**D208 Task 2: Logistic Regression**

Cali Ciccarelli

Dr. Straw

07/07/24

**Part One**

**Research Question:**

What variables make a customer more or less likely to churn?

**Goals:**

This analysis aims to identify which variables are best at predicting customer churn. Are some variables better at predicting customer retention? This is vital information that will help stakeholders make essential decisions for marketing, advertisements, and more. Keeping customers is highly important and profitable so the analysis will benefit the company.

**Part Two**

**Assumptions:**

* The dependent variable must be binary. This means the dependent variable can only have two outcomes. I will use “Churn,” which has a yes/no result. This is a binary variable.
* The observations cannot be related. This means the data records can not be dependent on one another/can’t be repeated. This can be verified by reviewing a residual plot and ensuring there are no patterns.
* Amongst the independent variables, there should be no multicollinearity. This means the independent variables should not be correlated with one another.
* There should be no extreme outliers. Extreme outliers should be removed or replaced.
* There must be a linear relationship between the independent variables and the logit of the dependent variable. The relationship should not be “bell” shaped.
* Sample size must be large. Predictions cannot be accurately made with a small sample. Our data has 10,000 records, so this is certainly sufficient.

(Bobbit, 2020)

**Benefits of Python:**

“Data analysis with Python is much more efficient. It allows data to be cleaned on a wide scale using a code applicable to the data set. It also allows for quick and accurate data analysis using the Python packages for statistical calculations and graph/model creations. The code can also be set to automate and analyze as updated data comes in.” (Ciccarelli, 2024) For logistic regression in particular, python offers packages that allow the regression to be completed with a few lines of code.

**Method Purpose:**

Logistic regression is most appropriate for my research question regarding what variables impact the churn variable. Logistic regression requires the dependent variable to be binary, which aligns with the churn variable. Linear regression would not work for this specific analysis since churn is not a continuous variable. With logistic regression, we can see what variables increase the odds of predicting whether or not a customer churns.

**Part Three**

**Data Cleaning:**

First, I checked for duplicated records. This returned none. Next, I went on to check for missing values. I found that ‘InternetService’ was the only variable with n/a values. It is likely that these n/a values were meant to represent that the customer did not have internet service with the provider. With this assumption, I filled the n/a values with a new " None " category. I then dropped several columns that likely will not affect the Churn rate so that I can better focus on the variables that may have an effect. I also checked for outliers. While I did find several, they seemed to be within reason. I decided to leave them in the data set because I felt they were valid and not extreme, so they should not negatively impact the analysis. Last, I checked the unique values for each categorical variable. This helped me see what outputs there were for each categorical variable. I ensured there were no typos/errors or similar outputs with slightly different names.

**Summary Statistics:**

Below, I have attached the output for the summary statistics. The statistics show the mean, min, max, std, and quartile values for the numerical values. The most common response and the number of unique variables are shown for the categorical variables.

