Segmentation and Profiling

Tracey Zicherman

2024-06-27

Executive Summary

In storing customer data, it is generally well known and accepted that the company is in possession of valuable but unrefined information about customer characteristics and behavior that is waiting to be analyzed and transformed into actionable intelligence.

In particular, a valuable way of making use of this information is the development of customer segments, categories into which customers tend to belong. These segments are determined through modeling, in such a way that customers within segments tend to share similar characteristics with each other, and have different characteristics between the segments. From these segments, profiles can be developed regarding particular tendencies of customers within each segment, which can be used to inform strategic company decisions in a number of areas, from marketing strategies to retention efforts.

This project developed and profiled customer segments with the aim of understanding customer retention, making use of information from company data stores on 5000 customers, with 60 features measured in total. The data was carefully checked and processed to ensure it was clean before further investigation and analysis. During this process, additional derivative features were created, in particular, a feature related to customer tenure was used to create a high retention feature, which identified high retention customers as those with tenure time greater than or equal to the 75% percentile (59 months) out of the maximum seen (72 months).

Two segmentation methods were compared, a supervised method, which had access to the high retention feature in order to learn a statistical relationship between this and the other customer features, and an unsupervised method, which did not have access to the high retention feature, enabling it to discover unforeseen patterns amongst the other features.

The second, unsupervised method, proved better at customer segmentation as measured by three key criteria: its ability to make use of all available features, its ability to clearly differentiate customer segments, (i.e. minimize their overlap), and to concentrate high and low retention customers within segments. The segments thus identified were then

employed to develop customer profiles. The profiles thus developed painted a nuanced picture of customer behavior and characteristics.

The lowest retention segment, segment 4, had only a 9% retention rate, meaning that 91% of customers in this segment had tenure less than 59 months. The profile that emerged from this segment was of large households (presumably families with children), with lower news consumption, lower political participation, low income (median $30K), lower education, but also lower spending and debt.

Segment 4 customers tended to be employed in sales and to belong to unions more frequently than other segments. They also tended to have lower adoption of technology, owning fewer devices, and having fewer service add-ons and displayed more active lifestyles.

The first high retention segment, segment 6, had a 46% retention rate. This profile tended to consist of couples with much higher income (median $137K), almost 4 times the median as the low retention segment 4, and had higher rates of retirement, news consumption and political participation. Customers with this profile tended to have a better overall financial picture, displaying the lowest levels of total debt and credit debt, while maintaining high spending.

The highest retention segment, segment 8, had a 49% retention rate. This profile also tended to consist also of couples with the highest income levels (median $249K), lowest retirement rate (0%), and had high rates of news consumption and political participation but also much higher spending and debt levels. They tended to have the highest rates of technology adoption and service add-ons, indicating high value as customers.

Depending on company priorities, an obvious strategy is to target marketing efforts on customers in segment 4 to boost retention efforts, and to focus retention efforts (for example, through rewards programs) on customers in segment 6 and 8. Another natural strategy could be to target marketing efforts towards potential customers sharing segment 6 and 8 characteristics, thinking of these as potential “high value customers”.

There is an important caveat in interpreting these results, namely the use of customer tenure, measured in months, to determine which customers are low versus high retention. Clearly, it is unknown, without additional information, whether a customer has a low tenure because they have not been with the company long, or because they left quickly.

Therfore, the available customer features revealed a significant, intrinsic limitation. There was unfortunately no feature which could provide the necessary information to distinguish genuine customer attrition, that is identify customers who leave soon versus customers who have just joined.

While this project made the best use of what was available, it is highly recommended that the company track additional customer features, primarily features related to *when* customer engagement with the company first began (for example, beginning of subscription or contract services), when it became dormant (for example, end of services) and when a customer can be considered to have churned (e.g. time in database while inactive). Such additional data could easily improve efforts to identify long versus short term and high versus low retention customers and potentially dramatically improve future customer segmentation and profiling efforts.

Methodology

In this project we performed a segmentation and profiling analysis with the following steps:

1. ***Data Cleaning***: The data was cleaned and prepared for manipulation and modeling.
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 segments. 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 generate 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 consisted of 5000 customers and 60 customer features. During the preprocessing phase, these features were subjected to standard checks, cleaned and prepared for modeling. Missing values were imputed using the standard strategies of using the mode for categorical features and the mean for numerical features.

Feature Engineering

New derivative features were then created, and preexisting features transformed. The pre-existing feature `PhoneCoTenure` indicated the number of months a customer has been with the company, and was used 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.

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)

EDA and Feature Selection

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

Overall, useful and potentially novel insights through segmentation and profiling were the main criteria for feature selection. In particular, it was assumed stakeholders and decision makers may be interested in identifying customers 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, features 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.

The specific interest in using k-means segmentation as an unsupervised method, due to its ability to detect unknown patterns, was the next most important criteria, and also 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 constructed, categorical feature characteristics for high and low retention customers were identified and discussed.

