Customer Churn
Tomasz Wojcik, Piotr Baran, Janny Tam

Introduction

Churn plays a vital role in almost every field. It is a measurement the business uses to see how many of its customers they lose in a given time frame. It can provide insight into various other metrics such as quality of service, pricing levels, satisfaction, and how a business compares to its competitors.

Churn is significant to the telecommunications industry, as it has one of the higher average churn rates. In 2020, the telecom industry's average churn rate was 21% in the United States (Statista, 2021). For telecom companies, roughly 20% of revenue is spent on capturing new customers and retaining current ones. To provide context on how churn affects a telecom business, In 2017, Telus released information on their retention costs. They advised that the average cost of keeping an existing customer was $11, while acquiring a new customer costs on average $521 ("Churn is Breaking the Telecoms Market: Here's How to Fix It," 2018). With the difference being this vast, companies must minimize their churn rates.

Our Data

Our dataset is an IBM-created data set located on Kaggle, which was created to predict customer behavior. There are 17 categorical variables and three numerical variables. The data consists of over 7000 customers, including churn, products that each customer utilizes, accounting information such as contract type and payment method, and demographics such as gender, if the customer is over the age of 65, and if they have a partner or dependants.

The Churn variable will be our target variable, where it is if the customer had canceled their Telco services in the last month. Using clustering, various classification models, and market basket analysis, we will see if we can predict customer patterns for those who are more likely to cancel their services.

Exploratory Data Analysis

By viewing our Power BI dashboard outputs included in our Appendix, we look at what the average customer looks like at Telco, along with what the average churned customer. Some highlights of our Exploratory Data include:

Full Customer Data Set	Churn Only Data Set
Demographics ● **32.37** Month Tenure ● **7043** Customers ● Younger Customers	**Demographics** ● **17.98** Month Tenure ● **1869** Customers ● Less likely to have a Partner
Products ● **90%** Phone Service ● **78%** Internet Service ● Secondary Service Range from **37% to 56%**	**Products** ● **90%** Phone Service ● **94%** Internet Service ● Secondary Service Range from **17% to 31%**
Accounting ● **$64.76** Average Monthly Charge ● **$2283.30** Average Lifetime Value ● **55%** Month to Month Contract	**Accounting** ● **$74.44** Average Monthly Charge ● **$1531.80** Average Lifetime Value ● **89%** Month to Month Contract

Clustering

As there are a finite amount of different services that Telco offers, there will be different types of customers with different needs when it comes to their phone, internet, and streaming needs. With clustering, we will try to extract any insights into how the customers behave when purchasing different combinations of services and draw any insight into the Churn patterns for these groups.

As our data is a mix of numerical and categorical variables, we will be unable to use a k-means clustering algorithm. We will be using the k-medoids algorithm or the Partitioning Around Medoids method. The information on how to use this method, how it works, its nuances, along with coding examples using the cluster and Rtsne libraries, is provided in the Medium article, "Clustering on mixed type data - A proposed approach using R," written by Thomas Filaire in 2018. We have also removed the Churn variable so that our clustering does not pick up any patterns due to the churn, as we want to see if we can predict which groups are more likely to churn.

Partitioning Around Medoids

With the k-medoid algorithm, we use Gower distance to compute the averages of the differences between the different entries. Using the Gower distance, we then calculate the Silhouette coefficient to determine the optimal amount of clusters. As Thomas mentions, k-medoid acts like k-modes, where "it produces a "typical individual" for each cluster" and is also memory intensive.

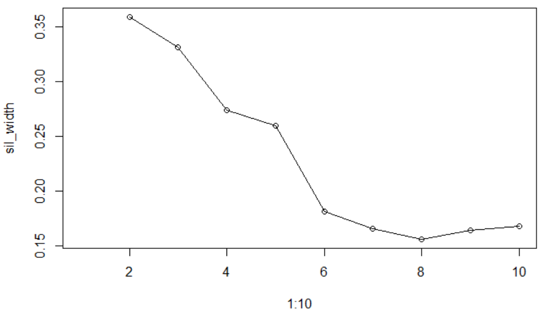
Gower Distance

Gower distance measures the average difference in each component (numerical, categorical) for each entry. The computation is different for each component, and the range of the distance will be between 0 and 1. Gower Distance is calculated for us using the daisy function, which also normalizes the data points.