count 10000.0000

mean 2.0877

std 2.1472

min 0.0000

25% 0.0000

50% 1.0000

75% 3.0000

max 10.0000

Name: Children, dtype: float64

count 10000.000000

mean 53.078400

std 20.698882

min 18.000000

25% 35.000000

50% 53.000000

75% 71.000000

max 89.000000

Name: Age, dtype: float64

count 10000

unique 5

top Divorced

freq 2092

Name: Marital, dtype: object

count 10000

unique 3

top Female

freq 5025

Name: Gender, dtype: object

count 10000

unique 2

top No

freq 7350

Name: Churn, dtype: object

count 10000.000000

mean 10.001848

std 2.976019

min 0.099747

25% 8.018214

50% 10.018560

75% 11.969485

max 21.207230

Name: Outage\_sec\_perweek, dtype: float64

count 10000.000000

mean 12.016000

std 3.025898

min 1.000000

25% 10.000000

50% 12.000000

75% 14.000000

max 23.000000

Name: Email, dtype: float64

count 10000.000000

mean 0.994200

std 0.988466

min 0.000000

25% 0.000000

50% 1.000000

75% 2.000000

max 7.000000

Name: Contacts, dtype: float64

count 10000.000000

mean 0.398000

std 0.635953

min 0.000000

25% 0.000000

50% 0.000000

75% 1.000000

max 6.000000

Name: Yearly\_equip\_failure, dtype: float64

count 10000

unique 3

top Month-to-month

freq 5456

Name: Contract, dtype: object

count 10000

unique 2

top No

freq 5166

Name: Port\_modem, dtype: object

count 10000

unique 2

top No

freq 7009

Name: Tablet, dtype: object

count 10000

unique 3

top Fiber Optic

freq 4408

Name: InternetService, dtype: object

count 10000

unique 2

top Yes

freq 9067

Name: Phone, dtype: object

count 10000

unique 2

top No

freq 5392

Name: Multiple, dtype: object

count 10000

unique 2

top No

freq 6424

Name: OnlineSecurity, dtype: object

count 10000

unique 2

top No

freq 5494

Name: OnlineBackup, dtype: object

count 10000

unique 2

top No

freq 5614

Name: DeviceProtection, dtype: object

count 10000

unique 2

top No

freq 6250

Name: TechSupport, dtype: object

count 10000

unique 2

top No

freq 5071

Name: StreamingTV, dtype: object

count 10000

unique 2

top No

freq 5110

Name: StreamingMovies, dtype: object

count 10000.000000

mean 34.526188

std 26.443063

min 1.000259

25% 7.917694

50% 35.430507

75% 61.479795

max 71.999280

Name: Tenure, dtype: float64

count 10000.000000

mean 3392.341550

std 2185.294852

min 155.506715

25% 1236.470827

50% 3279.536903

75% 5586.141370

max 7158.981530

Name: Bandwidth\_GB\_Year, dtype: float64

count 10000.000000

mean 172.624816

std 42.943094

min 79.978860

25% 139.979239

50% 167.484700

75% 200.734725

max 290.160419

Name: MonthlyCharge, dtype: float64

count 10000.000000

mean 3.490800

std 1.037797

min 1.000000

25% 3.000000

50% 3.000000

75% 4.000000

max 7.000000

75% 4.000000

max 8.000000

**Univariate and Bivariate visuals** are in the pdf file of the code. The output is saved with the code itself. This includes the univariate visuals of all variables used in the initial regression model and the bivariate visuals of each independent variable with ‘Churn.’

**Data Transformation:**

I transformed all categorical data into numerical data to best support the success of my model. I used one hot encoding to create dummy variables for my categorical variables that cannot be ranked nominally, such as “Marital” and “Gender.” The dummy variables will tell me whether or not the category is present for the customer. 0 represents a variable that is not present, and 1 represents a present variable. I also replaced the yes/no variables with 1s and 0s. I found this necessary for my model to interpret the binary variables correctly.

**CSV File, see attached.**

**Part Four**

**Initial Logistic Model:**

If you find the screenshot quality poor, please view the model on the attached PDF of the code and output.

A screenshot of a document

Description automatically generated

**Model Evaluation Metric:**

I used variance inflation factors to search for multicollinearity amongst the explanatory variables. VIF values over 5 signify a high correlation with other variables (Bobbit, 2020). I ran the VIF test several times and removed variables one by one to avoid excessive data loss. With this method, I ended up removing “Bandwidth\_GB\_Year”, “MonthlyCharge”, “Email”, “Outage\_sec\_perweek”, “Phone”, and “Age” due to multicollinearity. Next, I used the model's P values to eliminate variables. Again, I eliminated variables one by one. I looked for independent variables with a P value over .05. This would show that the independent variable is not significantly significant for detecting churn.

Please see the attached code for the output. This includes the VIF tests and the elimination logit models.

**Reduced Logistic Model:**

A screenshot of a computer

Description automatically generated

**Part Five:**

**Compare the models:**

Looking at the “Psuedo R Squared” value, it appears that the initial model is a slightly better fit than the reduced. This is unlikely due to the amount of multicollinearity in the first model. The multicollinearity causes the initial model to be invalid. The reduced model is more successful in predicting churn. Since the r-squared value is closer to 1 than 0, the model is a good fit (Mikko, 2019). The model predicts the churn variable well, but it is not a perfect result.

