Tonie Cooper

Master Data Analytics

[Tcoo158@wgu.edu](mailto:Tcoo158@wgu.edu)

D208 Task 2 Assessment

A1) What customers are at a high risk of churn?

A2) Stakeholders will be able to identify what customers are at a high risk of churning. This will allow stakeholders to focus on those areas and make decisions that will have a positive impact on those consumers.

B1) Assumptions of a logistic regression model include (Statistics Solution 2021):

Does not require a linear relationship between dependent and independent variables.

Residuals do not need to be normally distributed

Homoskedasticity is not a requirement

Dependent variables aren’t measured on a ratio or interval scale

There can be little to no multicollinearity between the independent variables (IBM 2022)

B2) The benefits of using Python and Jupyter notebooks are:

Python uses a large system with many packages for data science and machine learning. Python operates quickly when compared to R or MATLAB. Python is a cross-platform tool so it can be used with Windows PC’s or MacBooks. This is very important as both operating systems are widely used across the world. Finally, Python has a large support system and is among the most popular data science programming languages used today (Qureshi 2020, Swaminathan 2019).

B3) Logistic regression is the appropriate technique to analyze churn because churn is a binomial categorical variable. By using the independent variables provided a logistic regression model will help us understand the increased probability of churn as we add or subtract different independent variables (Navlani 2019).

C1) My preparation and manipulation goals will include:

1. Backing up my data to my computer. 2. Import CSV file into Jupyter Notebooks. 3. Rename dataset to ‘churn\_df’. 4. Correct any misspellings or typos to have a uniform dataset. 5. Find outliers that may distort results and remove if necessary. 6. If any missing data fill in using mean, median or mode.

Churn (Yes or No) will be the most relevant variable to this process. Churn will be used as our categorical target variable.

Finally, the predictor variables from the customer survey will be relevant to the decision-making process and deciding what variables increases and decreases the churn rate.

C2) Statistical Summary:

The dataset originally consists of 50 columns and 10,000 records. The user ID and demographic columns will be removed from the data frame as little to no importance is placed on those variables. Next, binomial columns were changed to 1/0 to allow for the analysis and computations included in the model. 34 numerical columns are now remaining, and no null or NA/missing data points are present. Created box plots and histograms which revealed a normal distribution for Monthly\_Charge, Outage\_sec\_perweek and Email. Histograms created for Bandwidth\_GB\_Year and Tenure showed a binomial distribution which demonstrates a linear relationship in the scatterplot.

C3) Steps to Prepare Dataset:

Import dataset

Rename survey variables to a recognizable name

Get description of dataframe

View summary statistics

Drop less meaningful columns

Check for missing, null or NA data. Input data using mean, median or mode or remove outliers if they won’t have a significant effect on the data.

Create dummy variables to change categorical data to numerical data.

Create univariate and bivariate visualization

Extract prepared dataset as CSV file

Code is attached\*

D1) There’s an increase in pseudo–R as it went from 0.4473 to 0.5296 as the categorical dummy variables were added. This shows an explanation of variance is within the categorical data.

The initial multiple regression model:

Y = -5.8583 – 0.0395 \* children + -0.0069 \* age + 1.199e-07 \* Income – 0.0020 \* Outage\_sec\_perweek – 0.0015 \* Email + 0.0301 \* Contacts – 0.0308 \* Yearly\_equip\_failure + 0.7956 \* DummyTechiw – 2.295 \* DummyContract + 0.161 \* DummyPort\_modem – 0.0796 \* DummyTablet – 1.4252 \* DummyInternetService – 0.3157 \* DummyPhone – 0.2908 \* DummyMultiple – 0.3461 \*DummyOnlineSecurity – 0.5125 \* DummyStreamingTV + 0.1126 \* DummyPaperlessBilling – 0.2043 \* Tenure + 0.0461 \* MonthlyCharge + 0.0013 \* Bandwidth\_GB\_Year – 0.0167 \* TimelyResponse + 0.0143 \* Fixes – 0.0158 \* Replacements – 0.025 \* Reliability – 0.0341 \* Options – 0.0309 \* Respectfulness + 0.0047 \* Courteous – 0.009 \* Listening

