**D208 Logistic Regression**

**Nhi Le**

**Western Governors University**

**August 24, 2023**

**Part I: Research Question**

**A1.** **Research Question**

What are the features that significantly impact churn?

**A2.** **Goals**

From this analysis, the stakeholders can understand what factors affect churn to see where the services can be improved. Moreover, they will be able to predict which customers have high risks of retention. Then, the customer support team can contact those customers and have strategies to improve their experience with the company.

**Part II: Method Justification**

**B1.** **Summary of assumptions**

The four assumptions of a logistic regression model:

* The dependent variable is binary, for example Yes or No, True or False.
* Among explanatory variables, they are not too highly correlated with one another.
* There are no extreme outliers or influential observations in the data set.
* The sample size needs to be sufficiently large (Zach, 2020)

**B2.** **Tool benefits**

In this analysis, I will be using Python. Its benefits are: Firstly, Python has a lot of libraries and packages that can help me to create my predictive model with simple syntax. I can use Python from cleaning data using NumPy and Pandas libraries, to visualizing data using Matplotlib and Seaborn, to machine learning with Scikit-learn (Pruciak, n.d.). Secondly, Python has been around for a long time and has a global IT community. I can find support and solutions quickly for my codes if I need it, which helps me to complete this analysis effectively and accurately.

**B3.** **Appropriate technique**

Logistic regression is an appropriate technique because the research question is for Churn. Churn is a binomial variable, with Yes and No values. By using logistic regression, we can use multiple independent variables to predict the probability of churn. This technique helps us to understand which features when we add or remove will affect the likelihood of churning. Therefore, through the model, the stakeholders can see if an independent variable has positive or negative relationship with the probability of churn, so they can have plans to improve the services to reduce retention.

**Part III: Data Preparation**

**C1.** **Data cleaning goals**

My data cleaning goals are to detect missing data, duplicate data, and outliers, then decide to treat them with appropriate methods:

* Import dataset churn\_clean.csv into Jupyter Notebook.
* Get information (column names, data types), and statistical details (count, min, max, mean, std, percentile) of the dataset.
* Detect duplicates and delete the duplicated records if there are any.
* Find missing data and impute missing data with meaningful measures of central tendency (mean, median, or mode).
* Find outliers and treat them by removing them, retaining them, excluding them, or imputing them with the median.
* Rename the Item1 – Item8 columns to easily recognized names (For example: ‘Item1’ renamed to ‘TimelyResponse’)

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

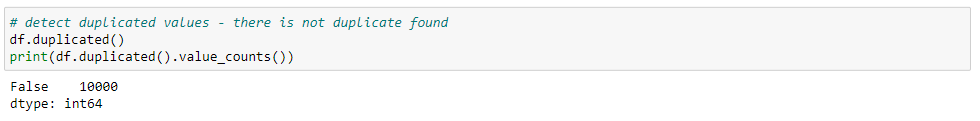
Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated



A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer program

Description automatically generated

**C2.** **Summary statistics**

The dependent variable used to answer the research question was ‘Churn’. It is a categorical variable as the values are ‘Yes’ or ‘No’. It recorded whether the customers canceled service last month. There were 10,000 records in total with 7350 records for ‘No’.

The dependent variables used to answer the research question all have 10,000 records:

