MAT-8452
Dr. Frey
Hai Du
July 12, 2020

MAT 8452 - Categorical Data Project
Telco Customer Churn Analysis

1. Introduction

Customer churn is a major problem and one of the biggest concerns for telecommunication industry. Due to the cost of retaining an existing customer is much lower than acquiring a new one, the company is seeking to develop a churn prediction model to assist telecom operators in predicting which customers are most likely to lose. Therefore, identifying the factors that increase customer churn is important to build this model. The purpose of this project is to explore these data more deeply, utilizing nonparametric statistical methods to do so. Through the course of this analysis, new insights will be offered as to the types of indicators that influence churn and charge, as well as attempting to compare both nonparametric and parametric methods.

2. Data exploration

The raw data contains a total of 7043 observations and 21 variables. However, not all of the predictors are useful for our models, specifically customerID, and thus we remove it first. We also have some missing values in this dataset, they are a total of 11 missing values from TotalCharges. By looking at the dataset, I realized that all of missing value are churn=" No", which is over-represented in the data. We will just delete those rows from the working data.

3. Statistical Tests and Analytical Methodologies

The first of the nonparametric methods I implemented with this dataset dealt with the Monthlycharge variable. I wanted to run a test to see if there any difference in monthly charge between customer retention and customer churn. To do this, I first checked their distributions with density and box plot, which do not follow the normal distribution and have different scale and location. Using Kolmogorov-Smirnov test, I was able to test if Monthlycharge comes from a specified distribution. This test assumes that the data are continuous and come from a random sample. The K-S test gives us the test statistic of 0.25 with p-value < 0.0001, which is less than the significant level of 0.05. Hence, we reject the null hypothesis and conclude that customer retention and customer churn do not come from the same distribution. I also ran the Sign Test to see if my data matched up with national averages. To do this, I searched online and found that the 50th percentile monthly charge in the U.S. is 65.5 dollars. As previously mentioned, I am treating these values as a random sample. The hypotheses were:

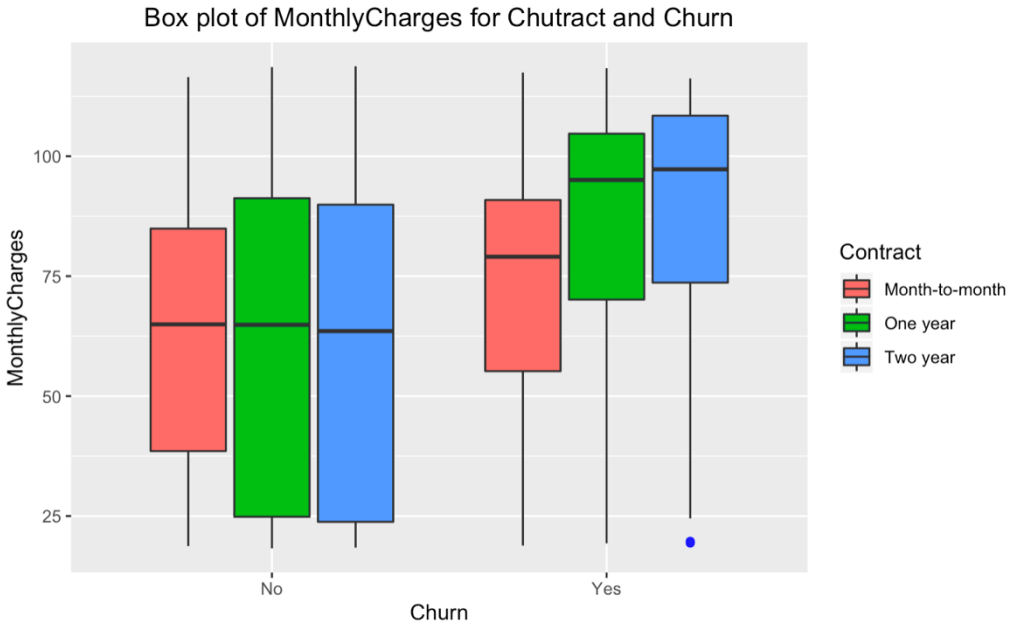
customer retention: $\begin{pmatrix} H_0\colon \theta_{0.5} < 65.5\ dollar \\ H_A\colon \theta_{0.5} > 65.5\ dollar \end{pmatrix}$ customer churn: $\begin{pmatrix} H_0\colon \theta_{0.5} > 65.5\ dollar \\ H_A\colon \theta_{0.5} < 65.5\ dollar \end{pmatrix}$

For both customer churn and retention, the Sign Test showed a significant result, meaning that the 50th percentile of lost customers will pay more than 65.5 dollars per month, and the 50th percentile of loyal customers will pay less than 65.5 dollars per month, at the 5% level. I also ran the parametric one-sample t-test and saw a similarly strong evidence respect to each test when comparing MonthlyCharge values with my sample means.