Optimal k-medoids (Silhouette coefficient)

After computing the Gower Distance for each entry, we will need to see if we can have any insight into which number of k will be optimal for our clustering. Using the Silhouette coefficient, we can determine which k will be optimal for our clustering. The Silhouette coefficient takes the average distances between entries in a cluster and compares them to the other clusters' distances. Using the PAM function in the cluster library, we can determine that we should be looking at a two k-medoid cluster. However, we have also analyzed three and four k-medoid clusters to draw more insight into our data.

Results



After reviewing our clusters, there is an interesting pattern that we can extract from our clusters of data sets. These results are based on a random selection of 3500 entries from our data set. We will be highlighting the main distances between the clusters while introducing pseudonyms for the groups of customers.

2-k Model

With our two k-medoids models, we seem to have two pretty good separate clusters; however, there seems to be a slight overlap between the two groups. Here is a breakdown of the two groups:

Cluster One - 2544 Entries	**Cluster Two - 956 Entries**
Full Service Customers	Phone Only Customers

- **100%** of customers have Internet Service - **88%** of customers in this cluster have a phone line, while roughly half have multiple lines. - **65%** of customers are on a month-to-month contract - **33% Churn**	- **20%** of customers have Internet Service - **95%** of customers have a phone line, where only 28% have multiple lines - **50%** of customers are on a 2-year contract - **7%** Churn

Within these groups, the main difference is that one cluster of customers mainly uses Telco's Phone Service, while the other also uses the Internet Service. There is also a big difference between the contact types between the group and the Churn Rates.

3-k Model

Compared to our 2-k model, our 3-k model provides a more apparent contrast between customers.

Cluster One - 1728 Entries	Cluster Two - 1011 Entries	Cluster Three - 761 Entries
The Essentials Customer	The Bundle Customer	Phone Only Customers
- **20** Month Tenure - **100%** Internet Services - Secondary Services range from **26% to 35%** - **85%** Month to Month - **45%** Churn	- **56** Month Tenure - **100%** Internet Services - Secondary Services range from **65% to 75%** - **50%** 2-year Contract - **12%** Churn	- **30** Month Tenure - **0%** Internet Services - **0%** Secondary Services - **45%** 2-year Contract - **7%** Churn

The 3-k means model breaks down our full-service customers into two clusters, The Essentials Customers, which seem only to purchase the basics—The Bundle Customer, which is more likely to be using all of the Secondary Services. Following the pattern, the Churn rates contrast the groups, as the groups with higher 2-year contract rates have lower Churn rates.

4-k Model

Lastly, our 4-k clustering model further dials into the cluster of customers that use the Internet Service. We will not be including the cluster with phone-only customers in our table, as it has the same characteristics as Cluster Three in our 3-k Model. Clusters one, two, and four have 100% Internet Service.

Cluster One - 1102	Cluster Two - 852	Cluster Four - 786
The Essentials Customer	The Streamer Customer	The Bundle Customer
- **17** Month Tenure - Secondary Services range from **15% to 29%** - **83%** Month to Month - **37%** Churn	- **29** Month Tenure - Secondary Services is at **37%**, where Streaming is at **70%** - **79%** Month to Month - **46%** Churn	- **59** Month Tenure - Secondary Services are between **63%** to **79%** - **61%** 2-year Contact - **7%** Churn

The third group of customers, The Streamer Customer, which like The Essentials Customer, which has a low utilization rate to the Secondary Services, has a very high rate of Streaming Services. As other models also show, there is a contrast between the Churn rates, and there is a connection to the contact type and full utilization of Telco's product line.

With these results, we can determine that there may be a pattern in which specific types of customers are more likely to Churn, and Telco could use this information to concentrate on lowering the Churn rate for these specific groups.

Classification

Our second question is if we are able to find a model that can accurately predict if a customer will churn or not. We created and compared five classification models, logistic regression, LDA, pruned classification tree, GBM, and XG Boost, to find the best model. Each model will be trained on the same 70 percent of data and tested on the remaining data points. Because our second question is to see which variables are the most significant in predicting churn, it is crucial to find a model with high accuracy and high sensitivity. This is because if we can find a relatively high true positive rate model, it means that the model works well in correctly identifying if a customer will churn and, in turn, will find the most significant variables to predict churn.