D2) Based on the model above with a pseudo-R-value of 0.5296 we have coefficients with an insignificant p-value at the 0.05 level. Those variables will be removed and variables with a significant p-value will be selected to run against the DummyChurn variable

D3) See attached coding

E1) The reduced model explains 52% of variance. Most of the dummy variables used have a negative value. This shows that with each additional service a customer adds that rate at which a customer will churn is reduced by those variables respected amount.

In the best interest of the company the client should focus on providing more services and making it clear what services are available to the consumer.

E2) See attachment

E3) See attachment

F1) Results:

Y= -6.1973 – 0.0391 \* Children + 0.0070 \* Age + 0.7970 \* DummyTechie – 2.2895 \* DummyContract + 0.1598 \* DummyPort\_modem – 1.4240 \* DummyInternetService – 0.3193 \* DummyPhone - 0.2964 \* DummyMultiple – 0.3555 \* DummyTechSupport – 0.2049 \* Tenure + 0.0463 \* MonthlyCharge + 0.0013 \* Bandwidth\_GB\_Year

This analysis resulted in several inverse relationships in the predictor values such as Children and DummyContract. This is important to identify because with an increase in those negatively correlated variables customers are less likely to leave the company. DummyContract and DummyInternetService shows the strongest negative relationship while Age shows the largest positive correlation at 0.7970.

The coefficients suggest that for every 1 unit of:

Children – Churn will decrease by 0.0391

Age – churn will increase 0.0070

Techie - churn will increase 0.7970

Contract- Churn will decrease by 2.2895

Port\_modem - churn will increase 0.1598

InternetService - Churn will decrease by 1.4240

Phone - Churn will decrease by 0.3193

Multiple - Churn will decrease by 0.2964

TechSupport - Churn will decrease by 0.3555

Tenure - Churn will decrease by 0.2049

MonthlyCharge - churn will increase 0.0463

Bandwidth\_GB\_Year - churn will increase 0.0013

The limitation of this study is the assumption of linearity between the dependent variable and the independent variables. Next, this model is sensitive to outliers so those must be removed or changed to not disrupt the results of the model. Finally, the data set is a bit small. We can make decisions based on what we were given, but a larger data set collected over several months will allow this model to create a better image of what best to focus on to reduce the churn rate.

F2) The model above shows an inverse relationship between churn and several of the predictor values. The recommended course of action based off this knowledge is to retain their current customer base by offering more services and making it clear to the consumer what services are available to them.

Additionally, it’s recommended to do everything within their power to retain customers gained by promptly fixing customers problems and ensuring the equipment used is of a high quality to reduce the amount of equipment replacements.

G) Video

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=87174d04-42c9-45b6-ae52-ae98010072af>

I) Citations

*Assumptions of logistic regression*. Statistics Solutions. (2021, August 11). Retrieved May 2, 2022, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

Navlani, A. (2019, December 16). *Python logistic regression tutorial with Sklearn & Scikit*. DataCamp Community. Retrieved May 2, 2022, from https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python

Qureshi, A. (2020, October 13). *The importance of R in data science*. Medium. Retrieved May 2, 2022, from https://towardsdatascience.com/the-importance-of-r-in-data-science-6b394d48fa50#:~:text=Put%20simply%2C%20it%20is%20Pearson's%20correlation%20coefficient%20(r).,1%20and%20a%20%2B1).

Swaminathan, S. (2019, January 18). *Logistic regression - detailed overview*. Medium. Retrieved May 2, 2022, from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

*What is logistic regression?* IBM. (n.d.). Retrieved May 2, 2022, from https://www.ibm.com/topics/logistic-regression