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
CommuteTime,HouseholdSize,TownSize,CardItemsMonthly,DebtToIncomeRatio,HHIncome,CarsOwned,TVWatchingHours,Region,TotalDebt,CardSpendMonth,HHIncome,CarValue,TechOwnership,NumAddOns
```

Segmentation Methods

Supervised Decision Tree Segmentation

The decision tree supervised-learning-based segmentation method was used to find detectable useful patterns between customer features selected for segmentation and the target customer feature `HighRetention`, thereby potentially 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 and association with the binary customer retention feature `HighRetention`.

Several decision trees were fit, and the optimal decision tree was chosen which balanced model complexity and accuracy.

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, used to assign observations to these segments based on the decision rules.

Note that the decision rules corresponding to this decision tree only involve a small subset of the segmentation features. 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.

Unsupervised k-Means Segmentation

To keep the k-means segmentation truly unsupervised, we wished to suppress any information related to customer retention in the learning algorithm. Accordingly, the tenure-related features `PhoneCoTenure`, `HighRetention` were omitted. The intention was that the resulting segmentation should contain meaningful information about high and low customer retention thus adding weight to the segmentation pattern being discovered , since it contained no assumptions or information about retention, and yet such associations were discovered independently. This indeed turned out to be the case.

Scale Data

Given that the segmentation features were measured on vastly different scales, following standard procedures, the segmentation features were standardized.

Elbow Method for Selecting Number of Segments

The elbow method for selecting k was employed, and the value $k = 8$ selected, balancing 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. A good deal of trial-and-error justified this choice, revealing 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).

See the appendix for plots related to decision tree and k-means segmentation.

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 method separate 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 low and high retention customers into some segments, and low and high retention customers into others?

Feature Space Utilization

The decision tree model only used $5/15 \approx 33\%$ 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 $10/15 \approx 67\%$ 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 feature means across all 8 segments.

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 therefore ostensible, from each other.

	Decision Tree	***k* Means**
Total Segment Separation	5.342	10.956
Average Segment Separation	0.382	0.783

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 appear 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.

Tree Segment	% High Retention	kMeans Segment	% High Retention
4	43.5	8	48.8
2	40.0	6	45.9
8	28.0	1	27.2
3	23.7	2	26.4
5	11.8	5	25.8
7	10.4	3	23.3
1	10.4	7	19.7
6	8.3	4	8.6

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 across all features.
3. ***Better high and low retention separation of customers*** into individual segments.

Segmentation with Preferred Solution

Having selected k-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