Logistic Regression

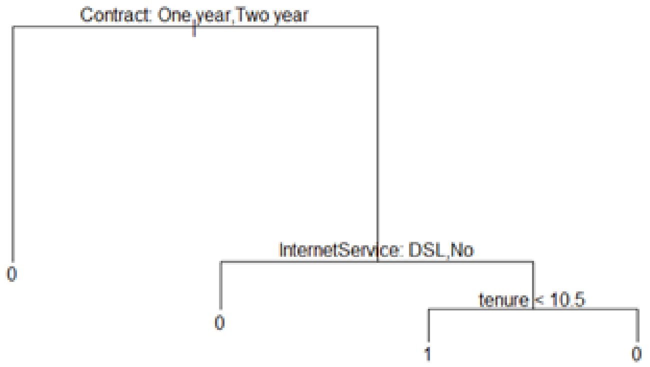
We received an accuracy of 80.85% from our logistic regression model with a true positive rate of 68.62%. This model will determine which variables have the most considerable significance by looking at the p-values. The variables with the most considerable significance, smallest p-value, are tenure, contract type, multiple lines, paperless billing, and total charges. Variables like tenure, contract type, and using multiple lines relate to how satisfied or loyal they are to a company. While if they are using paperless billing and total charges have to do with the financials.

LDA Model

Using the LDA model, we received an accuracy of 80.57% and a true positive rate of 56.60%. The variables with the largest absolute value in the coefficients of linear discriminants will be the most significant. Surprisingly, we see that InternetServicesFiberOptic plays the most prominent role in determining if a customer will churn or not. With a significantly lower sensitivity, this model does not run well in predicting if a customer has churned. This can be explained with the most significant variable being InternetServicesFiberOptic which we believe should not be a significant predictor.

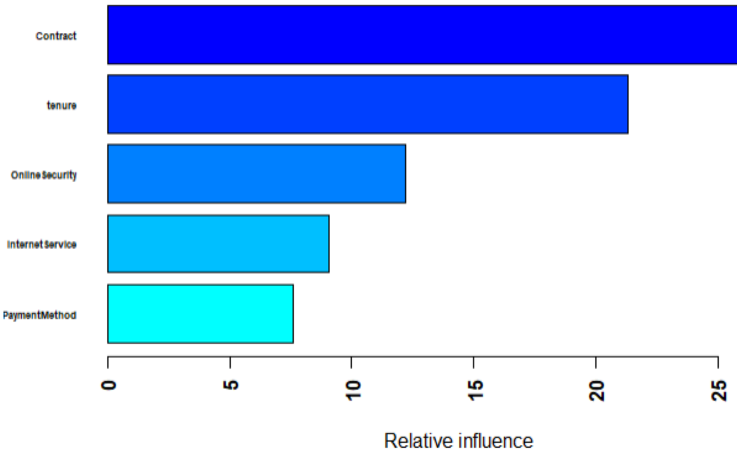
Pruned Classification Tree

The next model we used is a pruned classification tree. We found that the best tree size is when there are 4 terminal nodes. However, we discovered that the model's accuracy does not change if there are 4 or 5 terminal nodes, but we will use 4. Our classification tree's accuracy with 4 terminal nodes is 77.41%, and the sensitivity is 31.25%. The low accuracy can be explained with the first split in the tree; if a customer does not have a one-year or two-year contract, the model has predicted that a customer has churned. However, it is not always the case that month-to-month customers will churn, and therefore our accuracy goes down. Overall, we see that the variables that play the biggest role in determining churn are contract, internet service, and tenure, consistent with what we see in the two models above.



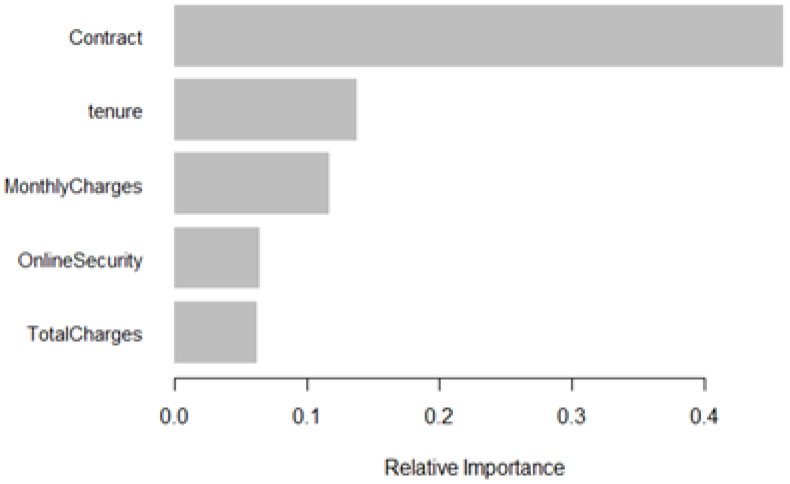
GBM Model

From our GBM model, we will be using the multinomial distribution, which will perform on ten cross-validation folds and shrinkage of 0.1. The model can predict if a customer will churn or not 80.90% of the time and has a true positive rate at 69.89%. The variables that play the largest significance in predicting churn will be found in the relative influence bar graph. Following our other models, contract and tenure play the biggest role. However, in this model OnlineSecurity and InternetService play the second biggest role.



