SMU®

Data Science Machine Learning DS7331.402

Amber Burnett
Lance Dacy
Shawn Jung
Jeremy Otsap

Lab 3

April 12, 2020

Table of Contents

Business Understanding

The purpose of the data is to provide recommendations how to retain existing customers from a North American Telco provider. In telecommunications, the estimated cost of new customer acquisition is approximately 5x higher than retaining an existing customer.

Furthermore, only a third of customers switch carriers due to lower prices; more and more factors such as dissatisfaction with quality of service, advancing technology and media features, competitors having better cellular coverage, and poorly implemented loyalty programs are all contributing to customer attrition.

Using data science techniques, the goal was to determine an algorithm that could predict the customers who are most likely to leave our provider (churn). Naturally, the most efficient algorithm is desired, but likely requires trial and error.

Metrics are capable of measuring the outcome of the algorithm, allowing the data team to select the appropriate model based on the metrics desired. Most of the metrics focus on time and space. If an algorithm can solve the problem at hand efficiently (or the least amount of time and effort consistently as deployed), then it is the one likely selected.

If an algorithm can solve the problem utilizing less resources (memory or space), then it is likely a candidate as well. The data team has to weigh the final solution against these metrics.

In order to understand relationships within the large data set, the data team gravitated towards two types of clustering methods; Agglomerative and Hierarchical. The business understanding of this process can be succinctly described:

- Based on dissimilarities between the data objects, groupings of objects are identified (clustered). A type of dissimilarity can then be deduced for the target objects.
- The results of this method can be shown by a tree diagram, focusing on taxonomic relationships (a dendrogram). This diagram demonstrates the progressive grouping of the data as it iterates over the clustering algorithm. The process then evolves to categorize the data into a suitable number of classes.
- The dendrogram then represents a hierarchy of partitions. A partition is then selected by truncating the tree at an appropriate level; that level depends on either user-defined constraints or more objective criteria presented by the algorithm.
- The trees and partitions can then be evaluated to gain insight into how the data are best related or classified for future analysis or perhaps create new variables to analyze based on the clustering exercise.
- The final result will help the team provide proper customer segmentation based on statistically significant data relationships.

Data Understanding

Describe the meaning and type of data

A more detailed analysis of the original data set can be found in previous documents provided. Naturally as the data team performs analysis, the original dataset evolves.

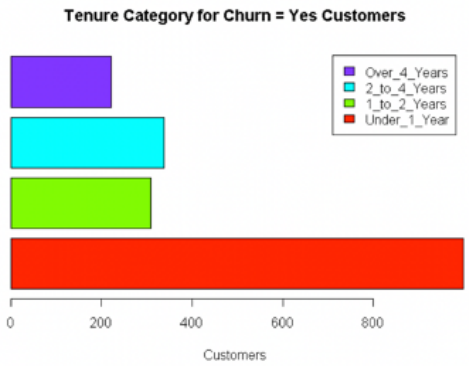
Still at play is the fact that there are 11 missing values in the TotalCharges column. The team deduced that these missing values correspond to 0 in the tenure field and thus assigned that assumption to the values. The conclusion is that these are essentially new customers that have just begun their contract for the first time. Using values from the MonthlyCharges, the team is able to impute these values for TotalCharges that were missing. The 11 missing values are innocuous to the overall process, but imputed nonetheless. There are no duplicated data or outliers in this dataset.

The goal of this particular modeling exercise was to examine what services would benefit or entice the customers (preventing their churn). Preventing data leakage was a great concern given the manipulation of the data thus far as well as what the customer's total spend would be in the future. The monthly spend is known so that can be deduced to help determine if the customers are cost conscientious.

The important variable of Tenure will remain, providing how long the customer has been in their contract. The Contract variable itself may be a secondary business outcome tied to tenure. The team decided not to include it with Tenure, classifier; hoping to prevent skewing the data. The Payment Method and the Paperless Billing variables do not have much business value and thus were excluded in this analysis.

Visualize the any important attributes appropriately.

Of the customers who are in the "Yes" Churn category, more customers are within their first year of service, than the other 3 categories combined. This proved to be a statistically significant data point for classification and clustering.



