D208 Predictive Modeling Task 2

The telecommunications industry has many providers to choose from and customers can switch between providers at free will. For that reason, telecommunication companies must prioritize on retaining existing customers before they consider joining a competitor. When a customer stops using a specific provider’s product or service for a certain time frame, it is called churn (WGU Data Dictionary). The goal of this analysis is to find trends in customers that churn and in addition, reduce the amount of churn in the future.

The main question that will be tackled in this analysis will be “is it possible to calculate the odds of a customer to churn based off their demographics and the amount of service they are provided?” Telecommunications company provide numerous services ranging from phone, TV streaming, and more. The dataset given supplies data for services that each customer signed up for, customer account information and demographics. The objective is to narrow down customer characteristics to find statistically significant correlation to churn. If the variables are significant, then a logistic model can be created to calculate the odds of a customer leading to churn in the next month. With a successful model, the telecommunications company can take action to try and retain the customer for a longer tenure.

The method that will be used for this analysis will be logistic regression. The assumptions that are made before working on this specific method is that the response variable is dichotomous, and the distribution is binomial. In this analysis, both the response variables and predictor variables are categorical. Thus, making the question that is being answered a classification problem. Each customer will be put in a reference group based off the predictor variables, and that will calculate the probability of them churning. The goal is to calculate odds of a customer churning based on the predictor variables of the customer’s demographic and service history.

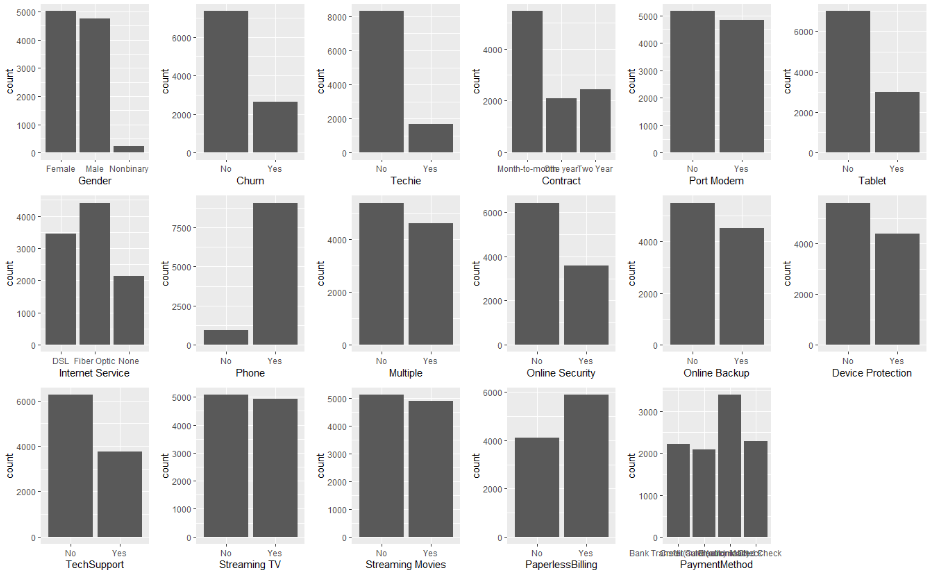
For this analysis, the program R will be used. R has many packages to use during the data preparation phase. With the packages “tidyverse” and “dpylr” the data can be prepared through data manipulation. After that data has been manipulated a summary statistic can be run through the data frame. The package “ggplot2” will help create visuals for univariate, bivariate distributions, and more. Using the “glm” function, R can also create the logistic regression model. Once the model is constructed, the initial model will be fitted, altered, and reduced to find the most successful model. Another benefit of R is that the program has a package to create a confusion matrix. Overall, R helps in making statistical analysis more convenient.

The goal during the data preparation phase is to confirm assumptions made for logistic regression by calculating summary statistics of predictor variables and creating distribution visuals. After loading the dataset given with the assignment, the first step is to create a data frame with only a unique identifier, response variable and predictor variables. For this analysis, the response variable is the variable “churn.” The initial predictor variables will be Gender, Techie, Contract, Port modem, Tablet, Internet Service, Phone, Multiple, Online Security, Online Backup, Device Protection, Streaming TV, Streaming Movies, Paperless Billing, and Payment Method. Once the data frame has been made, the data type of each variable will be changed into a factor. This is to allow the variable to be recognized as a binary when creating a logistic model. After the data is manipulated, the summary function can draw the variables summary statistics. The next step is to create visuals of the data’s distributions. Lastly, it is important to report any interesting findings before creating the model.