XGBoost Model

Our last and best model uses XG Boost uses the objective function binary: logistic, boosting step size of 0.1, maximum depth of tree at three, and a maximum number of iterations at 100. The accuracy is 80.90%, and the sensitivity is 70.16%. This will be our best model because it has the highest accuracy but also the highest sensitivity. Like all other models, the variable with the most significant relative importance is contract followed by tenure and monthly charges.



Results

In all the models, we constantly see that contract and tenure play the most significant role in predicting if a customer will churn or not. Other variables that play a factor are MonthlyCharges, online security. Therefore, we can make recommendations to the company based on this information. For new customers, we recommend finding incentives that will make customers sign for more extended contracts and will, in turn, have a higher tenure. The company could create an incentive to sign contracts with lower monthly charges. This can be seen in the Telus Easy Payment program, where they allow a customer to spread out their device's cost over 24 months instead of paying the upfront cost at once. We suggest rewarding customers for staying with the company longer, which can be rewards such as lower monthly charges or offering additional services such as free online security or tech support. The Telus Bring it Back program lowers the upfront cost of a new device if the customer brings back their old Telus device, thus encouraging customers to stay within Telus to get a new device for cheaper.

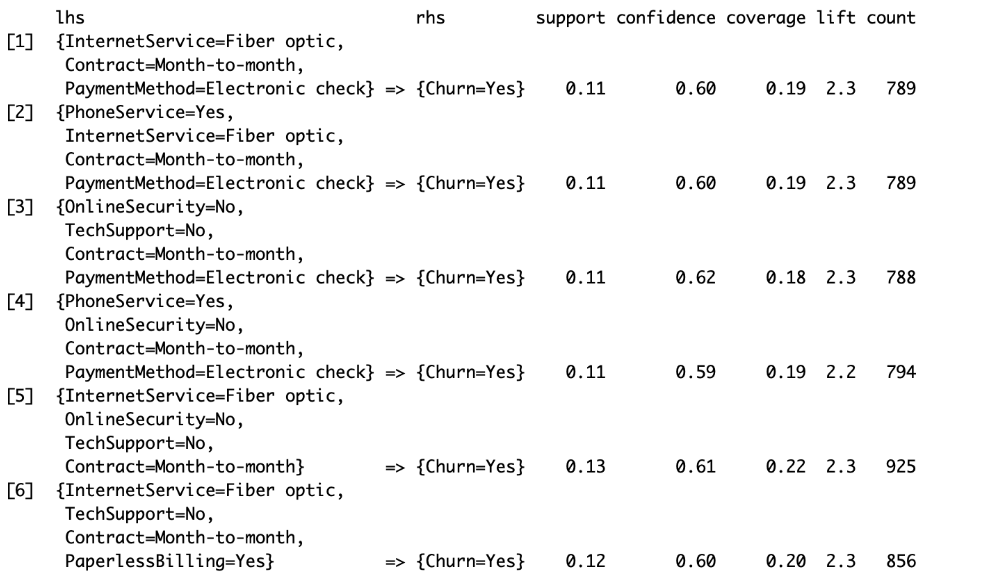
Implementing Clusters

The next question is if the accuracy of our XG boost model will increase if we include clusters. To do this, we included the column indicating what cluster a customer is in. Our accuracy does not increase when running the XG boost model with 2, 3, and 4 models. This concludes that including the clusters does not improve our model..

Model	Accuracy
Base Model	80.90%
2 Clusters	80.74%
3 Clusters	80.74%
4 Clusters	80.46%