* The 10 continuous variables including:
* ‘Children’, ‘Age’, ‘Income’ are demographic variables on billing statement for each customer. The average number of children that customers had was 2. The age range of customers was 18 to 89. The average income was 39806.93.
* ‘Email’ is numeral variable to record the number of marketing or correspondence emails sent. On average, customers received 12 emails, and the maximum emails customers got was 23. ‘Contacts’ is numeral variable for how many times customer contacted technical support. Customers contacted support 1 time on average.
* ‘Outage\_sec\_perweek’ shows system outages in the customer’s neighborhood’s average of seconds per week. The longest time that the customers experienced system outages was 21 seconds per week.
* ‘Bandwidth\_GB\_Year’ (the average yearly amount of data used, in GB, per customer) is a continuous variable. This variable has the mean of average of data usage of a year was 3392.34 GB.
* ‘Tenure’ is numerical variable to record how many months the customer has been with the provider. The customers have used our services from 1 to 72 months. ‘MonthlyCharge’ is the monthly charge for the customer. Customers paid us 172.62 per month. ‘Yearly\_equip\_failure’ is numeral variable to show the number of time customer’s equipment failed and needed to reset or replaced last year. This did not happen a lot, but some customers had to experience this up to 6 times.
* The 23 categorical variables including:
* 8 categorical variables reflect customer’s satisfaction ratings on a scale of 1 to 8 (1 = most important, 8 = least important): ‘item1’ – Timely response, ‘item2’ – Timely fixes, ‘item3’ – Timely replacements, ‘item4’ – Reliability, ‘item5’ – Options, ‘item6’ – Respectful response, ‘item7’ – Courteous exchange, ‘Item8’ – Evidence of active listening. The values are from 1 to 8. The mean of each variable is around 3.50.
* ‘Gender’ is categorical variable to reflect the gender of customer. There are 3 unique values for gender.
* ‘Techie’ has Yes/No value. This categorical variable reflects if the customer thinks that they are good at technology. ‘No’ was the top value.
* ‘Contract’ is categorical variable on what kind of contract customer has ‘Month-to-month’, ‘One Year’, or ‘Two Year’. 5456 customers had Month-to-month contracts.
* ‘Tablet’ is categorical variable answering if the customer has a tablet. ‘Port\_modem’ is categorical variable answering if the customer has a portable modem. The values are ‘Yes’ or ‘No’. The top values for these 2 variables were ‘No’.
* ‘InternetService’ shows customer’s internet service provider. There were 3 unique values.
* ‘Phone’, ‘Multiple’, ‘OnlineSecurity’, ‘OnlineBackup’, ‘DeviceProtection’, ‘TechSupport’, ‘StreamingTV’, ‘StreamingMovies’, ‘PaperlessBilling’ are services that the company provides. The values of these variables are ‘Yes’ or ‘No’ to reflect if the customer signed up for. Top values for ‘Phone’ and ‘PaperlessBilling’ were ‘Yes’, the rest had the top values as ‘No’.

Here are the screenshots of the summary statistics output of the variables (including: count, mean, std, min, 25%, 50%, 75%, max):

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a screen

Description automatically generated

**C3.** **Visualizations**

Univariate visualizations:

A screenshot of a computer screen

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

A graph with a bar and a line

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

Bivariate visualizations:

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**C4.** **Data transformation**

My data transformation goals are to make sure I can use all the independent variables for the analysis. To make that happen, I will need to encode categorical values into numerical values since I can only use numerical values in the logistic regression.

* Drop variables that will not be needed for the analysis.
* Create dummy variables for categorical variables.
* Encode categorical values to numerical values: For those variables with Yes/No values, the dummy value is 1 for Yes and 0 for No. For the Gender variable, it has Male, Female, and Nonbinary. The DummyFemale is 1 when Gender is Female and else it is 0. Contract has 3 values: Month-to-month, One Year, and Two year. DummyMonthtoMonth is 1 when Contract is Month-to-month, else it is 0. InternetService has 3 values: Fiber Optic, DSL, and None. DummyFiberOptic is 1 when InternetService is Fiber Optic, else it is 0.
* Spot-check the statistical details of the dataset to make sure categorical values are encoded correctly.
* Drop those categorical values from the data set.
* Extract the prepared dataset as CSV file named ‘churn\_prepared’.

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated



**C5.** **Prepared data set**

The prepared data set will be submitted as ‘churn\_prepared2.csv’ along with this doc file.

**Part IV: Model Comparison and Analysis**

**D1.** **Initial model**

# Create the initial logistic model

log\_reg = smf.logit('DummyChurn ~ Children+Age+Income+Outage\_sec\_perweek+Email+Contacts+Yearly\_equip\_failure+Tenure+MonthlyCharge+Bandwidth\_GB\_Year+TimelyResponse+TimelyFixes+TimelyReplacements+Reliability+Options+Respectfulness+Courteous+Listening+DummyFemale+DummyTechie+DummyPort\_modem+DummyTablet+DummyPhone+DummyMultiple+DummyOnlineSecurity+DummyOnlineBackup+DummyDeviceProtection+DummyTechSupport+DummyStreamingTV+DummyPaperlessBilling+DummyMonthtoMonth+DummyFiberOptic+DummyStreamingMovies', data=df).fit()

log\_reg.summary()