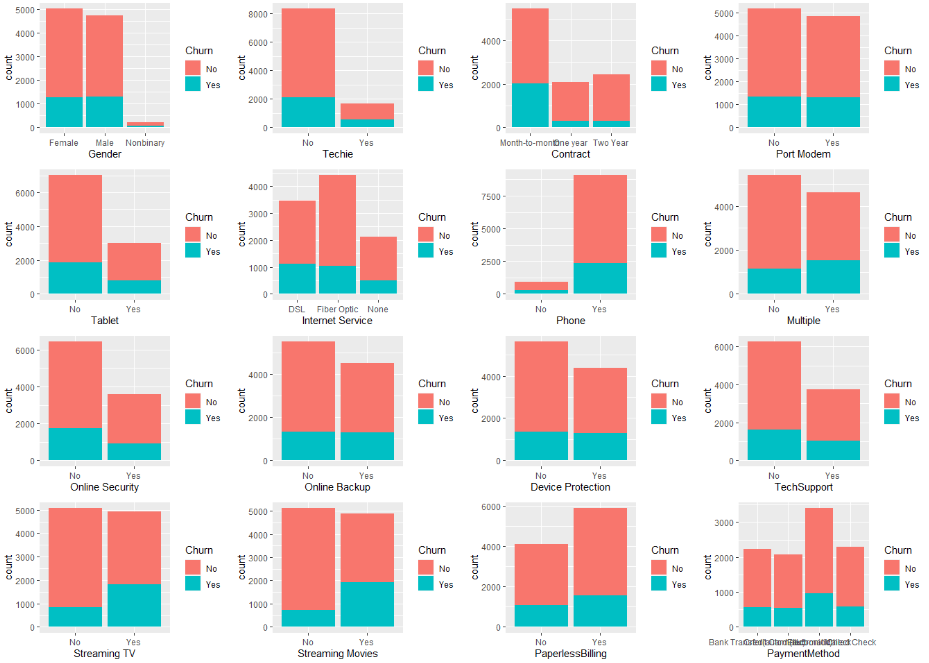
The summary statistics of the variables mentioned before have a few interesting findings. First off, there is about a 25% churn rate in the dataset. Secondly, most customers do not identify as a techie and most customers also do not own tablets. Lastly, 90% of customer receives phone services from the company. Below are screenshots of the visuals made from R and packages used were “ggplot2” and “gridExtra.”