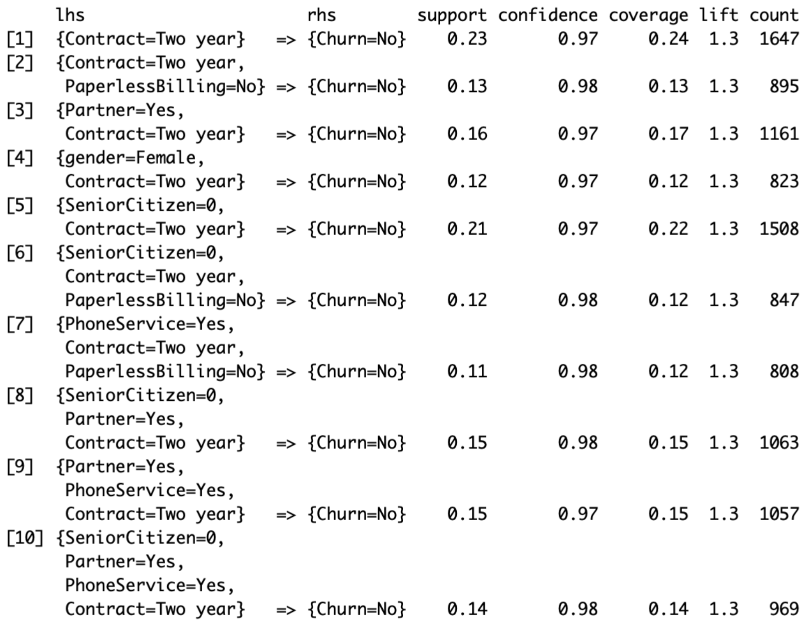
Market Basket Analysis

We used association rules with the market basket analysis to see if we could find any interesting relationships between the variables. Specifically, we wanted to see if there were any rules or characteristics to target to lessen customer churn. Initially, when we ran the data through the apriori function, we obtained roughly 53,000 rules. This was too many rules to look through, so to narrow our focus, we adjusted our code so that the consequent rule was either "Churn=Yes" or "Churn=No." This would allow us to specifically look at any rules that are associated with customers who churn or not. Also, to further reduce the number of rules, we adjusted the support and confidence levels to give us the most robust rules.



```
    lhs                                    rhs         support confidence coverage lift count
[1] {InternetService=Fiber optic,
     Contract=Month-to-month,
     PaymentMethod=Electronic check} => {Churn=Yes}    0.11       0.60     0.19  2.3   789
[2] {PhoneService=Yes,
     InternetService=Fiber optic,
     Contract=Month-to-month,
     PaymentMethod=Electronic check} => {Churn=Yes}    0.11       0.60     0.19  2.3   789
[3] {OnlineSecurity=No,
     TechSupport=No,
     Contract=Month-to-month,
     PaymentMethod=Electronic check} => {Churn=Yes}    0.11       0.62     0.18  2.3   788
[4] {PhoneService=Yes,
     OnlineSecurity=No,
     Contract=Month-to-month,
     PaymentMethod=Electronic check} => {Churn=Yes}    0.11       0.59     0.19  2.2   794
[5] {InternetService=Fiber optic,
     OnlineSecurity=No,
     TechSupport=No,
     Contract=Month-to-month}        => {Churn=Yes}    0.13       0.61     0.22  2.3   925
[6] {InternetService=Fiber optic,
     TechSupport=No,
     Contract=Month-to-month,
     PaperlessBilling=Yes}           => {Churn=Yes}    0.12       0.60     0.20  2.3   856
```

As pictured above, for customers who churned, we set a confidence level of at least 0.59 and a support level of at least 0.11. With these parameters, we were able to achieve a manageable number of rules to analyze. Because our data resulted in so many rules, we found that most of them had a support of 0.10. Increasing the support to 0.11 eliminated many of the rules, and the higher the support, the more prominent that rule is amongst the dataset. Also, we wanted the highest possible confidence interval and found that at the 59% confidence, we achieved a significant reduction in rules. Unsurprisingly, most customers who churned were just on a month-to-month contract. By looking at some of the rules' characteristics, this group didn't seem to use secondary services (like online security, back-ups, tech support, etc.). One interesting note was that many of these rules included the fiber optic option in their internet service. This caught our eye because, generally speaking, fiber optics usually provides the fastest internet speed compared to, in this case, DSL (or no internet service at all).



```
     lhs                      rhs        support confidence coverage lift count
[1]  {Contract=Two year}   => {Churn=No}    0.23       0.97     0.24  1.3  1647
[2]  {Contract=Two year,
      PaperlessBilling=No} => {Churn=No}    0.13       0.98     0.13  1.3   895
[3]  {Partner=Yes,
      Contract=Two year}   => {Churn=No}    0.16       0.97     0.17  1.3  1161
[4]  {gender=Female,
      Contract=Two year}   => {Churn=No}    0.12       0.97     0.12  1.3   823
[5]  {SeniorCitizen=0,
      Contract=Two year}   => {Churn=No}    0.21       0.97     0.22  1.3  1508
[6]  {SeniorCitizen=0,
      Contract=Two year,
      PaperlessBilling=No} => {Churn=No}    0.12       0.98     0.12  1.3   847
[7]  {PhoneService=Yes,
      Contract=Two year,
      PaperlessBilling=No} => {Churn=No}    0.11       0.98     0.12  1.3   808
[8]  {SeniorCitizen=0,
      Partner=Yes,
      Contract=Two year}   => {Churn=No}    0.15       0.98     0.15  1.3  1063
[9]  {Partner=Yes,
      PhoneService=Yes,
      Contract=Two year}   => {Churn=No}    0.15       0.97     0.15  1.3  1057
[10] {SeniorCitizen=0,
      Partner=Yes,
      PhoneService=Yes,
      Contract=Two year}   => {Churn=No}    0.14       0.98     0.14  1.3   969
```