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

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%. Profiles were developed for these segments, with features grouped by conceptual similarity.

We highlight notable takeaways from this profiling, and refer the reader to the appendix for detailed tables of results (for ease of interpretation, figures were rounded to the nearest integer value). In particular, we omit discussion of profile features related to company internals (service add-ons and account features).

Demographics

Segment 4 was predominantly male (56%), unmarried (91%), had lower amounts of education (median 14 years). The overwhelming majority were not retired (97%) and the most frequently occurring employment type was sales (37%). A small percentage were union members (16%), while minorities subscribed to news (35%), voted (46%) and were members of a political party (36%). This segment had the largest households by far, with a median of 5 members per household.

Segment 6 was gender balanced (50/50%), majority married (52%), and had higher amounts of education (median 16 years). Again, the overwhelming majority were not retired (97%) and the most frequently occurring employment type was professional (32%). A similar percentage were union members (15%), while a majority subscribed to news (65%), voted (63%). A similar percentage were political party members (40%), and households were smaller, with a median of 2.

Segment 8 skewed female (59%), majority unmarried (47%), and was more educated (median 16 years). None were retired (100%) and the most frequently occurring employment type was labor (32%). Fewer were union members (13%), while a majority subscribed to news (68%), voted (69%). A similar percentage were political party members (43%), and households were also small, with a median of 2.

