned tree.

##### Plots and Analysis

***SIDEKICK: TO DO - Add Distribution Plot***

***SIDEKICK: TO DO - Add Stacked Bar Plot***

#### Unsupervised *k*-Means Segmentation

In general, unsupervised learning methods are employed to capture novel or unexpected relationships amongst features and observations, that is, without the assumptions implicit in the selection of a special “target” to associate the remaining features to. The goal that at least some of these segments should provide some insight about high and low retention customers, that is, that some segments should prove to contain more high or low retention customers than others.

In our case, in order for the method to remain truly unsupervised, we wished to suppress any information related to customer retention in the learning algorithm. Accordingly, the tenure-related features PhoneCoTenure and HighRetention were omitted. The hope was that in doing so, the resulting segmentation would contain meaningful information about high and low customer retention *thus adding weight* to the segmentation pattern thus discovered (since it contained no assumptions or information about retention, and yet such associations were discovered independently).

We mentioned previously that *k*-means segmenting can only use numerically encoded features. This is because it relies on a notion of distance between points in the feature space, which does not apply to non-numeric features.

##### Scale Data

Given that the features are measured on vastly different scales, it is possible that large scale features can have undue influence on the resulting *k*-means segmenting. Following standard procedures, the segmentation features were thus standardized to reduce this risk.

##### Elbow Method for Selecting Number of Segments

Naively when using *k*-means to perform segmentation, the practitioner must choose the number *k* of segments beforehand, but it is usually better if this choice can be informed. The elbow method exists for this purpose. A plot of *k* versus a measure of “homogeneity” of the segmented observations is generated – specifically, this is “total within-segment variation”.

The elbow method looks for a elbow (kink) in the *k* vs total within segment variation graph, and selects the minimum just before the kink. Similar to decision tree pruning, this choice of *k* is thought to provide a nice balance between model complexity, in this case, as measured by number of segments,and accuracy, in this case, as measured by how well the segmented observations “group together”, i.e. how low the variation is within the segment.

Note that the decision tree segments a

***SIDEKICK: TO DO - Elbow Plot Goes Here***

There is no clear kink, so we are freer to choose a value of ourselves, we will select . This still balances complexity with the need to detect differences between segments, and provides more fine-grained information when considering high-vs-low retention customers than a smaller number of segments.

Furthermore, a good deal of trial-and-error, and justified this choice of *k* by revealing that it provided a good separation of high vs. low retention customers by segment, as well as a good separation of features, as determined by the variance (spread) across segments of the segment means for each feature (more below).

Incidentally, this choice of *k* was only weakly informed by the fact that it somewhat facilitated direct comparison with decision tree segments, although not explicitly, since the statistical comparison measures (primarily variance of segment means) would just have easily worked for different numbers of segments.

##### Plots and Analysis

***SIDEKICK: TO DO - Add Distribution Plot***

***SIDEKICK: TO DO - Add Stacked Bar Plot***

## Findings

### Evaluate Segmentation Solutions

In order to evaluate and compare the segmentation solution, we relied on the following evaluation criteria:

1. ***Segment Feature Space Utilization***: How well does the segmentation method make use of the available features?
2. ***Segment Separation***: How well does the segmentation methodseparate different clusters from each other?
3. ***High and Low Retention Discrimination***: How well does the segmentation method allow us to differentiate high and low retention customers by segment, that is, how well does it place and high retention customers into some segments, and low retention customer into others?

#### Feature Space Utilization

As mentioned previously, the decision tree model only used 5/15 33% of the total number of features, whereas *k*-means intrinsically makes use of all features.

***Conclusion***: Due to the omission by the decision tree model of 10/15 67% of the total number of features, *k*-means has clearly better segment feature space utilization.

#### Segment Separation

To measure the degree to which the segments are well-separated from each other, we use two statistical measures of separation, the total and average variance (across segments) of the segment means, which we call “total separation” and “average separation”. Specifically, this is the sum and the average of the variances across all features of the featire means across all 8 segments.

The reasoning for the use of this measure is as follows. Since the mean of each feature is a measure of it’s “center” (and indeed, the centroid of each segment is the vector of feature means), the variance of means capture how far the within-segment means of each feature are from their overall center.

The sum and average variance of the feature means taken over all features captures how far the within-segment feature centers are from their overall center, and thus ostensible, from each other.

***Conclusion***: *k*-means is clearly better at separating segments, with higher total and average segment separation.

#### High and Low Retention Discrimination

To determine how well the segmentation methods separate high and low retention customers, we look at the overall picture provided by proportion of high segmentation customers per segment, per method. When considering these results, note that there is no natural correspondence between the numbers assigned to each segment by each method.

We notice that *k*-means appears better able to separate high retention customers, with segments 6 and 8 having roughly 46% and 49%, respectively. The decision tree also has two segments with high percentages of high retention customers, namely 2 and 4, but these are lower, at roughly 40% and 44% respectively.