What is even worse than having churn is not knowing why your customers churned. The monthly charge of losing customers is relatively high. What is the cause? I was interested in knowing if monthly charge differs among co-renting partner. Since we have two independent treatments with equal variance and shape, I wanted to use a Rank Sum test to see if there is difference between having partner and without partner. The Wilcoxon Rank Sum Test gave us

the W test statistic as 6954306 with p-value < 0.0001. We reject H0 and conclude Ha. That is, we have enough evidence at 0.05 level to conclude that having a partner will increase monthly charge. Upon investigating the 95% confidence interval for the difference, I am 95% confident that having partner tend to be between four and six dollars higher than non-partner. When comparing this result to the parametric two-sample t-test, the similar result is achieved. To fully rely on the t-test, however, I would have to make sure each monthly charge is in fact normally distributed, which may or may not be met due to the left skewed.

Since my data analysis was focused on the reason behind the costumer churn, I also wanted to compare the monthly charge variable between the churn and contract. First, as we would be expected, the churn rate of month-to-month contract customers is much higher than the longer contract customers. Customers who are more willing to commit to longer contracts are less likely to leave. From a personal point of view, I think the price concessions are the main reason for me to sign a long-term contract, Therefore, I am interested in exploring the relationship between monthly contract and annual contract in term of scale(variability) and location(median), as well as what price point can customers be better retained. To do this, I would like to use Ansari-Bratest, Kruskal–Wallis, Rank Sum and Jonckheere-Terpstra test to analyze this part of data. Before running any tests, I



assume that the any k samples of interest are independent and of equal shape for both customer retention and customer churn. The contracts are inherently independent.

First, I used the Ansari-Bradley and normal theory test to check if monthly contract and an annual contract are identically distributed. I assume that the median of monthly charge for different contracts are the same in customers retention, however they are not normal distribution due to the left skewed. Both parametric and nonparametric test gave the p-value < 0.0001. We reject H0 and conclude that one year has greater variability associated with it than does the monthly contract. Second, I used Rank-Sum Test to see if there any difference in contract between the One-Year and Two-Year. This gave p-value 0.2695>0.05, we retain H0 and do not have enough evidence to conclude at 0.05 level that the two populations are not equal. When comparing this result to the parametric two-sample t-test, the p-value is relatively small, we reject H0 and conclude that the two populations are not equal, but likely are less reliable than the A-B results due to the abnormal tails. Third, I wanted to compare monthly charge for all different contract in customer retention. Using Kruskal-Wallis and ANOVA F test, both tests gave p values > 0.05, we fail to reject H0 and cannot conclude Ha. That is, we do not have enough evidence at 0.05 level to conclude that contracts are different. The box plot above shows that the monthly contract is roughly similar to other contracts. In order to verify the results of K_W test, I developed the nonparametric approach to find confident interval for these shifts, which are CI_{12}(-1.9, 0.15), CI_{13}(-0.4, 0.9) and CI_{23}(-0.25, 1.05) (1=month-to-month, 2=One year, 3= Two year). Since the confidence interval is so close to 0, we can make a conclusion that we have weak evidence indicating that the shifts may be equal to 0. Finally, for customer churn, I was interested in knowing whether monthly charge increases as the contract term increases. Using

Jonckheere-terpstra test, I found p-value < 0.0001, so we have enough evidence at 0.5 level to conclude that monthly charge increases as the contract term increases.

With the three numeric variables I used throughout this analysis, I also wanted to see if total charges had any sort of association with each tenure. To accomplish this, I made scatterplot, and found that there is a fairly strong, positive, linear association between total charges and tenure. To verify this result, I ran Spearman's and Pearson correlation test, both tests concluded that these two variables are linearly related. Again, Pearson correlation test may less likely reliable than Spearman's test due to the heteroscedasticity. Since there were so many categorical variables in this dataset, I naturally wanted to test for association between these variables. I decided to determine if there exists a correlation between the churn and the services customers chooses. This sort of analysis lends itself well to Chi-Square Tests for Independence. For comparison, I examined both parametric and nonparametric Chi-Square Tests. Under the conventional significance level of 0.05, both tests reject the null hypothesis and conclude that there exists an association between churn and services customers choose except PhoneService. By looking at the data, we know that more than 90% of customers have choose PhoneService, so there would not make any significant point.

Finally, I was interested in finding the scale effect of how variables can affect out total charges. I used the variables tenure, churn, and PhoneService to address this problem. Since multiple linear regression requires normal distribution for the data, we cannot simply apply that to our analysis as the variables are all skewed and not normally distributed. Therefore, I applied the bootstrap method for this multiple linear regression and found that 95% bootstrap confidence interval for the slope of tenure, churn, and PhoneService are (78.1, 81.1), (493.3, 573.9), (755.8, 869.3).

3. Conclusion

In summary, there appeared to be several findings that could provide useful insight into the customer churn analysis. First of all, the churn rate of our customers has reached 1/4, of which 95% have not signed long-term contracts. We find that customer churn tends to have higher monthly charge, while loyal customers are relatively low. Moreover, for the loyal customers, the price of their contract does not change much, which is around 63. This shows that setting the price at 63 is more conducive to retaining customers. There are several reasons for the customer churn. One of the reasons is the tenure of customers in the company. Customers who have been with the company longer or have paid more in total are less likely to churn, and more likely they are to become loyal customers. Finally, the demand of customers for different services is positively correlated with the total charge.