Ownership

All segments 4, 6, 8 were majority home (86%, 99%, 99% respectively) and car owners (86%, 99%, 99%), though the percentages for high retention segments 6, 8 were much higher. All owned similar numbers of cars (median 2, 2.5, 2) though segment 6 owned slightly more. Car values were much different across segments, with segment 4 having much lower value (median $16.2K) than segment 6 (median $68.5K) and segment 8 (median $87.5K), which had the highest value.

Behavior

Segment 4 was much more active than 6, which was more active than 8 (58%, 39%, 33% respectively), and although all watched similar amounts of television segment 4 watched slightly less (median 20, 21, and 21 hrs). All segments showed similar levels of pet ownership (median 2 pets).

Finances

Segment 4 had much lower levels of income than segments 6 and 8 ($30K, $137K, $249K). Segment 6 had the lowest levels of total debt, while segment 8 had much higher levels (median $2.2K, $1.1K, $38.8K, for 4, 6, 8 respectively), and the same pattern held for credit debt (median $710, $319, $1259, for 4, 6, 8) and percentage in loan default (34%, 15%, 60%). Segment 4 had the lowest levels of monthly credit card spending compared to 6 and 8 (median $2.5K, $4.5K, $4.7K), while 6 purchased slightly more credit card items per month (med 10, 11, 10). It is worth noting that segment 6 appears to have the soundest finances, when considering both income spending and debt levels.

Technology

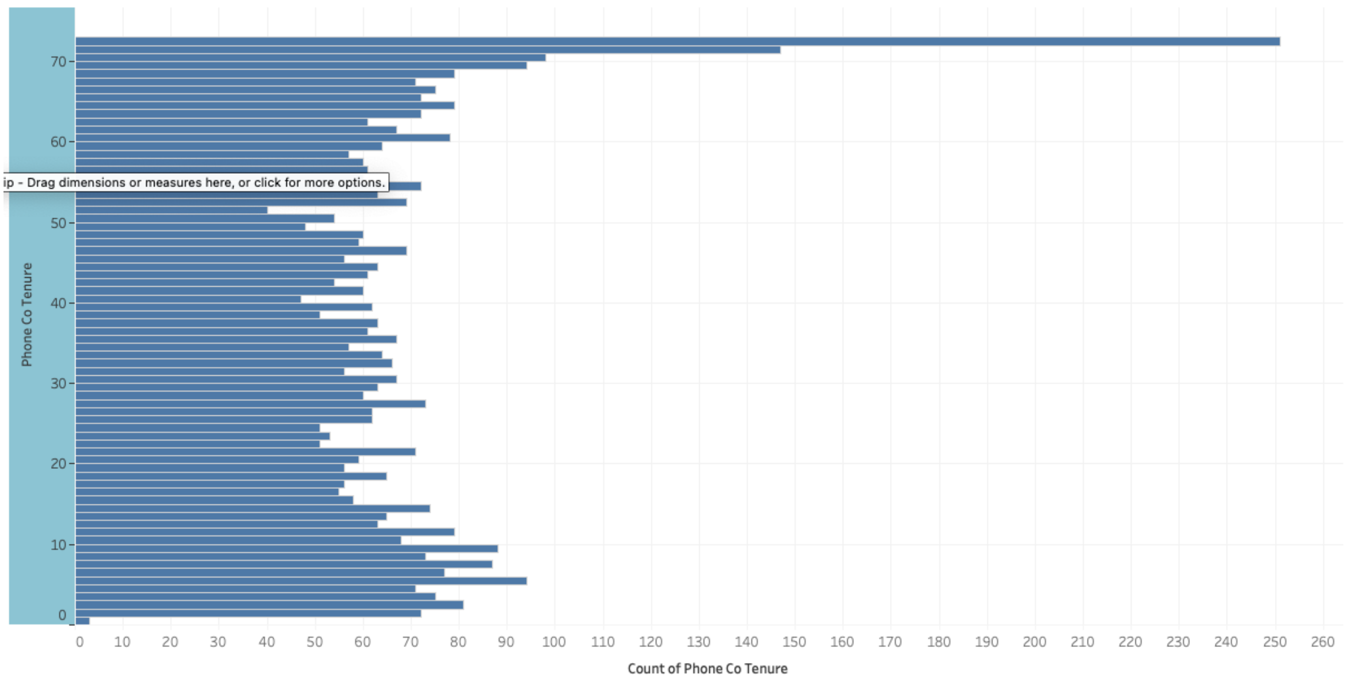
Segment 8 showed the highest levels of overall ownership of technological devices (med 2, 2, 2.5) with the highest ownership of game systems (54%, 50%, 60%), PCs (66%, 69%, 76%), mobile devices (52%, 47%, 59%) and fax machines (12%, 24%, 48%)

