Telecom Customer Database: Segmentation & Profiling Project

Cullen Nauck

12/18/2019

Merrimack College

DSEG5113

Contents

[Table of Figures 3](#_Toc27693546)

[Executive Summary 4](#_Toc27693547)

[Overview 4](#_Toc27693548)

[Business Goals 4](#_Toc27693549)

[Data Description 4](#_Toc27693550)

[Methodological Summary 4](#_Toc27693551)

[Results 4](#_Toc27693552)

[Recommendations 4](#_Toc27693553)

[Core Report 5](#_Toc27693554)

[Overview 5](#_Toc27693555)

[Business Goals 5](#_Toc27693556)

[Data Description 6](#_Toc27693557)

[Data Quality 6](#_Toc27693558)

[Data Types 6](#_Toc27693559)

[Methodological Summary 7](#_Toc27693560)

[Segmentation Solution Logic 7](#_Toc27693561)

[Results 7](#_Toc27693562)

[Segmentation 7](#_Toc27693563)

[Customer Segment Profiles 7](#_Toc27693564)

[Business Segment Profiles 7](#_Toc27693565)

[Random Forest 7](#_Toc27693566)

[Appendix 7](#_Toc27693567)

# Table of Figures

**No table of figures entries found.**

# Executive Summary

## Overview

## Business Goals

## Data Description

## Methodological Summary

## Results

## Recommendations

# Core Report

## Overview

For this project, a 5,000 customer database was given with various consumer information and telecommunications spending information. The goal for this project was to develop a customer segmentation that can support effective and economically sound customer retention efforts. The business goals developed for this project was to determine what segments would be in the market for high-speed internet, unlimited mobile data and/or unlimited voice packages (or a bundle of all three). By segmenting the database into distinct clusters, we can more efficiently target customers based on their previous history.

To start with, some feature engineering was required to recode and reformat variables into data sets that could be used for the analysis. Some of this refactoring included cleaning the null values and reformatting the monetary values to be floating point numbers. Also, some variables were converted from total spending over the customers tenure into *per month* spending. This was done to get a better feel of how much each customer was spending on a month-to-month basis, so customers of varying tenures could be compared and analyzed against each other. For instance, for voice per month, the voice over tenure spending values were divided by the phone company tenure for each customer, giving a *per month* voice spending value.

Next, some of the continuous variables went through a process of binning into similar groups (primarily based on the quartile ranges). This was done to be able to view the cross-segment differences more clearly and effectively. Finally, some of the nominal categorical data was coded numerically for use in the k-means clustering method.

Before running the clustering and segmenting the data, the variables were split into various groups. The first group was the clustering variables, which included general overview variables, debt behavior variables and finally communications behavior variables. These variables were chosen for the basis of the segmentation analysis. The second group was the business profile variables, which included internet and TV behavior, mobile data behavior and voice behavior variables. These variables would be used to compare the segments to determine high and low value segments.

Once the data was cleaned and refactored, the k-means clustering approach was used to segment the data into 5 segments. This process included scaling the clustering variable data and then running the k-means clustering technique (using the scikit-learn package in Python). After performing this analysis, the segmentation solution was evaluated statistically and practically. The 5 segments were compared using crosstabs and pivot tables both on the clustering variable data and the business profile variables. This gave an in-depth overview of the segment profiles and cross-segment differences in order to arrive at the objective of determining which segments would be in the market for various packages.

Finally, using the segments determined for each customer, a classification model was built using a Random Forest. This was done so that this model can be taken to the remaining customers outside this small dataset to target specific customers for packages. The random forest was evaluated using training and test data from the 5,000 customer database. The accuracy of the model was evaluated by running the model on the test data.

## Business Goals

The goal for this project is to determine segments that can be targeted for packages including high-speed internet, unlimited mobile data and/or unlimited voice packages (or a bundle of all three). To do this, spending and various telecom behavior was gathered for each customer to determine their previous history. By determining which segments are more inclined to internet, mobile phones and/or voice, we can better know which upgrades and packages will have a higher probability of retaining customers. Since we have a broad customer range over various markets, it is important that we recognize the needs of each segment separately rather than assume all customers want the same thing.