Above, we can see some of the key rules for customers who did not churn. For this group, we set a confidence interval of at least 0.95 and a support level of at least 0.11. We applied the same logic to these parameters as we did to the customers who churned (see the previous paragraph). Again, unsurprisingly the customers that churned the least were the ones that were on the 2-year contracts. From what it looks like, a lot of those customers were using the phone services from Telco. Quite a few rules also stated that many individuals were not senior citizens and had partners. A lot of these findings were also reinforced by our findings from our exploratory data analysis.

Results

Because we had a much higher confidence level for customers who did not churn, it would be better to focus on and create strategies from those rules. To summarize, Telco should try and push for 2-year contracts as this will inevitably decrease the customer churn the most. One of the other more prominent rules was that 21% of customers who did not churn signed a 2-year contract and were not senior citizens, with a confidence level of 97%. This would be useful for Telco as it would narrow down the demographic. For the future, maybe focus on a marketing strategy to target that senior citizen group and increase that group's chances of not churning. Finally, because fiber optics were pretty prominent in the association rules for customers who churned, maybe Telco should investigate why this is the case. Maybe customers who were on those month-to-month contracts didn't like the internet speed from the fiber optics.

Conclusion

Our analysis has come to a few conclusions that could provide Telco with valuable information in predicting customer churn and areas to focus on to reduce customer churn:

1. When we are grouping customers based on their service data, we find that the customers who have the higher churn rates are subscribed to phone or internet services but no secondary services like security, tech support, back-ups, etc. Also, customers who included streaming services noticed higher churn rates.

2. To better predict which customers would churn or not, using the XG Boost method produced the best results with an 80.9% accuracy and a 70.2% true positive rate. We also found that contract, tenure, and monthly charges had the highest relative importance in terms of which customers churned through this analysis.
3. We found that customers who churned were on month-to-month contracts and didn't entirely use the secondary services through market basket analysis. The customers who did not

churn were on the 2-year contracts used the phone service, and non-seniors with a 2-year contract generally did not churn.

Based on our analysis, we recommend that Telco find incentives that will make customers sign those long 1-2 year contracts. Maybe Telco could offer lower monthly charges when customers sign a long-term contract. Also, offering packages that provide secondary services could convince customers to stay longer as well. By offering a more comprehensive range of services, customer satisfaction may improve, and the churn rate may diminish. A strategic ad campaign targeting senior citizens, which makes up roughly 16.2% of the customers in the dataset, may improve the demographic's retention rate.

Limitations and Next Steps

Due to this project's time constraint, we weren't able to analyze more association rules because there were initially 53,000 of them. We limited our search insignificant rules by looking at consequent rules that resulted in "Churn=Yes" or "Churn=No" to save time. There may have been more interesting relationships in the ones that we had omitted.

In terms of the following steps to be taken, analyzing customer service data, surveys, call logs, and social media posts would greatly benefit. Customer satisfaction plays a significant role in churn, and understanding how a customer feels may mitigate or provide more information as to why they churn or not. Also, we could look into a cost-benefit analysis of having a loyalty program by providing rewards that don't necessarily have something to do with Telco (discounts at movie theatres/restaurants, etc.).

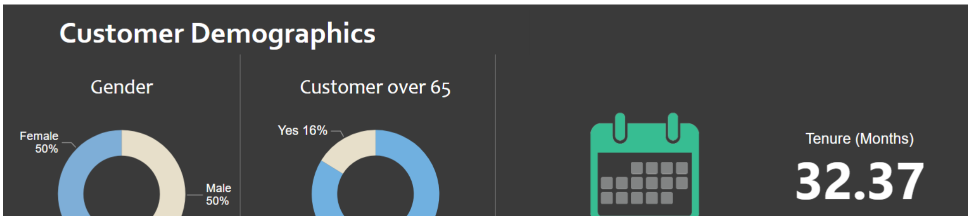
References

Statista (2021, February 09). Customer service: Churn rate by INDUSTRY U.S. 2020. Retrieved April 15, 2021, from https://www.statista.com/statistics/816735/customer-churn-rate-by-industry-us/