**Confusion Matrix:**

A screenshot of a graph

Description automatically generated

**Accuracy Calculation:**

A screenshot of a computer

Description automatically generated

Find the executable code in the attached file.

**Part Six:**

**Regression Equation:** ln(p/1-p) = -1.45 + 1.07(Techie) + 1.59(Multiple) + .44(DeviceProtection) + .26(TechSupport) + 2.87(StreamingTV) + 3.37(StreamingMovies) +.15(PaperlessBilling) -.10(Tenure) +.24(dmy\_Male) -3.24(dmy\_One\_year) -3.3(dmy\_Two\_year) -1.32(dmy\_Fiber\_Optic) -1.39(dmy\_None)

**Interpret Coefficients:**

* The log odds for Churn increase by 1.07 when a customer answers yes to “Techie”.

Using the calculation e^.03, we find that the rate a customer will churn increases by 1.07 when the customer is found to be a “Techie”.

* The log odds for Churn increase by 1.59 when a customer does have Multiple lines.
* The log odds for churn increase by .44 when a customer has device protection.
* The log odds for churn increase by .26 when a customer has tech support.
* The log odds for churn increase by 2.87 when a customer does stream TV.
* The log odds for churn increase by 3.37 when a customer does stream movies.
* The log odds for churn increase by .15 when a customer participates in paperless billing.
* The log odds for churn decrease by .10 for every unit (month) a customer stays with the company. (tenure)
* The log odds for churn increase by .24 if the customer is male.
* The log odds for churn decrease by 3.24 if a customer does have a one-year contract.
* The log odds for churn decrease by 3.3 if the customer does have a two-year contract.
* The log odds for churn decrease by 1.32 if a customer does have fiber optic.
* The log odds for churn decrease by 1.39 if a customer does not have internet service with the company.

**Statistical/Practical significance:**

The results interpret that the chances a customer will churn increases when a customer has the following: Multiple, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, PaperlessBilling, dmy\_male (customer is a male). The chances a customer will churn decreases when a customer has the following: dmy\_One\_year, dmy\_Two\_year, dmy\_Fiber\_optic, dmy\_none. The chance of churn also decreases each month a customer stays with the company. The R squared value supports that the model will be accurate more than half the time. The P values show that the variables left in the model are statistically significant to churn since they are all under .05.

**Limitations:**

Limitations of this model include not being able to determine causation. While I did remove variables to eliminate multicollinearity within the model, the causation of some of the variables may be more attributed to another variable rather than Churn. Particularly, customers could be leaving due to high costs, which could be why the streaming variable options can predict churn. The model should be looked at only as something that can predict whether or not a customer will churn. Unfortunately, the model is not able to predict Churn perfectly. The R-squared value is good but could be closer to one to signify a more substantial model.

**Recommendations:**

While we cannot prove causation, the results are statistically significant enough to begin playing around with the remaining variables. They could add incentives to stay with the company for a more extended period since increased tenure appears to help avoid churn. The company could look to push fiber optic internet since this seems to have better customer retention. They should continue collecting data after these changes to see if customers are less likely to churn. Another recommendation would be to start collecting data from those who churned and have them list the reason(s) they churned. If you give them a multiple-choice option, this data could be incredibly beneficial in helping predict churn better.

Bobbitt, Z. (2020, October 13). *The 6 Assumptions of Logistic Regression (With Examples)*. Statology. <https://www.statology.org/assumptions-of-logistic-regression/>

Ciccarelli, C. (2024, July 4). *D208 Task One: Linear Regression)*.

Martin, M. (2021, January 12). *5.7 Logistic Regression: Interpreting model coefficients* [Video]. YouTube. https://www.youtube.com/watch?v=tfJUbCGELCQ

Mikko R. (2019, August 23). *Pseudo-R2 in logistic regression* [Video]. YouTube. <https://www.youtube.com/watch?v=zGdQ8fbl6j4>