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

DummyChurn = -8.55 + -0.06\*Children + 0.01\*Age + 0.00\*Income + -0.002\*Outage\_sec\_perweek + -0.01\*Email + 0.06\*Contacts + -0.04\*Yearly\_equip\_failure + -0.32\*Tenure + 0.03\*MonthlyCharge + 0.003\*Bandwidth\_GB\_Year + -0.03\*TimelyResponse+ 0.003\*TimelyFixes + 0.03\*TimelyReplacements + -0.03\*Reliability + -0.03\*Options + -0.02\*Respectfulness + 0.004\*Courteous + -0.01\*Listening + -0.10\*DummyFemale + 1.09\*DummyTechie + 0.14\*DummyPort\_modem + -0.05\*DummyTablet + -0.28\*DummyPhone + 0.40\*DummyMultiple + -0.41\*DummyOnlineSecurity + -0.17\*DummyOnlineBackup + -0.19\*DummyDeviceProtection + -0.15\*DummyTechSupport + 0.93\*DummyStreamingTV + 0.16\*DummyPaperlessBilling + 3.43\*DummyMonthtoMonth + -1.02\*DummyFiberOptic + 1.18\*DummyStreamingMovies

**D2.** **Justification of model reduction**

The feature selection method I used to create the reduced model was backward stepwise elimination. I removed the least significant independent variables that had p-values greater than 0.05. Then I needed to verify if my model met logistic regression assumptions. One of the assumptions I thought was important to check was multicollinearity, by calculating VIF (Variance Inflation Factor). I needed to decide whether to keep or remove features with high VIFs. Finally, I created the reduced model making sure they have good Pseudo R-squared value and P-value for each feature is less than 0.05.

# create log\_reg1 without the features having P-values more than 0.05

log\_reg1 = smf.logit('DummyChurn ~ Children+Age+Tenure+MonthlyCharge+Bandwidth\_GB\_Year+DummyTechie+DummyPhone+DummyMultiple+DummyOnlineSecurity+DummyDeviceProtection+DummyStreamingTV+DummyPaperlessBilling+DummyMonthtoMonth+DummyFiberOptic+DummyStreamingMovies', data=df).fit()

log\_reg1.summary()

A screenshot of a computer screen

Description automatically generated

# After creating log\_reg1, I still saw DummyDeviceProtection with p-value > 0.05. So I created log\_reg2 without it.

# DummyPhone and DummyPaperlessBilling having p-values < 0.05 but still pretty close to 0.05

# I decided to exclude them from the model

log\_reg2 = smf.logit('DummyChurn ~ Children + Age+Tenure + MonthlyCharge+Bandwidth\_GB\_Year+DummyTechie+DummyMultiple+DummyOnlineSecurity+DummyStreamingTV+DummyMonthtoMonth+DummyFiberOptic+DummyStreamingMovies', data=df).fit()

log\_reg2.summary()

A screenshot of a data

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# Even MonthlyCharge has high VIF, I want to keep it in my model since it has low P-value in the model

# Without it, the Pseudo R-squared will reduce by 3.54% (From 61.84% to 58.30%)

# Create test model without MonthlyCharge

log\_reg\_test = smf.logit('DummyChurn ~ Children+Age+Tenure+DummyTechie+DummyMultiple+DummyOnlineSecurity+DummyStreamingTV+DummyMonthtoMonth+DummyFiberOptic+DummyStreamingMovies', data=df).fit()

log\_reg\_test.summary()

A screenshot of a computer

Description automatically generated

# create reduced model log\_reg4 from log\_reg3 without Bandwidth\_GB\_Year

log\_reg4 = smf.logit('DummyChurn ~ Children+Age+Tenure+MonthlyCharge+DummyTechie+DummyMultiple+DummyOnlineSecurity+DummyStreamingTV+DummyMonthtoMonth+DummyFiberOptic+DummyStreamingMovies', data=df).fit()

log\_reg4.summary()

A screenshot of a computer

Description automatically generated

**D3.** **Reduced logistic regression model**

# p-values for Children and Age > 0.05, I will remove them and create the final reduced model

# create reduced model

log\_reg = smf.logit('DummyChurn ~ Tenure+MonthlyCharge+DummyTechie+DummyMultiple+DummyOnlineSecurity+DummyStreamingTV+DummyMonthtoMonth+DummyFiberOptic+DummyStreamingMovies', data=df).fit()

log\_reg.summary()

A screenshot of a computer

Description automatically generated

DummyChurn = -8.49 + -0.11\*Tenure + 0.04\*MonthlyCharge + 1.05\*DummyTechie + 0.35\*DummyMultiple + -0.23\*DummyOnlineSecurity + 1.13\*DummyStreamingTV + 3.31\*DummyMonthtoMonth + -1.79\*DummyFiberOptic + 1.25\*DummyStreamingMovies

**E1.** **Model Comparison**

I use Pseudo R-squared to compare the initial and reduced logistic regression models. For the initial model, Pseudo R-squared value is 62.04%. For the reduced model, the Pseudo R-squared value is 61.02%. The Pseudo R-squared difference between these two models is not significant. Based on this metric, the initial model is a better fit than the reduced model. However, the initial model did not meet some of the logistic regression assumptions because it included observations with some extreme outliers and multicollinearity. After excluding more than half of the independent variables, the reduced model’s Pseudo R-squared was only reduced by 1%. Therefore, the reduced model is still a good fit.