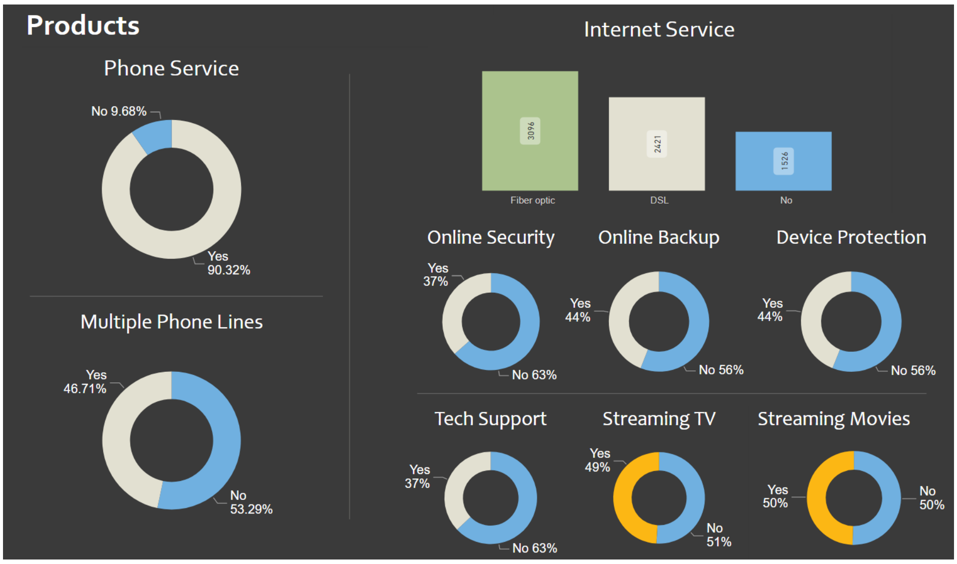
Churn is breaking the telecoms market: Here's how to fix it. (2018, September 27). Retrieved April 15, 2021, from https://telecoms.com/opinion/churn-is-breaking-the-telecoms-market-heres-how-to-fix-it/

Filaire, T. (2018, October 28). Clustering on mixed type data. Retrieved April 15, 2021, from https://towardsdatascience.com/clustering-on-mixed-type-data-8bbd0a2569c3