Modeling and Evaluation

Naturally, modeling and evaluating data is an iterative and responsive process that changes radically with small variations in the algorithm. For a more detailed view of the progression of the analysis, an HTML Workflow summary is provided. The modeling and evaluation process that was chosen is the Cluster Analysis (Option A).

The team has performed a cluster analysis using serval methods outlined in the repository provided. In addition, decisions were made for a suitable number of clusters for each method based on the algorithm results. Internal and external validation methods were used to describe and compare the clusters along with some detailed visualizations throughout the repository.

The information was prepared and a Gower Matrix constructed to handle the data to be appropriate for clustering analysis. Gower distance (a distance measure that can be used to calculate distance between two entity whose attribute has a mixed of categorical and numerical values) is calculated for each variable type and scaled between 0 and 1. A linear combination is then calculated to create the final matrix.

We will focus on hierarchical clustering as it lends itself to visual analysis, and thus more intuitive understanding. Agglomerative clustering starts with as many clusters as there are observations, with each of those observations being its own individual cluster. Then as the model progresses, more similar data points are discovered and grouped into clusters.

The goal of Agglomerative Hierarchical Cluster is to determine the longest distance between two points in each cluster (i.e. hierarchical cluster distance). There are 4 methods to determining this distance:

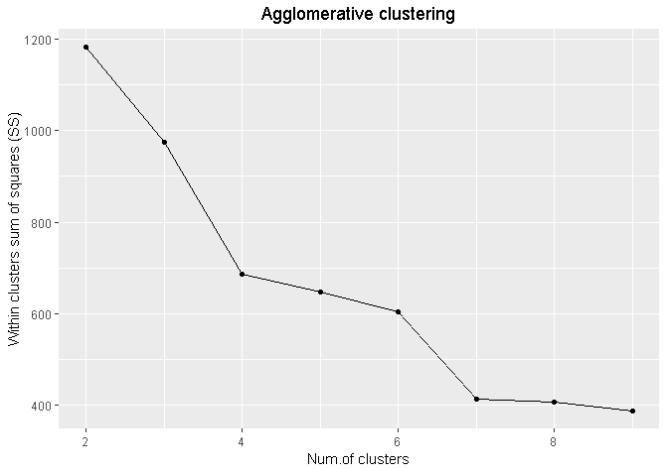
- **Complete**: pairwise similarity between all observations in clusters 1 and 2. This approach uses the largest distance as distance between the clusters. This is typically the default method in the analysis.
- **Single**: Similar to complete but uses the smallest distance between the clusters.
- **Average**: Also similar to complete, but uses the average distance between the clusters.
- **Centroid**: Calculates the centroid of each cluster and uses “centroid distance” which can be thought of as the multi-dimensional average of the distance between the clusters.

Given that the complete and average method tend to produce more "balanced" trees, they were the chosen method of this exercise.

The first step in this process is to select the optimal number of clusters. There are multiple methods to use, for this exercise, the team chose the following:

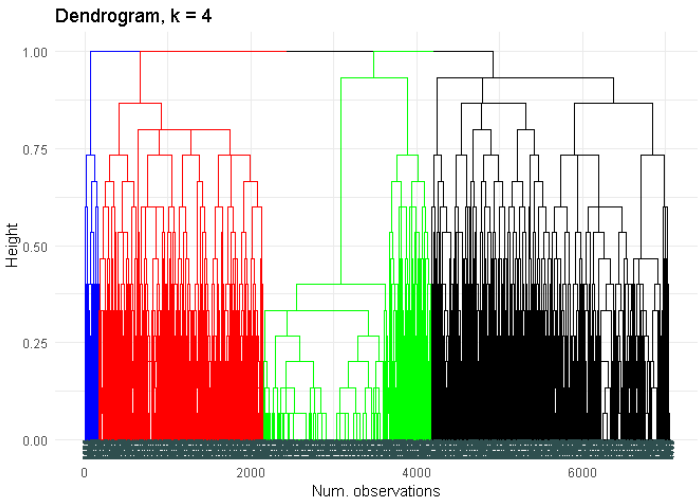
- **Elbow**: when compactness of clusters, or similarities within groups are most important for your analysis.
- **Silhouette**: displays a measure of how close each point in one cluster is to points in the neighboring clusters; basically measuring data consistency.