Geography

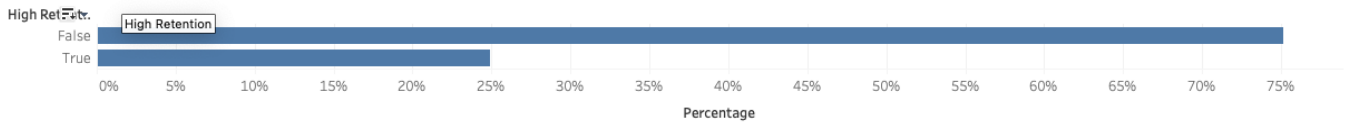
Segment 4 appearing in region 1 tended to live in smaller towns (town size 2) than those in regions 2,3,4 (size 3) and 5 (size 4), suggesting these customers were concentrated in urban areas in region 5. They displayed similar commute times across the regions 1-5 (median 25, 26, 26, 25, 26 min). Segment 6 appeared to favor similar sized towns in all 5 regions (2, 2, 3, 3, 3) and had smaller commutes in regions 1-4 than 5 (24, 24, 24, 23, 27). Segment 8 favored small towns in region 1 and larger towns in region 4 (1, 3, 2, 4, 3), with the commute time pattern corresponding to town size (med 23, 26, 25, 29, 25). High commute times for segment 8 in large towns in region 4 suggests predominantly suburban.

Appendix: Distribution Plots for PhoneCoTenure and HighRetention

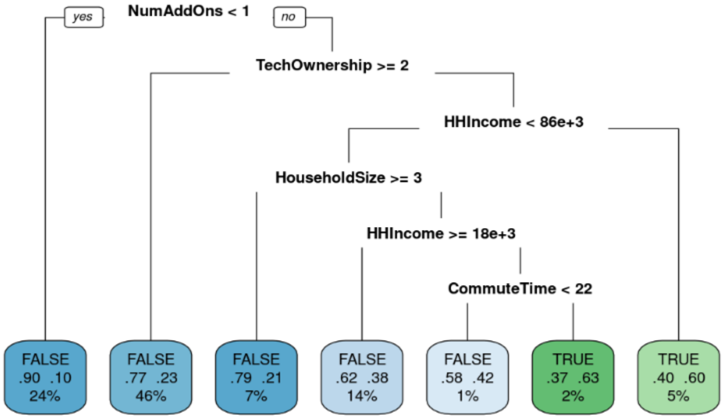
Histogram of Customer Tenure

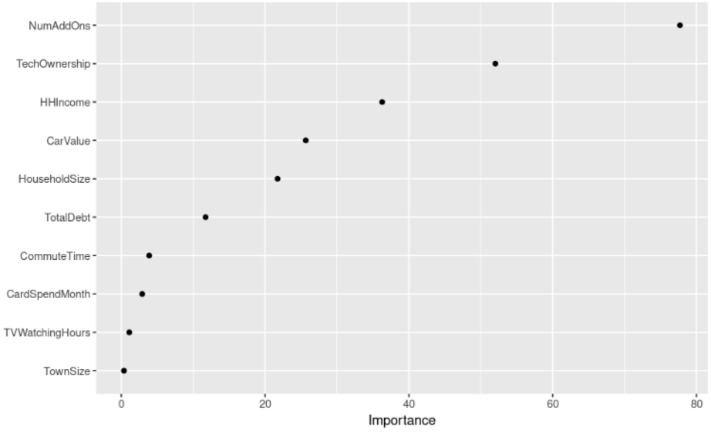


Percentage of High Retention Customers

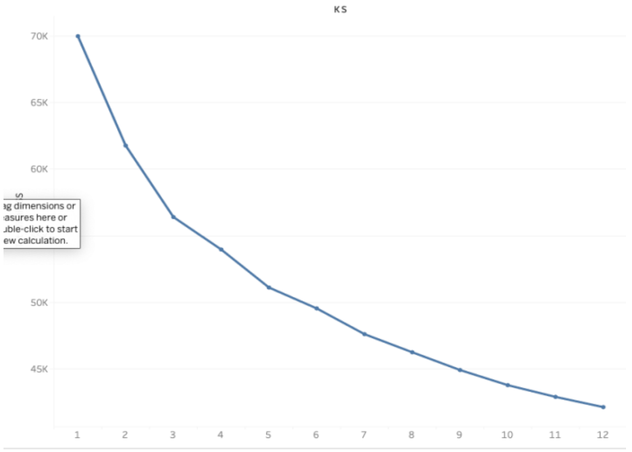


Appendix: Segmentation Method Plots

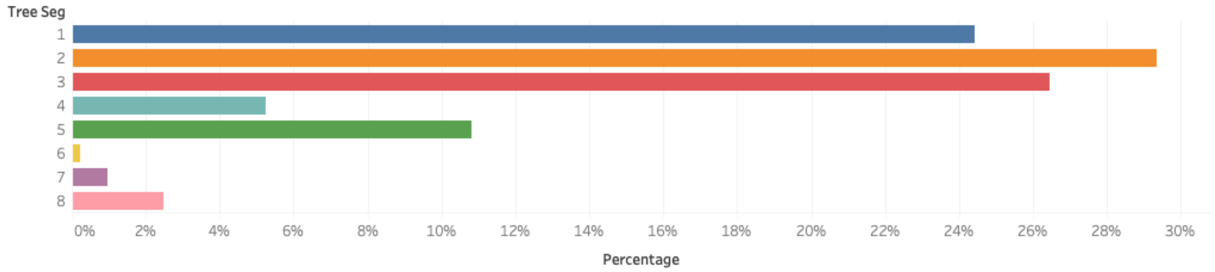




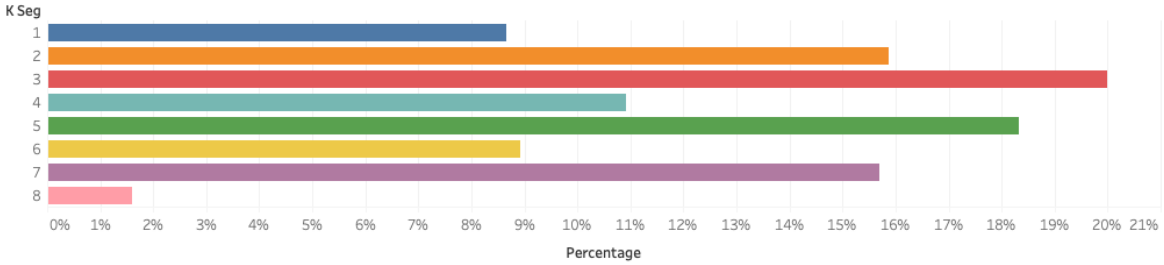
k-Means Elbow Plot



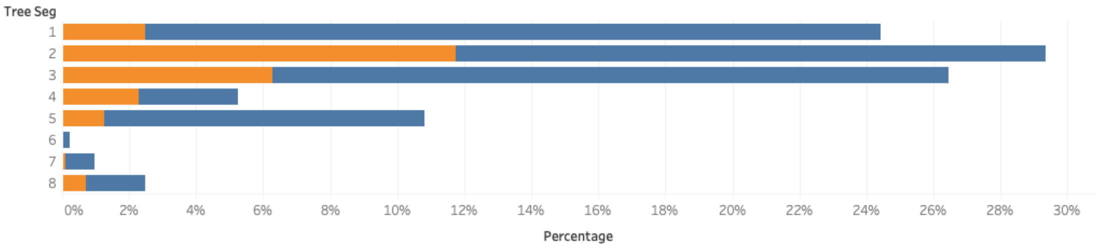
Percentage of Customers per Decision Tree Segment



Percentage of Customers per k Means Segment

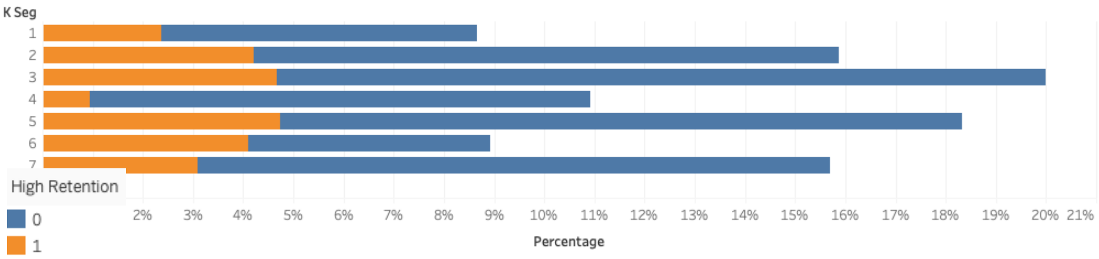


Percentage of Customers per Decision Tree Segment by Retention



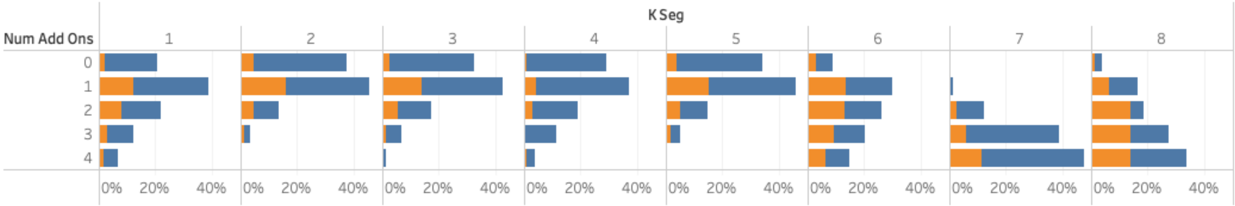
Appendix: Segment Distribution Plots for Segmentation Methods

Percentage of Customers per k Means Segment by Retention

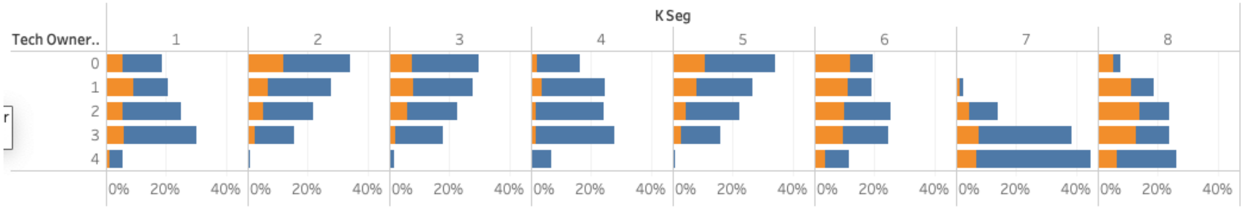


Appendix: Customer Segment Profile Visualizations

Number of Add Ons by Segment