## Data Description

The 5,000 customer database was provided in a .csv format with various data types. Variables included general information such as age, gender and region as well as spending and debt behavior. The 59 variables provided were truncated and reformatted into two smaller groups: demographic clustering variables and business profiling variables. In order to ensure the data was in the correct format for analysis, some feature engineering was done to clean data, bin data and code data into numerical values.

### Data Quality

Since the raw data was in various formats, some work was required to correct the data for analysis. The null (missing) values in analysis variables were replaced in different ways according to their variable type. The TownSize variable had two missing values, so it was assumed sufficient to simply replace these with the mode of TownSize. The missing Gender and JobCategory values were replaced with “Unknown”. Finally, the Card Spend, Voice and Data over tenure missing values were replaced with 0 because it was determined that a missing value simply meant the customer had not spent money in that category. Also, these monetary values were feature engineered to remove the ‘$’ and ‘,’ to format them into floating point numbers. The Internet variable contained 3 numeric values as well as ‘yes’ and ‘no’. It was assumed that these three values corresponded to ‘yes’ on Internet, and they were recoded in this way.

To more thoroughly capture spending trends over the customer database, the voice and data over tenure spending values were converted to a *per month* value by dividing by the total tenure with the phone company. This was done to get a better feel of how much each customer was spending on a month-to-month basis, so customers of varying tenures could be compared and analyzed against each other. For instance, for voice per month, the voice over tenure spending values were divided by the phone company tenure for each customer, giving a *per month* voice spending value.

Next, some of the continuous variable data went through a process of binning to gather the data into similar groups. To do this, statistical analysis was done on the continuous variable to determine quartile locations. These were then used to determine bin group regions across the range of the variable. The variables binned using this process were Age, DataPerMonth, HHIncome, EducationYears, CardSpendMonth, DebtToIncomeRatio, PhoneCoTenure, TVWatchingHours, and VoicePerMonth. These were binned to more clearly distinguish between the various segments.

Finally, for use in the k-means clustering approach, some of the categorical nominal variables had to be numerically coded. These included Gender, JobCategory and LoanDefault. This was done by simply assigning a particular category to a specific integer for each of these variables.

### Data Types

After determining the relevant variables for this analysis, these remaining variables were split into two groups: demographic clustering variables and business profiling variables. The first group was the clustering variables, which included general overview variables, debt behavior variables and finally one communications behavior variable. These variables were chosen for the basis of the segmentation analysis. The second group was the business profile variables, which included internet and TV behavior, mobile data behavior and voice behavior variables. These variables would be used to compare the segments to determine high and low value segments.

Within the demographic variables used for clustering were the general overview variables. These provided general demographic information for each customer in the database. They included age, town size, region, gender, job category, household income and education years. These variables were selected because it was expected that these could thoroughly capture the various demographic differences between segments of the customer set. Next, the debt behavior variables included debt to income ratio, card expenditures per month and loan default. These were included to capture some of the financial factors that can play into telecommunication usage. Finally, phone company tenure was included as a clustering variable so that we can get clarity into the differences between customers of various tenures with our company.

The business profiling variables were split into three groups: internet/streaming variables, mobile data variables and voice usage variables. The internet and streaming variables included PC, Game System and Fax ownership, as well as news subscriptions, TV watching information and finally internet access. Next, the mobile data variables included wireless data access, data spending per month and mobile phone ownership. Finally, the voice usage variables included voice spending per month, calling card, three way calling, multiline and voice mail usage.

## Methodological Summary

After the feature engineering and variable recoding was completed, the selected clustering variables could be used to segment the data set into groups. To do this segmentation a k-means clustering approach was used. The KMeans algorithm from the scikit-learn package in Python was used to perform this analysis. After the segmentation was completed, a classification model for the data set was built using a Random Forest.

