Customer Retention

Christopher Flowers

College of Information Technology, Western Governors University

D208 Predictive Modeling

Dr. Straw

November 14, 2021

**Part I: Research Question**

A.

1.  How can churned customers help reduce the attrition of customers?

2.   Below are some goals and objectives to pay attention to.

Does customer contract type depend on the customer's age?

Is churn affected by the loyalty of the customer?

Do paperless billings affect the churn?

Does having internet service affect the churn?

What variables give the greatest gauge of what customers will leave? The goal is to figure out what categories need to be focused on to mitigate the number of customers lost. Also what categories have the highest effect on the churn.

By doing this we can see if modern culture has any effect on customers leaving the company. This will allow the company to determine how it wants to create packages in the future to optimize customer retention and entice new customers as well.

**Part II: Method Justification**

B.

1.   Assumptions of logistic regression according to Assumptions of Logistic Regression(2021):

* Binary Logistic Regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal
* Requires observations to be independent of each other.
* Assumes linearity of independent variables and log odds.
* Requires larger sample size

2.   For this course, I have chosen to use the R programming language. Using R has benefits over other programming languages, such as Python. One of the benefits of using R programming is that it has a built-in graphical library, statistical analysis, and the ability to clean the data. It also can install and use additional libraries. R can run just statistical analysis, but it also can run a deep statistical analysis with fewer lines of code than other languages. Additionally, R has some shortcuts for traditional programming concepts like its shortcut for if-else.

3.   Logistic Regression was the appropriate technique to analyze the research question because of its binary or binomial response. This is the other appropriate because the dependent variable is a binomial response in Churn, which values are ‘Yes or No’. The primary objective is to find out what predictor variables are most likely to contribute to customers churning the customers. Also, it allows the ability to see how each variable impacts the dependent variable.

**Part III: Data Preparation**

C.

1.   The data preparation goals are to load the data set into a variable then clean and narrow the data set. This will be done by reading in the CSV file to import the datasheet into R. At this point, the data can continue to be prepared and can be manipulated as well. Continuing with the preparation, we will summarize the data to look for missing data and what data columns are relative to the research question. Then, the unnecessary columns will be removed. The goal of data manipulation is to see how the variables are distributed and to compare independent variables to the dependent variable. In manipulating the data, we will split the data set into two data sets to give a better sample size of the data. The showing of the bivariate and univariate data will also help to manipulate the data.

2.  This dataset has several continuous independent variables, but not all of these variables are needed for the logistic regression. The continuous independent variables needed are Children, Age, Income, Outage\_sec\_perweek, Email, Yearly\_Equip\_Failure, Tenure, Monthly Charge, Bandwidth\_GB\_Year, Item1, Item2, Item3, Item4, Item5, Item6, Item7, and Item8. These values majority of the time will be different from each other. There are several categorical independent variables most of which contain only two categories, which are either ‘Yes’ or ‘No’. These variables are Gender, Techie, Contract, Port\_modem, Tablet, InternetServices, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, and PaperlessBilling. All these variables are different than the other variables except for the dependent variable will also be categorical. The last categorical variable is going to be the categorical dependent variable and to use this to determine which variables played a part in customers churning. Yes, there is an identifier for all records because there are 10,000 records in the dataset and there are no NA’s in the dataset as seen in the summary. Whether those who have churned will play an effect on whether the customer will Churn in the future.

3.  To prepare this data for analysis, the first thing to do is to load the libraries used such as ggplot2, openxlsx, assertthat, dplyer, readr, and tidyverse. These libraries give the use of functions to be able to do certain things like openxlsx allows for the creation of an excel file. After the libraries are loaded the next thing to do is to read the excel file into R and place it into a variable, which can be manipulated, in this case, the variable is called churnData. Next is to use the str() to give us an idea of the data and data types present in the data. Now let’s take a look at the summary for the churnData dataset to give a total of each category for categorical data and continuous data will have minimums and maximums for each variable. The next thing to happen will be to remove any unneeded variables from the dataset. Most removed variables are going to be those in which are considered demographic data like City and County. But some other variables are not needed and will need to be removed as well to help reduce outliers in the data. After removing the unneeded columns, the next thing needed is to remove any N/A observations or other missing data. This is done by using the na.omit function and the complete.case function followed by using the glimpse function to take a look at the data to see if any data was removed by looking at the total row count. For readability reasons, we change the Item numbered columns to named columns. Next, we changed any columns with an underscore to something without the underscore. For example, Bandwidth\_GB\_Year, Port\_Modem, Outage\_sec\_perweek, and Yearly\_Equip\_Failure will change respectfully to Bandwidth, PortModem, OutageSec, and EuipFailure. The last thing done is to split the data into two different datasets test and train with the train being the larger set. This is to get a better sample size than the original sample size and then export all new datasets to excel.