Tech Ownership by Segment

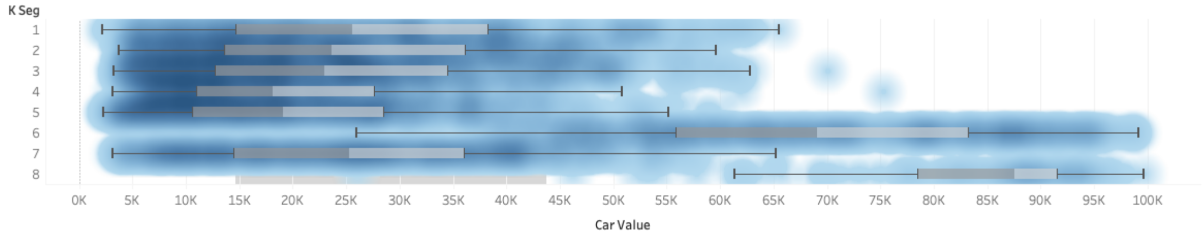


High Retention

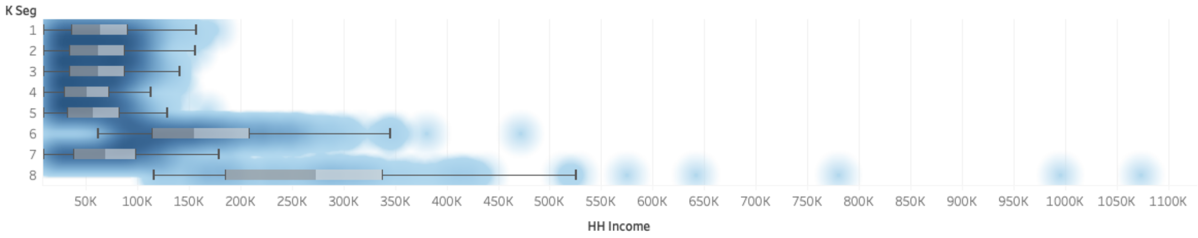
🟦 0

🟧 1

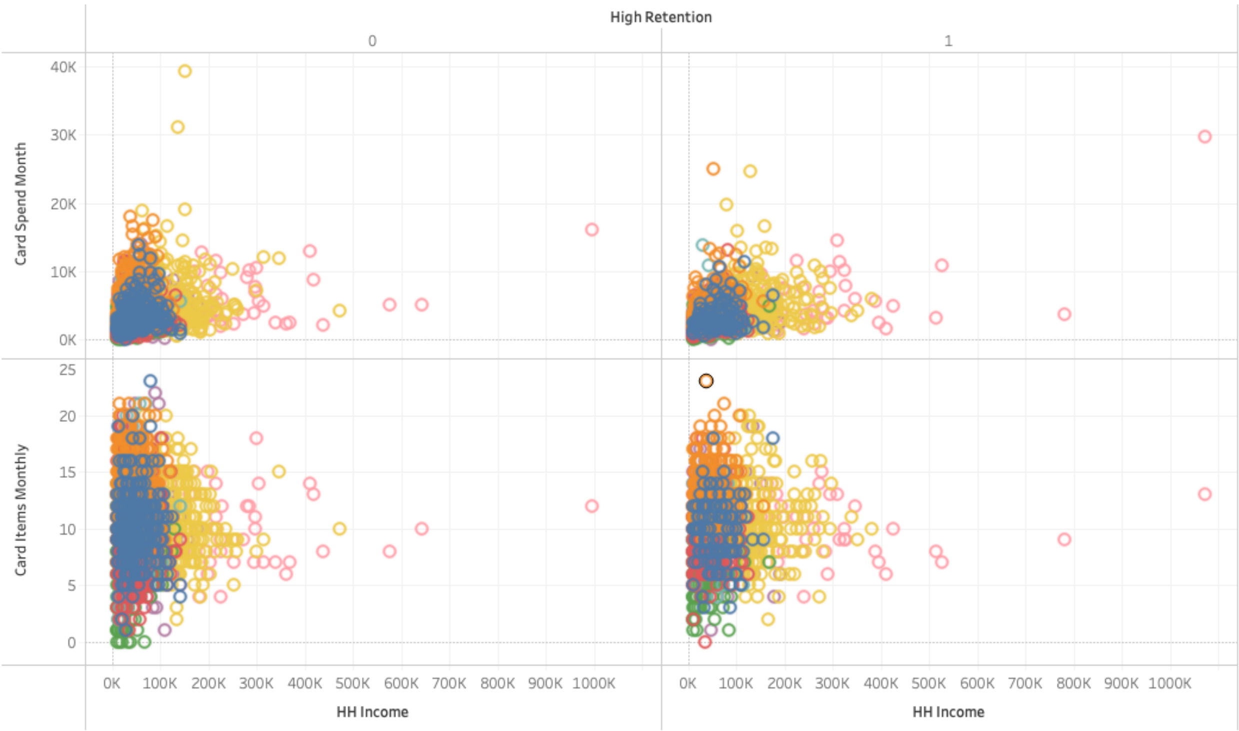
Household Car Value by Segment



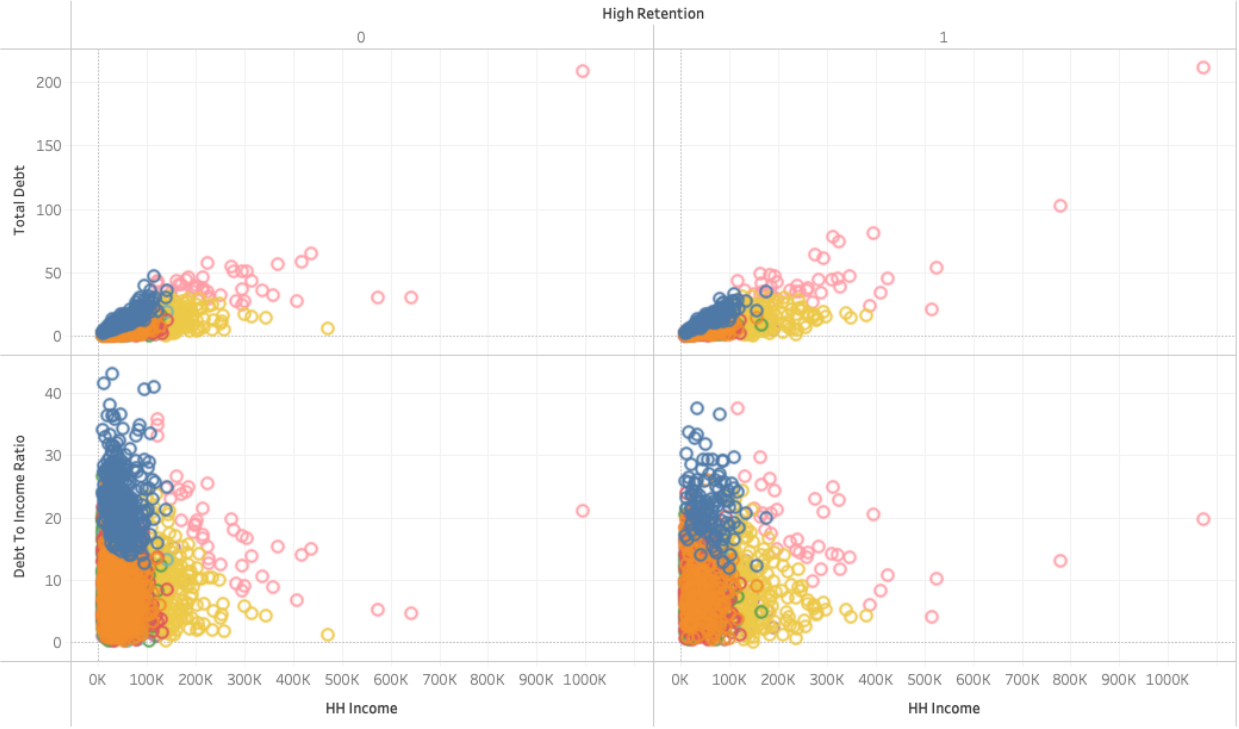
Household Income by Segment



Income vs. Card Spending and Monthly Items by Segment



Income vs. Total Debt and Debt-to-Income Ratio by Segment



Appendix: High and Low Retention Customer Profiles

Demographics

Segment	4 [Low Retention]	6 [High Retention]	8 [High Retention]
% High Retention	9%	46%	49%
% Female	44%	50%	59%
% Married	9%	52%	47%
Mode Job Category [%]	Sales [37%]	Professional [32%]	Labor [32%]
Median Years Education	14 years	16 years	16 years
% Retired	3%	3%	0%
% Union Member	16%	15%	13%
% News Subscriber	35%	65%	68%
% Votes	46%	63%	69%
% Political Party Member	36%	40%	43%
Median Household Income	5 persons	2 persons	2 persons

Ownership

Segment	4 [Low Retention]	6 [High Retention]	8 [High Retention]
% Homeowner	86	99	99
% Car Owner	86	99	99
Average Cars Owned	2	2.5	2
Median Car Value	$16,200	$68,500	$87,550

Behavior

Segment	4 [Low Retention]	6 [High Retention]	8 [High Retention]
% Active Lifestyle	58	39	33
Median TV Watching	20 hrs	21 hrs	21 hrs
Median Number Pets	1.5	2	2

Financial

Segment	4 [Low Retention]	6 [High Retention]	8 [High Retention]
Median Household Income	$30,000	$137,000	$249,000
Median Total Debt	$2,270	$1,150	$38,800
Median Credit Debt	$710	$319	$1,259

Segment	4 [Low Retention]	6 [High Retention]	8 [High Retention]
% Loan Default	34%	15%	60%
Median Card Spend	$2,450	$4520	$4,712
Avg Card Items Monthly	10	11	10

Technology

Segment	4 [Low Retention]	6 [High Retention]	8 [High Retention]
Median Tech Ownership	2 items	2 items	2.5 items