**E2.** **Output and calculations**

The confusion matrix is that Out of 3,000 :

* True Positive + True Negative = 2042 + 627 = 2669
* False Positive + False Negative = 172 + 159 = 331

The accuracy is 88.97%. The model performs well.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

**E3.** **Code**

I attached a Jupiter Notebook file for my code as ‘D208(2).ipynb’

**Part V: Data Summary and Implications**

**F1.** **Results**

The equation for the reduced model includes 9 independent variables:

DummyChurn = -8.49 + -0.11\*Tenure + 0.04\*MonthlyCharge + 1.05\*DummyTechie + 0.35\*DummyMultiple + -0.23\*DummyOnlineSecurity + 1.13\*DummyStreamingTV + 3.31\*DummyMonthtoMonth + -1.79\*DummyFiberOptic + 1.25\*DummyStreamingMovies

The coefficients suggest that for every 1 unit of:

* Tenure (The number of months that the customers stay with our services): DummyChurn will decrease 0.11 units.
* MonthlyCharge: DummyChurn will increase 0.04 units.
* DummyTechie (Customers are techie): DummyChurn will increase 1.05 units.
* DummyMultiple (Customers have multiple lines): DummyChurn will increase 0.35 units.
* DummyOnlineSecurity (Customers have online security): DummyChurn will decrease 0.23 units.
* DummyStreamingTV (Customers have Streaming TV): DummyChurn will increase 1.13 units.
* DummyMonthtoMonth (Customers have month-to-month contracts): DummyChurn will increase 3.31 units.
* DummyFiberOptic (Customers have Internet Service with Fiber Optic): DummyChurn will decrease 1.79 units.
* DummyStreamingMovies (Customers have Streaming Movies): DummyChurn will increase 1.25 units.

P-values for all independent variables are at or close to 0.000. The p-values are very low, which proves that the reduced model is statistically significant. The model is also practically significant because the accuracy is high, at 88.97%, which means that the predicted churn results are highly accurate. Therefore, we can use this model to predict which customers will likely leave our services.

The limitation of this analysis is that the data set is small and only contains data from the last year. If we had the data in the time series for at least 3 years, we could track and find more accurate trends.

**F2.** **Recommendations**

The features that affect churn are Tenure, MonthlyCharge, DummyTechie, DummyMultiple, DummyOnlineSecurity, DummyStreamingTV, DummyMonthtoMonth, DummyFiberOptic, and DummyStreamingMovies. Firstly, I see that the longer the customers stay with our service, the less likely retention will happen. However, higher monthly charges can lightly increase churn. Therefore, to reduce churn, the company should reduce costs for customers when they are loyal customers. We can give customers one time or recurring discounts after a specific amount of time. Sales and marketing can look further into this and have discounts/ pricing strategies. Secondly, the more services customers have, the higher likelihood of churn can happen. This might suggest that our services do not work well together, or we do not provide sufficient data capacity for customers to use the services smoothly. We need to investigate this and if it is because of data capacity, we will need to create data plans for customers to reduce churn. Lastly, customers with month-to-month contract increase the probability of retention. Therefore, selling customers one-year or two-year contracts can help to keep customers with us.

**Part VI: Demonstration**

**G. Panopto demonstration**

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=eaebf8f5-6f9e-44c6-bf3c-b06900206eed>

**H. Sources of third-party code**

*Logistic regression in python with statsmodels*. Andrew Villazon. (n.d.). https://www.andrewvillazon.com/logistic-regression-python-statsmodels/

GeeksforGeeks. (2023, January 10). *Detecting multicollinearity with VIF - python*. GeeksforGeeks. https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/

Zach. (2022, June 8). *How to perform logistic regression in Python (STEP-by-step)*. Statology. https://www.statology.org/logistic-regression-python/

GeeksforGeeks. (2023b, March 20). *Logistic regression using python*. GeeksforGeeks. https://www.geeksforgeeks.org/ml-logistic-regression-using-python/

**I. Sources**

Zach. (2020, October 13). *The 6 assumptions of logistic regression (with examples)*. Statology. https://www.statology.org/assumptions-of-logistic-regression/