D208 – Predictive Modeling (Task 2)

Morrell J. Parrish

Western Governors University

Table of Contents

[A. Research Question 3](#_Toc108117628)

[A2. Objective or Goals 3](#_Toc108117629)

[B. Assumptions Summary 3](#_Toc108117630)

[B2. Benefits Of Chosen Analytical Tool(s) 3](#_Toc108117631)

[B3. Chosen Technique Explanation 4](#_Toc108117632)

[C. Data Preparation Description 4](#_Toc108117633)

[C2. Summary of Statistics 4](#_Toc108117634)

[C3. Data Preparation Steps 5](#_Toc108117635)

[C4. Univariate and Bivariate Visualizations 8](#_Toc108117636)

[C5. Churn Data Set 12](#_Toc108117637)

[D. Model Comparison and Analysis 12](#_Toc108117638)

[D2. Justification of Based Variable Selection Procedure and Model Evaluation Metric 15](#_Toc108117639)

[D3. Logistic Regression Model 16](#_Toc108117640)

[E. Data Set Analyzation 18](#_Toc108117641)

[E3. Logistic Regression Model Code 19](#_Toc108117642)

[F. Summary 19](#_Toc108117643)

[F2. Recommended Course of Action 20](#_Toc108117644)

[G. Panopto video recording 20](#_Toc108117645)

[References 21](#_Toc108117646)

D208 – Predictive Modeling (Task 2)

# A. Research Question

During this course of research, we will determine which variables are indicators or indicates which customers are more likely to “churn” or terminate their services?

## A2. Objective or Goals

The objective of this analysis is to perform logistic regression on our churn dataset and define which variable(s) within our churn dataset can be indicators for churn. “The churn rate, also known as the rate of attrition or customer churn; is the frequency in which consumers discontinue doing business with a company. It is commonly represented as the percentage of service subscribers who cancel their memberships within a specified time frame” (Frankenfield, 2022). Defining which variables are indicators businesses can focus on improving in those areas and decrease their churn rate.

# B. Assumptions Summary

“A logistic regression model is used to estimate the relationship between a dependent variable and one or more independent variables, but it is used to make a prediction about a categorical variable versus a continuous one” (Lawton, 2022).  A logistic regression model makes the following assumptions:

* The response variable is ***binary*** *- (takes on 2 possible outcomes)*
* The observations are independent
* There is no multicollinearity among the explanatory variables
* There are no extreme outliners
* There is a linear relationship between explanatory variables and the logit of the response variable
* The sample size is sufficiently large

## B2. Benefits Of Chosen Analytical Tool(s)

The chosen analytical tool for this analysis will be *Python (PyCharm)*. Both *Python* and *R* have strengths and weaknesses; the dataset used in this analysis contains 10000 observations and 50 variables. Both *R* and *Python* have packages/libraries which allow you to cleanse, manage, transform, and perform analysis and statistics. Another reason we will be using *Python*

is because it’s simple and has a very versatile programming style.

## B3. Chosen Technique Explanation

A logistic regression model will allow us to perform regression analysis on a dependent variable that has binary characteristics, and in this case that would be our dependent variable “churn”; also, a logistic regression model will allow us to add or remove variables, by doing this this will help determine if they have a positive or negative impact on predicting if a customer churns or stays.

# C. Data Preparation Description

To use the churn dataset in our analysis we will first need to prepare the data.

The following steps were taken to prepare the dataset for analysis:

* download the churn dataset
* determine which variables will be used in the analysis
* import the dataset into *PyCharm*
* remove independent variables, demographics, and personal identification variables not being used in the analysis
  + caseorder, customer\_id, interaction, UID, city, state, county, zip, lat, lng, population, timezone, job, email, contacts
* determine if any outliners exist and remove them