% Owns Game System	54%	50%	60%
% Owns PC	66%	69%	76%
% Owns Mobile Device	52%	47%	59%
% Owns Fax	12%	24%	48%

Add-ons

Segment	4 [Low Retention]	6 [High Retention]	8 [High Retention]
% Wireless Data	22%	37%	58%
% Voice Mail	26%	38%	56%
% Three Way	43%	63%	76%
% Pager	19%	33%	52%
% Electronic Billing	34%	37%	46%
% Call Waiting	42%	64%	81%
% Caller Id	44%	60%	78%

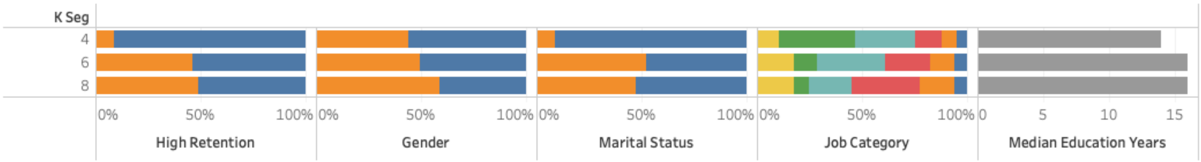
Account

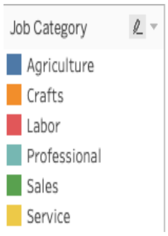
Segment	4 [Low Retention]	6 [High Retention]	8 [High Retention]
% Multline	36%	68%	86%
% Equipment Rental	34%	39%	45%
Avg Equip Rental Last Month	$7	$16	$30
Avg Data Last Month	$11	$16	$22
Avg Voice Last Month	$30	$57	$69

Geography

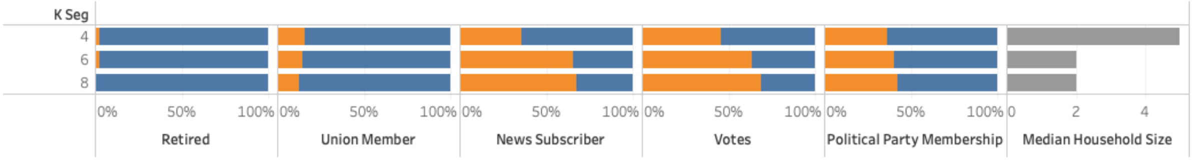
Region	Segment	1	2	3	4	5
Med Commute Time	4 [Low Retention]	25	26	26	25	26
	6 [High Retention]	24	24	24	23	27
	8 [High Retention]	23	26	25	29	25
Med Town Size	4 [Low Retention]	2	3	3	3	4
	6 [High Retention]	2	2	3	3	3
	8 [High Retention]	1	3	2	4	3

High and Low Retention Customer Profiles: Demographics 1

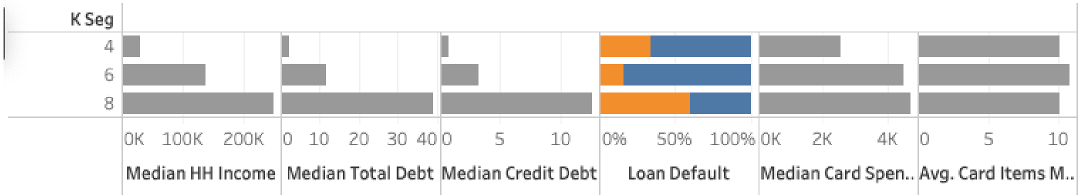




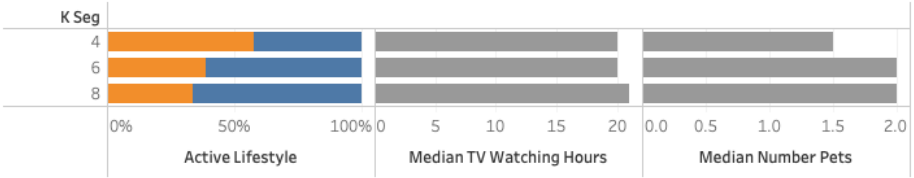
High and Low Retention Customer Profiles: Demographics 2



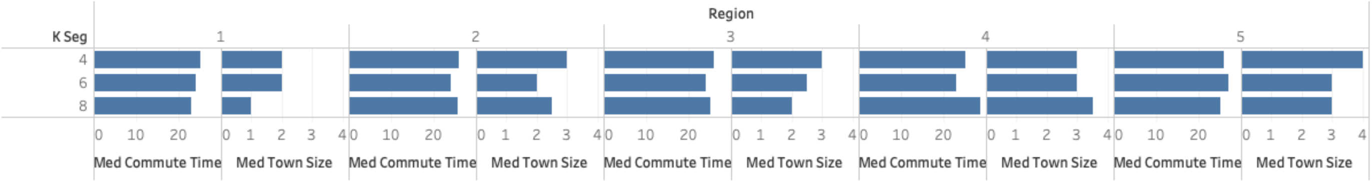
High and Low Retention Customer Profiles: Financial



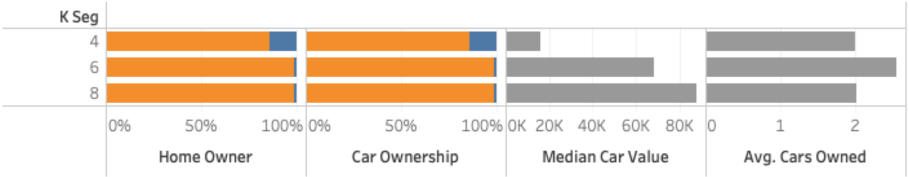
High and Low Retention Customer Profiles: Behavior



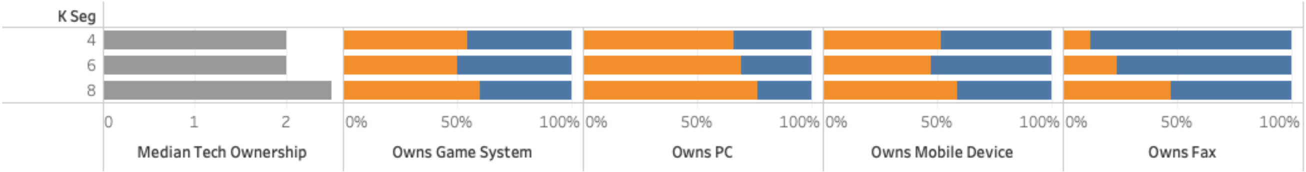
High and Low Retention Customer Profiles: Geography



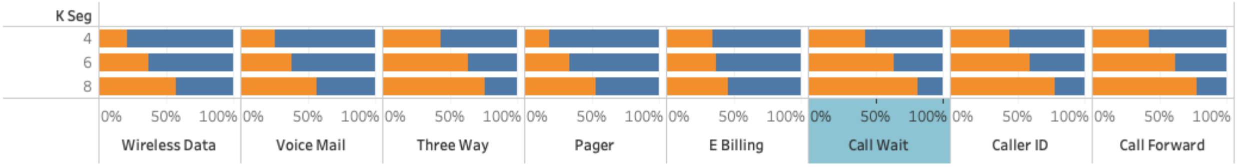
High and Low Retention Customer Profiles: Ownership



High and Low Retention Customer Profiles: Technology



High and Low Retention Customer Profiles: AddOns


K Seg
Wireless Data
Voice Mail
Three Way
Pager
E Billing
Call Wait
Caller ID
Call Forward