Summary Statistics Figure

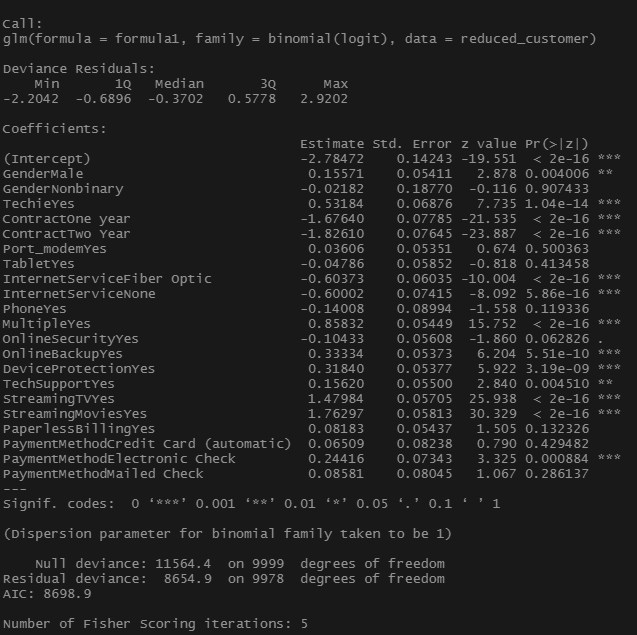


Univariate Distribution Figure

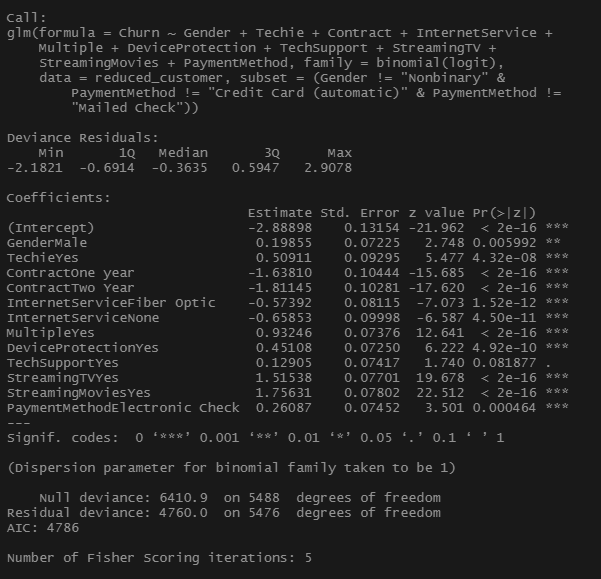


Bivariate Distribution Figure

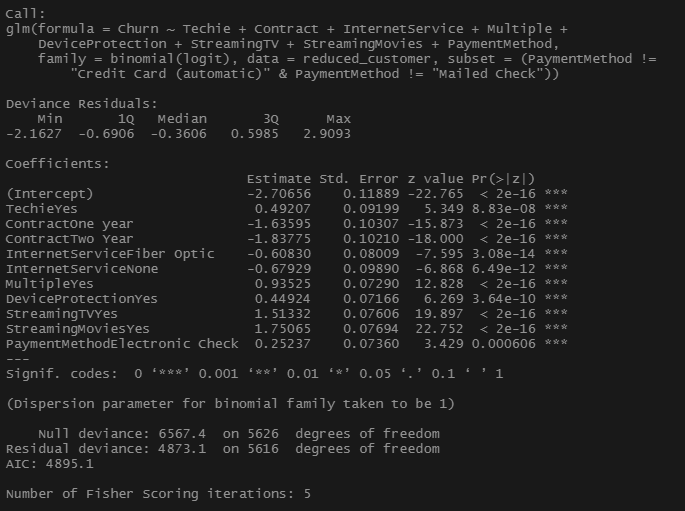
The first logistic regression model included the response variable “Churn” and all the predictor variables that were listed earlier. One of the objectives for the analysis was to narrow down the customer characteristics to those that are statistically significant to the variable churn. To reduce the model, any variable that was not statistically significant in the initial model was removed. Looking at the summary for the reduced model, there is a drastic drop in both null and residual deviance. The change in the model, led to an improvement of fit for the data. Now taking it a step further, another model can be made by removing the variable “Tech Support” and the variable “Gender.” This would make the next reduced model only contain variables that are 0.99 statistically significant. The summary of the second reduced model shows that the residual deviance has increased. This indicates that the removal of the variables “Gender” and “Tech Support” led to an increase in variance. For that reason, it is safe to assume that the first reduced model is a better fit for the data. After running the first reduced model into a confusion matrix, it is shown to have a 78% accuracy. The model has a sensitivity of .9046 and specificity of .4590. This means the model is great at detecting true positives but not true negatives. However, even if the model is not great at detecting true negatives or people that end up to churn, the model can still be used for the telecommunication company’s advantage.



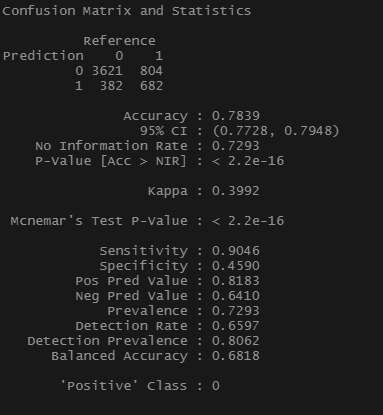
Initial Logistic Regression Model Figure



First Reduced Logistic Regression Model



Second Reduced Logistic Regression Model



Confusion Matrix of predictions made from the first reduced logistic model

The results of the analysis have made a logistic regression equation that has a 78% accuracy. The logistic regression equation is:

Churn = -2.89 + 0.19 (GenderMale) + 0.51 (TechieYes) + (-1.68) (One Year Contract) + (

1.81) (Two Year contract) + (-0.57) (Fiber Optics) + (-0.66)(No Internet) + 0.93 (MultipleYes) + 0.45 (Device Protection) + 0.13 (Tech Support) + 1.52 (Streaming TV) + 1.75 (Streaming Movies) + 0.26 (Electronic Check Payment).

Looking at the coefficients from the equation, variables that have a positive relationship to churn are customers that identify as male, identify as techie, have multiple lines, use services such as device protection, tech support, TV streaming, movie streaming, and uses electronic check payments. On the other hand, the variables that have a negative relationship are customers that are on contract, use fiber optics internet or does not have the internet service at all. The positive relationship will increase the likelihood of churn while the opposite is said for the negative relationships. The model of the analysis is limited to its own reference group. It only considers the customer demographics and characteristics. This model does not reflect how a customer views the company and their service. The benefits of the model are that it shows which aspect of a customer will increase the chances of them to churn and which aspects will retain the customer.

With the model, the strategy the telecommunication should take is to prioritize the male techie demographic, sign customers to contracts, and push fiber optic services. The telecommunications company can filter which customers are likely to churn and send out discounts, promotional deals, and other ads to retain the customer. They can start by targeting customers that identify as male and techie because those groups have a higher chance to churn. In addition, the company will need to focus on customers that are up for contracts by reaching out to them by phone or email with a new offer. Furthermore, the chance of churn can be lowered at the start if sales agents in the company prioritize signing a new customer onto a long-term contract. The analysis shows that customer with fiber optics internet is more likely to stay with the company. Making this service more available to customers will both increase sales and the likelihood of retaining the customer. The company can develop different strategies and prioritize a selected customer group when using the logistic regression model.

References

WGU Data Dictionary. D207 D 208 D Churn Data Consideration and Dictionary. Retrieved January 25, 2021, from https://tasks.wgu.edu/student/001407536/course/20900013/task/2786/overview