The decision tree appears better at separating low retention customers than *k*-means. For *k*-means, only one segment has a low percentage of high retention customers (hence a high percentage of low retention customers), namely segment 4 at roughly 8%, while for the decision tree there are 4 segments with low percentages of high retention customers, between roughly 8-11%.

***Conclusion***: These results are somewhat mixed, however, given the potentially higher value in identifying high retention customers, we give the advantage to *k*-means.

#### Comparison and Segmentation Method Selection

After careful investigation, it was determined to use the k-means segmentation for the following reasons:

1. Better feature space utilization, at 100% vs. 33% of available features.
2. Better segment separation, as measured by higher sum and average variances of segment means of all features collectively.
3. Better high and low retention separation of customers into individual segments.

### Segmentation with Preferred Solution

Having selected -means segmentation, the resulting eight segments were visualized and investigated, and the results used to build the corresponding customer profiles.

In this report, some general observations are made about the eight segments and their corresponding profiles, and provide visualizations. The discussion then focuses on the two high retention segments and one low retention segment.

### Overview of Segmentation Results

***SIDEKICK: TO DO - Overview of Results goes here***

See appendix B for summary statistics on the customer profiles, namely, the median values of the numerical segmentation features and the mode of categorical features.

***SIDEKICK: TO DO - Additional Plots Go Here***

### High and Low Retention Segments

As mentioned. Segments 6 and 8 had much higher retention than other segments, about 47% and 49%, respectively. Segment 4 had much much lower retention than other segments, at approximately 9% \*\*\*

## Conclusions

## REMOVE: Summary Statistics of Decision Tree and k-Means Segments

***SIDEKICK: TO DO - Remove This From Final Report!*** ***SIDEKICK: TO DO - Remove This From Final Report!***

## REMOVE: Detailed Summary Statistics for k-Means Segments

***SIDEKICK: TO DO - Remove This From Final Report!***

## Appendix A: High and Low Retention Segment Summary Stats

### Numerical Features

### Categorical Features

## # A tibble: 3 × 5  
## kSeg CustomerID\_mode Gender\_mode JobCategory\_mode MaritalStatus\_mode  
## <fct> <chr> <chr> <chr> <chr>   
## 1 4 4459-VLPQUH-3OL Female Sales Married   
## 2 6 8959-RZWRHU-ST8 Female Professional Unmarried   
## 3 8 0649-TBFJFL-QU4 Male Labor Married

## # A tibble: 3 × 5  
## kSeg PoliticalPartyMem\_mode Retired\_mode UnionMember\_mode LoanDefault\_mode  
## <fct> <chr> <chr> <chr> <chr>   
## 1 4 No No No No   
## 2 6 No No No No   
## 3 8 No No No Yes

## # A tibble: 3 × 5  
## kSeg LoanDefault\_mode CarBrand\_mode CarOwnership\_mode HomeOwner\_mode  
## <fct> <chr> <chr> <chr> <chr>   
## 1 4 No Domestic Own Yes   
## 2 6 No Foreign Own Yes   
## 3 8 Yes Domestic Own Yes

## # A tibble: 3 × 5  
## kSeg OwnsFax\_mode OwnsGameSystem\_mode OwnsMobileDevice\_mode OwnsPC\_mode  
## <fct> <chr> <chr> <chr> <chr>   
## 1 4 No Yes Yes Yes   
## 2 6 No No No Yes   
## 3 8 No Yes Yes Yes

## # A tibble: 3 × 5  
## kSeg ActiveLifestyle\_mode NewsSubscriber\_mode Votes\_mode CallerID\_mode  
## <fct> <chr> <chr> <chr> <chr>   
## 1 4 Yes No No No   
## 2 6 No Yes Yes Yes   
## 3 8 No Yes Yes Yes

## # A tibble: 3 × 5  
## kSeg CallForward\_mode CallingCard\_mode CallWait\_mode EquipmentRental\_mode  
## <fct> <chr> <chr> <chr> <chr>   
## 1 4 No Yes No No   
## 2 6 Yes Yes Yes No   
## 3 8 Yes Yes Yes No

## # A tibble: 3 × 5  
## kSeg EquipmentRental\_mode Multiline\_mode Pager\_mode ThreeWayCalling\_mode  
## <fct> <chr> <chr> <chr> <chr>   
## 1 4 No No No No   
## 2 6 No Yes No Yes   
## 3 8 No Yes Yes Yes

## # A tibble: 3 × 5  
## kSeg VM\_mode WirelessData\_mode CreditCard\_mode EBilling\_mode  
## <fct> <chr> <chr> <chr> <chr>   
## 1 4 No No Visa No   
## 2 6 No No Disc No   
## 3 8 Yes Yes Disc No