## C2. Summary of Statistics

There are 23 continuous variables, 27 categorical variables and 10000 observations; however, upon further review of the statistical summary some abnormalities can be noted in the following areas children, outage sec perweek, yearly equip failure, tenure, income, monthly charge, and bandwidth gb year. For the purpose of this study we will not use all 50 variables; the *(18)* categorical variables we will use during this analysis are: area, marital, gender, churn, techie, contract, portmodem, tablet, internetservice, phone, multiple, onlinesecurity, onlinebackup, deviceprotection, techsupport, streamingtv, streamingmovies, paperlessbilling, paymentmethod; the *(16)* continuous variables are children, age, income, outage sec per week, yearly equipment failure, tenure, monthly charge, and bandwidth gb year, item1 (timelyresponse), item2 (fixes), item3 (replacements), item4 (reliability), item5 (options), item6 (respectfulness), item7 (courteous), and item8 (listening). The below figure is the summary of statistics for the continuous variables.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Column1 | **Count** | **Mean** | **STD** | **Min** | **25%** | **50%** | **75%** | **Max** |
| **Children** | 10000 | 2.0877 | 2.147200446 | 0 | 0 | 1 | 3 | 10 |
| **Age** | 10000 | 53.0784 | 20.69888156 | 18 | 35 | 53 | 71 | 89 |
| **Income** | 10000 | 39806.92677 | 28199.9167 | 348.67 | 19224.7175 | 33170.605 | 53246.17 | 258900.7 |
| **Outage\_sec\_perweek** | 10000 | 10.00184816 | 2.976019188 | 0.09974694 | 8.018214 | 10.01856 | 11.969485 | 21.20723 |
| **Yearly\_equip\_failure** | 10000 | 0.398 | 0.635953177 | 0 | 0 | 0 | 1 | 6 |
| **Tenure** | 10000 | 34.52618809 | 26.44306263 | 1.00025934 | 7.917693592 | 35.430507 | 61.479795 | 71.99928 |
| **MonthlyCharge** | 10000 | 172.6248162 | 42.94309411 | 79.97886 | 139.979239 | 167.4847 | 200.734725 | 290.160419 |
| **Bandwidth\_GB\_Year** | 10000 | 3392.34155 | 2185.294852 | 155.5067148 | 1236.470827 | 3279.536903 | 5586.14137 | 7158.98153 |
| **TimelyResponse** | 10000 | 3.4908 | 1.037797216 | 1 | 3 | 3 | 4 | 7 |
| **Fixes** | 10000 | 3.5051 | 1.034640536 | 1 | 3 | 4 | 4 | 7 |
| **Replacements** | 10000 | 3.487 | 1.027976981 | 1 | 3 | 3 | 4 | 8 |
| **Reliability** | 10000 | 3.4975 | 1.025816251 | 1 | 3 | 3 | 4 | 7 |
| **Options** | 10000 | 3.4929 | 1.024819309 | 1 | 3 | 3 | 4 | 7 |
| **Respectfulness** | 10000 | 3.4973 | 1.033585768 | 1 | 3 | 3 | 4 | 8 |
| **Courteous** | 10000 | 3.5095 | 1.028501595 | 1 | 3 | 4 | 4 | 7 |
| **Listening** | 10000 | 3.4956 | 1.028633292 | 1 | 3 | 3 | 4 | 8 |

Figure 1: Summary of Statistics

## C3. Data Preparation Steps

To use the churn dataset in our analysis we will first need to prepare the data:

* import the dataset into *Python (PyCharm)*
* view the dataframe’s description, structure, and data types
* view summary statistics
* evaluate the dataset, remove null or missing values
* remove any outliners
* remove demographics, and personal identification
  + caseorder, customer\_id, interaction, UID, city, state, county, zip, lat, lng, population, area, timezone, job, email, contacts
* convert binomial variables (yes/no to 1 and 0) to numerical variables
* view univariate and bivariate visuals

The following code below was used to prepare our data:

*# Load data set into Pandas dataframe*df = pd.read\_csv('churn\_clean.csv')  
  
*# Remove less meaningful demographic variables*df = df.drop(columns=['CaseOrder', 'Customer\_id', 'Interaction', 'UID', 'City', 'State', 'County', 'Zip', 'Lat', 'Lng', 'Population', 'TimeZone',  
'Email', 'Contacts', 'Job'])  
  
*# Display Churn dataframe*print(df)  
  
*# Rename last 8 columns*df.rename(columns={'Item1': 'TimelyResponse', 'Item2': 'Fixes', 'Item3': 'Replacements', 'Item4': 'Reliability','Item5': 'Options', 'Item6': 'Respectfulness', 'Item7': 'Courteous', 'Item8': 'Listening'}, inplace=True)  
  
*# Get column info*print(df.info())  
  
*# Describe Churn dataset*print(df.describe())  
  
*# Save stats summary to excel*df.describe().to\_excel('summary\_stat.xlsx', index=False)  
  
*# Create Seaborn boxplots for continuous variables*fig2, axs = plt.subplots(4, 2, figsize=(9, 9))  
  
sns.boxplot(y='Churn', x='Children', data=df, color="orange", ax=axs[0, 0])  
sns.boxplot(y='Churn', x='Age', data=df, color="gold", ax=axs[0, 1])  
sns.boxplot(y='Churn', x='Outage\_sec\_perweek', data=df, color="turquoise", ax=axs[1, 0])  
sns.boxplot(y='Churn', x='Yearly\_equip\_failure', data=df, color="red", ax=axs[1, 1])  
sns.boxplot(y='Churn', x='Tenure', data=df, color="pink", ax=axs[2, 0])  
sns.boxplot(y='Churn', x='Income', data=df, color="silver", ax=axs[2, 1])  
sns.boxplot(y='Churn', x='MonthlyCharge', data=df, color="green", ax=axs[3, 0])  
sns.boxplot(y='Churn', x='Bandwidth\_GB\_Year', data=df, color="darkblue", ax=axs[3, 1])  
  
plt.tight\_layout(pad=0.4, w\_pad=0.5, h\_pad=1.0)  
  
plt.show()  
  