Elbow Plot for Agglomerative Clustering



Given the sharp change in the sum of squares, 4 clusters appears to be the best for this analysis.

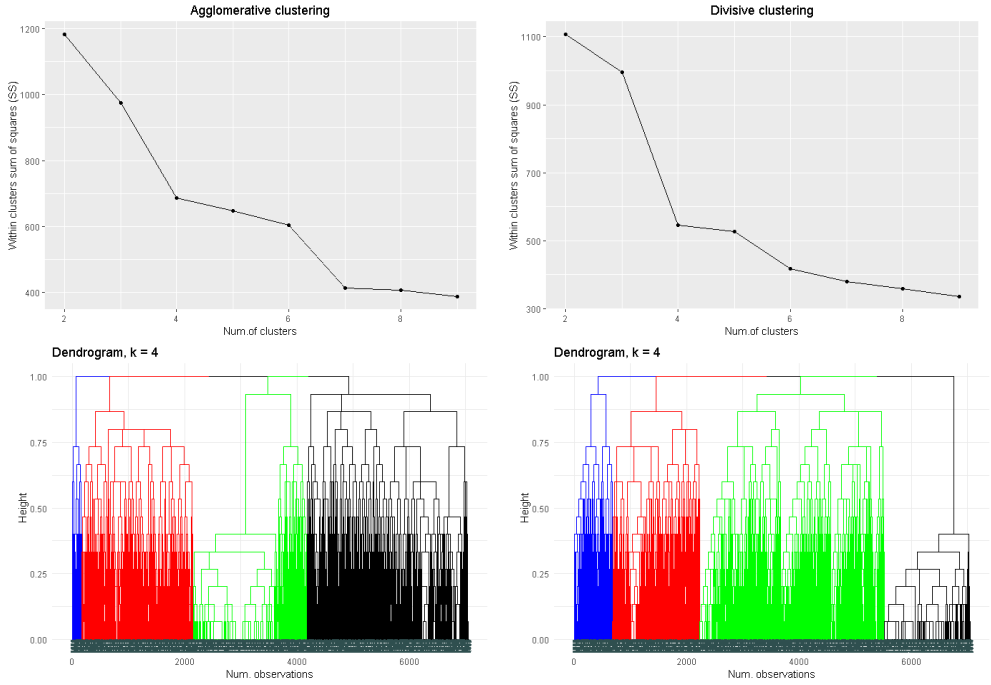
Dendogram for 4 Clusters



A key component of this analysis will be comparing the agglomerative (bottom-up) and divisive (top-down) methods of Hierarchical Clustering and noting the differences to arrive at a final decision for the model.

Divisive clustering assumes all data points are part of a single cluster, and then subdivides them by most dissimilar ones into separate clusters. In general, divisive is better for a smaller number of large clusters, while agglomerative is better for finding a large number of smaller clusters. After analyzing both clustering methods, the deduction can be made that 4 clusters are the best decision for this dataset.

Agglomerative vs. Divisive Clustering



Based on the detailed analysis provided in the repository, the conclusion can be made that:

- Customers who have medium or high monthly charges are people with phone services, internet services and streaming services
- Customers who have low monthly charges are people without phone, multiple line, internet, online security, device, technical support or streaming services.

Armed with this knowledge, the raw clustering data can be visualized; assisting the marketing team on target customers to begin the campaign for retention.

The distinction of Internet Only users, Phone Only users and Phone+Internet Bundle user group is apparent. This is also supported by extensive clustering analysis from Machine Learning algorithms. The existing customer segmentation is still valid and is validated by these techniques.