Text

Description automatically generated

Above is the loading of the libraries, reading the CSV file, running the str and summary functions. This screen is used to remove the variables that are not needed so that we can summarize the data to check that these variables have been removed. We begin by checking for missing data and checking the row count. Then, we rename some variables. Once the variables are renamed, we check the renaming.

Text

Description automatically generated

Results of loading libraries and reading the CSV file in.

Text

Description automatically generated

Results of the Str function.

A picture containing calendar

Description automatically generated

Table

Description automatically generated

Results of the summary of the churnData datasets.

A picture containing calendar

Description automatically generated

Summary of the churnData datasets after removing the demographic variables.

A picture containing graphical user interface

Description automatically generated

Summary of churnData dataset after removing the other variables.

Background pattern

Description automatically generated

Results of checking for missing data or N/A observations.

A picture containing calendar

Description automatically generated

Results of renaming variables.

Text

Description automatically generated

Splitting the dataset into two and looking at the two different datasets then exporting all datasets to excel.

Background pattern

Description automatically generated

Results of splitting the dataset and the results of the glimpse for the train.

Background pattern

Description automatically generated

A glimpse of test dataset and exporting the datasets to excel

4.

Text

Description automatically generated

Chart, bar chart

Description automatically generated

Text

Description automatically generated

Chart, bar chart

Description automatically generated

Text

Description automatically generated

Chart, bar chart

Description automatically generated

Text

Description automatically generated

Chart, bar chart, waterfall chart

Description automatically generated

Text

Description automatically generated

Timeline

Description automatically generated with medium confidence

Text

Description automatically generated

Timeline

Description automatically generated

A screenshot of a computer

Description automatically generated

Chart, bar chart

Description automatically generated

Bivariate

Text

Description automatically generated

Chart

Description automatically generated with medium confidence

Chart, box and whisker chart

Description automatically generated

Calendar

Description automatically generated

Text

Description automatically generated

Calendar

Description automatically generated

Calendar

Description automatically generated

Chart

Description automatically generated

Text

Description automatically generated

Chart, calendar

Description automatically generated

Chart, calendar

Description automatically generated

5.  Provide a copy of the prepared data set.



**Part IV: Model Comparison and Analysis**

D.

1.