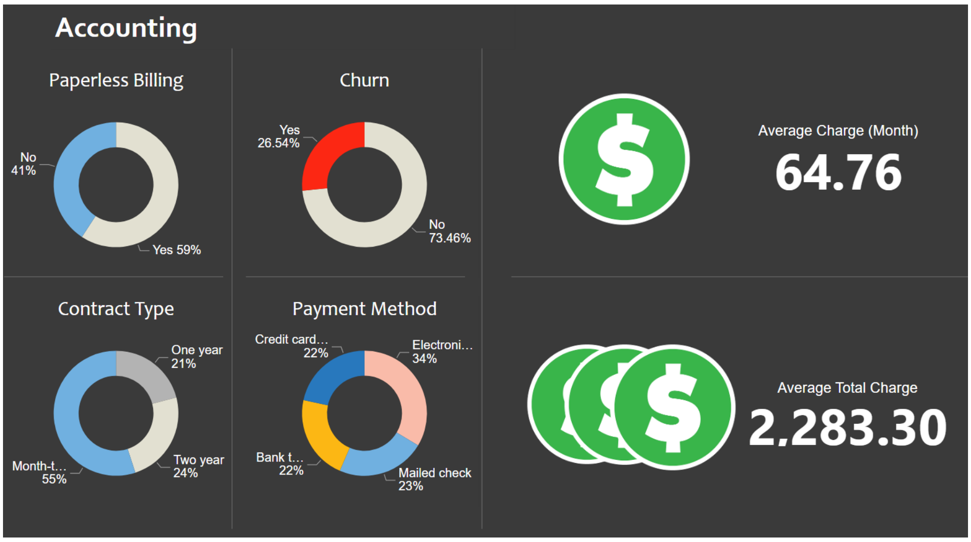
Appendix



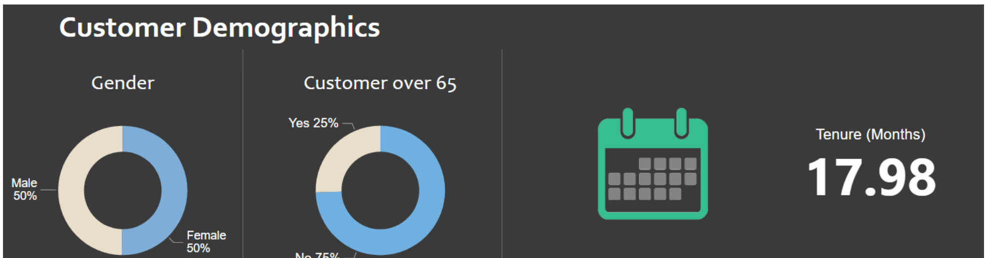
Telco Customer Demographics - Full Data Set



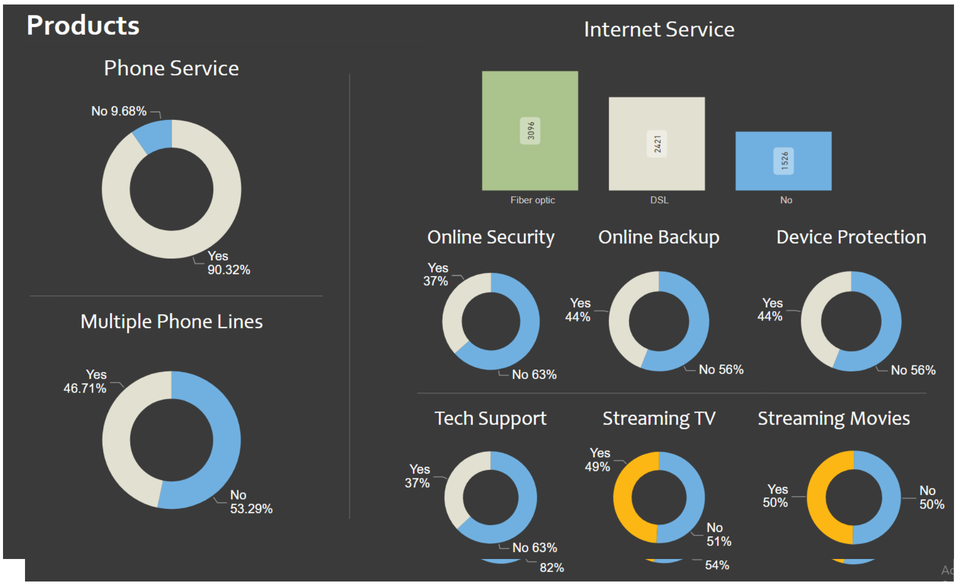
Telco Product Line Usage - Full Data Set



Telco Customer Accounting - Full Data Set



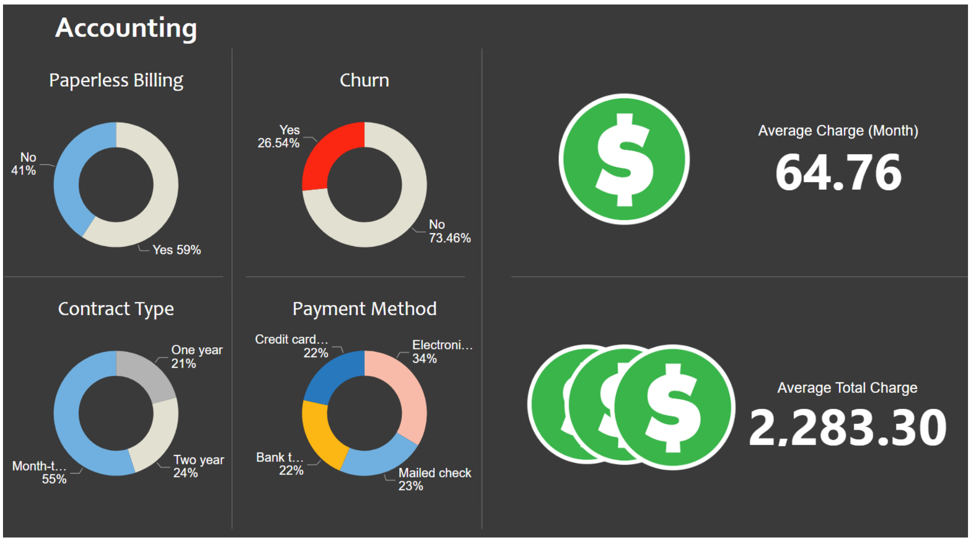
Telco Customer Demographics - Churned



Telco Product

Line

Usage - Churned



Teclo Customer Accounting - Full Data Set