AGGLOMERATIVE HCLUST	Cluster 1	Cluster 2	Cluster 3	Cluster 4		ROW TOTALS
tenureCAT Under_1_Year	679	1107	276	7	7043	2069
tenureCAT 1_to_2_Years	334	456	247	10		1047
tenureCAT 2_to_4_Years	464	613	502	45		1624
tenureCAT Over_4_Years	556	684	952	111		2303
monthlyCat LOW	2033	391	0	27	7043	2451
monthlyCat MED	0	2269	24	146		2439
monthlyCat HIGH	0	200	1953	0		2153
Gender Male	1042	1386	1039	88	7043	3555
Gender Female	991	1474	938	85		3488
SeniorCitizen NO	1902	2419	1431	149	7043	5901
SeniorCitizen YES	131	441	546	24		1142
Partner NO	1106	1619	866	50	7043	3641
Partner YES	927	1241	1111	123		3402
Dependents NO	1261	2079	1502	91	7043	4933
Dependents YES	772	781	475	82		2110
PhoneService NO	507	8	0	167	7043	682
PhoneService YES	1526	2852	1977	6		6361
MultipleLines NO	1184	1679	523	4	7043	3390
MultipleLines NO PHONE SVC	507	8	0	167		682
MultipleLines YES	342	1173	1454	2		2971
InternetService NO	1526	0	0	0	7043	1526
InternetService DSL	507	1705	36	173		2421
InternetService FIBER OPTIC	0	1155	1941	0		3096
OnlineSecurity NO	335	1836	1261	66	7043	3498
OnlineSecurity NO ISP	1526	0	0	0		1526
OnlineSecurity YES	172	1024	716	107		2019
OnlineBackup NO	329	1827	871	61	7043	3088
OnlineBackup NO ISP	1526	0	0	0		1526
OnlineBackup YES	178	1033	1106	112		2429
DeviceProtection NO	359	1926	793	17	7043	3095
DeviceProtection NO ISP	1526	0	0	0		1526
DeviceProtection YES	148	934	1184	156		2422
TechSupport NO	329	1827	871	61	7043	3088
TechSupport NO ISP	1526	0	0	0		1526
TechSupport YES	178	1033	1106	112		2429
StreamingTV NO	370	2079	336	25	7043	2810
StreamingTV NO ISP	1526	0	0	0		1526
StreamingTV YES	137	781	1641	148		2707
StreamingMovies NO	368	2015	387	15	7043	2785
StreamingMovies NO ISP	1526	0	0	0		1526
StreamingMovies YES	139	845	1590	158		2732

Deployment

In conclusion, the data team can continue spending an inordinate amount of time analyzing the data provided and continue to fine-tune the segmentation for the desired retentive marketing campaign. The law of diminishing returns will naturally set in at some point. The goal is to spend as little time as possible, being as accurate as we can as well as making the model as simple as possible to support and maintain.

There are risks to all models and given the current information provided, the team feels confident in recommending the following customer segmentation for the target retention campaign. The ultimate goal would be to convert at risk customers to long-tenured customers.

The recommended segmentation along with their average monthly charges are:

- Fiber Phone Users ($91.20)
- DSL-Phone Users ($61.80)
- Simple Internet Users ($42.20)
- Phone Only Users ($21.10)

The business outcome of this exercise would suggest that if the marketing team could find a counter measure to at risk customers (such as “first 3 months free”), they could prioritize the lack of churn for those clients.

To support the deployment of this model would require simply the current infrastructure of the marketing team’s original dataset and scalable cloud infrastructure to handle the extreme processing power of the random forest algorithms used.

Given the current infrastructure exists for this exercise, the team would need to ensure the AWS support team allocated the required scalable architecture as the models are re-assessed. The model could be updated each month and perform reasonably within the confines of this analysis. As the data is evaluated in a live environment, the team would fine-tune the operations to ensure the persistence of the recommended customer segmentation. The target would be customers within their first year of contract, so a monthly run might not be required.

While automatic billing is related to long-term use, the marketing team could target these clients with a $10 off coupon when transitioning to auto-billing. Using this data and the model assist in narrowing down the target retention plans.

Instead of using broad-swathing programs, the marketing team could narrow their focus on the clients identified in this analysis to focus their efforts and impact the bottom line of the business earnings by saving costs of acquiring and onboarding new customers.

Appendix

In order to provide visualizations into the data as well as summary and advanced statistical modeling, many tools are used. These tools naturally required programming skills and data manipulation skills. In order to make this work as reproducible as possible, the following links to coding notebooks are provided.

Code Link: GitHub Complete Repository