Text

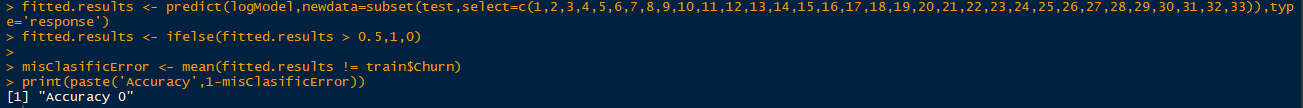
Description automatically generated

A picture containing text

Description automatically generated

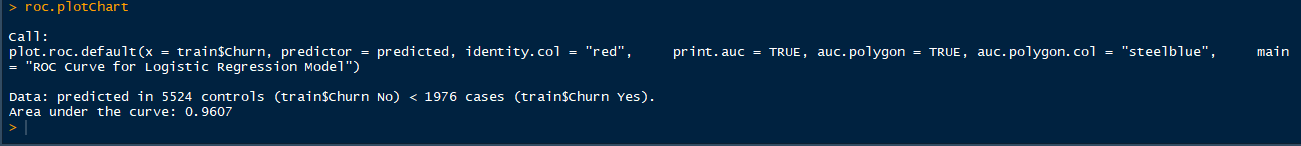
A picture containing text

Description automatically generated



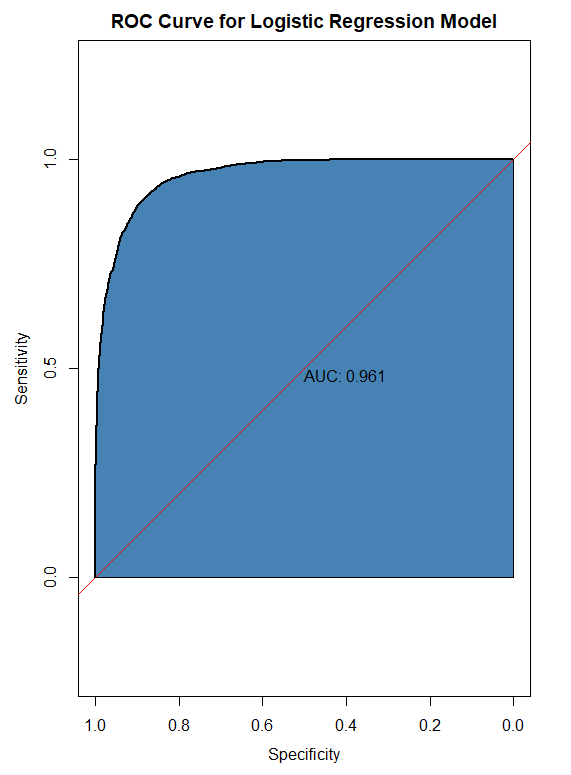
Text

Description automatically generated



Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

The residual of all predictor variables minimum is -2.81 with a maximum of 3.43. The intercept and 4 other variables have significant codes of 0.001. One variable has a significance code 0.01, three variables with a significance 0.05 and one variable with a significance code of 0.1.The estimate compared to the std error for the most part are comparable for most variables. But there other variables std error and estimate are completely noncomparable. The formula of the regression is y = -3.742+ 1.121 \* children – 8.511\* Age + 7.967 \* Income + 4.595\* GenderMale -3.796\*GenderNonBinary + 3.597 \* OutageSec – 3.065\* Email - 9.190 \*EquipFailure + 1.147 \* TechieYes – 3.285\* ContractOne Year – 3.477\* ContractTwo Year + 1.11\* PortModemYes – 8.183 \* TabletYes – 3.751\* InternetServiceFiber Optic – 2.164\*InternetServiceNone – 2.767\* PhoneYes + 3.234\* MultipleYes+ 9.316\* OnlineSecurityYes -7.093\* OnlineBackupYes- 9.211\* Device ProtectionYes- 3.151\*TechSupportYes+1.424\* StreamingTVYes +1.372\*StreamingMoviesYes + 2.233\* PaperlessBillingYes + 1.615\* Tenure + 5.156\* MonthlyCharge – 3.366\* Bandwidth – 2.522\* TimelyResponse – 2.574\* TimelyFixes +4.540 \*TimelyReplacements – 4.937\* Reliability – 1.134 \* Options – 1.133\* RespectfulResponse - 1.690 \*CourteousExchange + 1.229\* ActiveListening.

2.  Using Husson’s FactoMineR function RegBest we can determine the best variables to use in the reduced logistic regression model(Husson 2021). Based on this the best statistical variables are Churn, StreamingMovies, and Tenure. Tenure and StreamingMovies have both p-value significant codes 0.01 then Churn and the intercept of p values significant codes of 0.001.

3.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated with medium confidence

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, pie chart

Description automatically generated

The reduced model is made up of only three variables, the two independent variables, and one dependent variable. The equation of the model is y =-0.4.64 +1.870\*StreamingMoviesYes -0.063\*Tenure.

E.

1.  Refer to D1 and D3 for residuals of each model. The logical variable selection was done by using Hussion’s FactoMineR’s RegBest function which statistically calculates the best variable to be used(Husson 2021). The model evaluation metric was completed by looking at the ROC value for both models separately and the results were close to one another. But the initial logistic regression was better has it had a score over 96% meaning 96 percent of the predicted values were predicted correctly compared to the nearly 87 percent predicted correctly in the reduced logistic regression model.

2.  Refer to the above to see any calculations and code output.

3.  All code is included above.

**Part V: Data Summary and Implications**

F.

1.  Based on the reduced logistic regression model using the following independent variables StreamingMovies, and Tenure. The ROC value of this model was 0.869, which would be approximately 87 percent of the predicted values are correct. This means that 98 percent of the model variables are correlated in the model. The equation of the model is y =-0.4.64 +1.870\*StreamingMoviesYes -0.063\*Tenure. In other terms coefficients mean StreamingMoviesYes will increase 1.870 units times Churn and Tenure will decrease 0.063 units times Churn. All variables are statistically significant to 0.001 and all have the same p-values of <2e-16. There could be some limitations to the analysis of the dataset used is too small and whether not splitting the dataset would have fixed these issues.

2.   For this course, we were asked to help a telecommunication company with the churn of its customers. In this report, we looked at customer churn of the telecommunication company to help mitigate the customers leaving the company. Based on the results from the reduced regression model using StreamingMoviesYes, and Tenure. All of these variables have the same significant codes and p values, meaning they have a direct relationship to the dependent variable of Churn. The recommendation to the company would be to focus on the customers who stream movies with less tenure. So offering a package or packages with higher bandwidth caps and or has a lower fee for going over the bandwidth cap to allow for more or faster movie streaming. Also, the company should look at giving discounts to customers who have been tenured longer with the company but also not forget about short-tenured customers. In conclusion, the company needs to focus on customers streaming movies and who have longer tenure because these customers are valuable to the company.

G.  <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=d18ecba4-b1ea-49ec-99af-ade4003fbdd8>

H.  None

I.   Sources

*Assumptions of logistic regression*. Statistics Solutions. (2021, August 11). Retrieved November 11, 2021, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/.

Francois Husson, J. J. (2021, January 8). *RegBest: Select variables in multiple linear regression in factominer: Multivariate exploratory data analysis and Data Mining*.

J.  Demonstrate professional communication in the content and presentation of your submission.