### Segmentation Solution Logic

Although it has various drawbacks since it is a Euclidian distance spaced model (which required the recoding of categorical values), k-means was determined to be sufficient for this level of analysis. K-means separates samples into a selected number of groups of equal variance by minimizing inertia. To perform this analysis, first the variables selected for clustering were scaled by applying a standard scaler to them. This was done to normalized the variables so that differences in orders of magnitude between the variables would skew the results. Next, an inertia plot was created to determine how the inertia values were affected by the number of clusters:

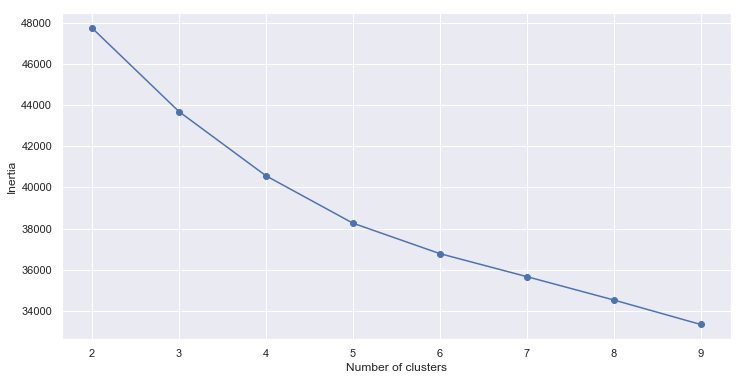


Figure : K-means Inertia Comparison

Although in this case the Inertia plot does not show a leveling off after a certain number of clusters, it was determined for this level of analysis that 5 clusters would be sufficient. The k-means function was run for 5 clusters based on the clustering variables previously discussed. After this was completed, the clusters were reviewed to evaluate the statistical and practical efficacy of the solution. The inertia for this 5 segment k-means clustering was determined to be 38260. The segments ranged in size from 1291 customers to 383 customers. Some of the cluster centers were compared against each other for the k-means variables to view the scatter of each cluster across the variable space.

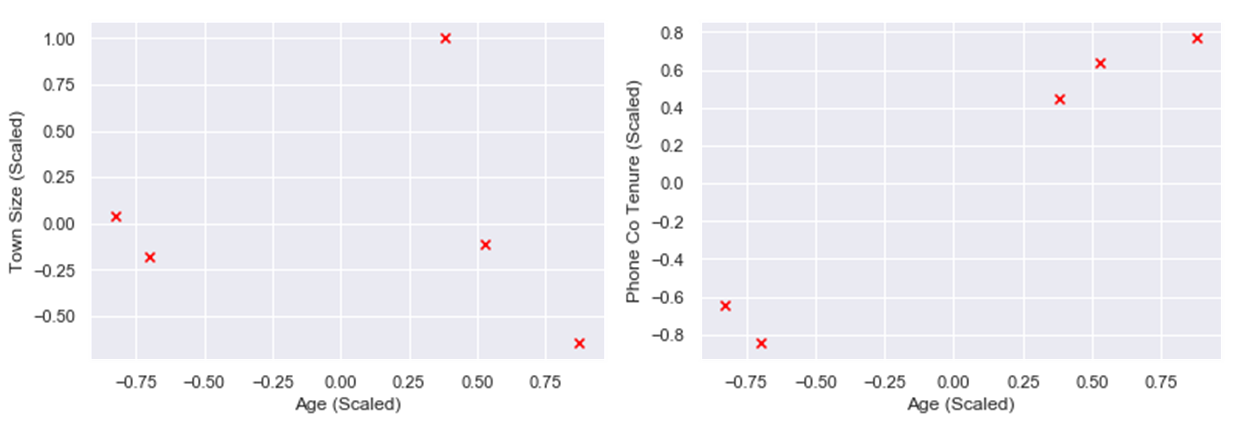


Figure : K-means Segment Scatter Plots

Though these are just two of many scatter plots that can be used to visualize the clusters, these show how each of the clusters fit into various segments of the market. Though some of the segments are relatively close to each other, this shows that the k-means clustering is

## Results

### Segmentation

### Customer Segment Profiles

### Business Segment Profiles

### Random Forest

## Appendix