*# Create Churn Visualizations with categorical variables*df3 = df[['Area', 'Gender', 'Techie', 'Port\_modem', 'OnlineSecurity', 'OnlineBackup', 'Marital', 'Contract', 'Tablet', 'InternetService', 'TechSupport', 'PaperlessBilling', 'StreamingTV', 'Phone', 'Multiple', 'StreamingMovies', 'PaymentMethod', 'Churn']]  
  
for index, category in enumerate(df3):  
 plt.subplots(1, 1, figsize=(6, 6))  
  
 order = sorted(df3[category].unique())  
 ax = sns.countplot(category, data=df3, hue='Churn', order=order)  
 ax.set\_ylabel('')  
  
 bars = ax.patches  
 half = int(len(bars) / 2)  
 left\_bars = bars[:half]  
 right\_bars = bars[half:]  
  
 for left, right in zip(left\_bars, right\_bars):  
 height\_l = left.get\_height()  
 height\_r = right.get\_height()  
 total = height\_l + height\_r  
  
 ax.text(left.get\_x() + left.get\_width() / 2., height\_l + 40, '{0:.0%}'.format(height\_l / total), ha="center")  
 ax.text(right.get\_x() + right.get\_width() / 2., height\_r + 40, '{0:.0%}'.format(height\_r / total), ha="center")  
  
plt.show()  
  
*# Convert binary variables (yes/no, female/male) to 0 or 1*df['DmyGender'] = [1 if v == 'Male' else 0 for v in df['Gender']]  
df['DmyChurn'] = [1 if v == 'Yes' else 0 for v in df['Churn']]  
df['DmyTechie'] = [1 if v == 'Yes' else 0 for v in df['Techie']]  
df['DmyContract'] = [1 if v == 'Two Year' else 0 for v in df['Contract']]  
df['DmyPort\_modem'] = [1 if v == 'Yes' else 0 for v in df['Port\_modem']]  
df['DmyTablet'] = [1 if v == 'Yes' else 0 for v in df['Tablet']]  
df['DmyInternetService'] = [1 if v == 'Fiber Optic' else 0 for v in df['InternetService']]  
df['DmyPhone'] = [1 if v == 'Yes' else 0 for v in df['Phone']]  
df['DmyMultiple'] = [1 if v == 'Yes' else 0 for v in df['Multiple']]  
df['DmyOnlineSecurity'] = [1 if v == 'Yes' else 0 for v in df['OnlineSecurity']]  
df['DmyOnlineBackup'] = [1 if v == 'Yes' else 0 for v in df['OnlineBackup']]  
df['DmyDeviceProtection'] = [1 if v == 'Yes' else 0 for v in df['DeviceProtection']]  
df['DmyTechSupport'] = [1 if v == 'Yes' else 0 for v in df['TechSupport']]  
df['DmyStreamingTV'] = [1 if v == 'Yes' else 0 for v in df['StreamingTV']]  
df['DmyStreamingMovies'] = [1 if v == 'Yes' else 0 for v in df['DmyStreamingMovies']]  
df['DmyPaperlessBilling'] = [1 if v == 'Yes' else 0 for v in df['PaperlessBilling']]  
  
*# Drop original categories*df4 = df.drop(columns=['Gender', 'Churn', 'Techie', 'Contract', 'Port\_modem', 'Tablet', 'InternetService', 'Phone', 'Multiple', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'PaperlessBilling'])  
  
print(df4.describe())

## C4. Univariate and Bivariate Visualizations

Text

Description automatically generated

Chart

Description automatically generated

Figure 2: Boxplots for Continuous Variables

Graphical user interface, text, application, email

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure : Bar-plots for Categorical Variables

## C5. Churn Data Set

The prepared dataset used for this analysis has been uploaded with the assessment file.

# D. Model Comparison and Analysis

With the variables identified in C2 we will create our initial logistic regression model.

Graphical user interface, text, application, email

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Figure 4: Classification Report

According to the classification report for our initial model, 73% of the customers predicted to churn did so. The model also only correctly predicted this outcome for 64% of those customers. Our *f1-score* of 68% indicates that this model does a semi-descent job predicting whether a customer will churn. There were 27 variables evaluated during this logistic regression model; however, not all of the variables are/were needed; we can use a heatmap to determine which variables are most useful.

Chart

Description automatically generated

Figure 5: Heatmap with all variables

With all of the variables, the heatmap is difficult to read; however, it should be noted that there is some correlation between tenure and bandwidth gb year. Because our heatmap is difficult to interpret visually, we will look at the logit summary to get a better picture of our variables.

Table

Description automatically generated

Table

Description automatically generated

Figure 6: Initial Logistic Regression Model Results

We will concentrate on the p-values; p-values less than 0.05 are a good candidate for further research; below is the logit regression results. After review our initial logistic regression model summary we will redice our model down to the following variables: bandwidth gb year, age, income, tenure, yearly equip failure, port modem, tablet, streaming tv, paperless billing, timely response, churn, fixes, and replacements.

## D2. Justification of Based Variable Selection Procedure and Model Evaluation Metric

Originally our dataset contained 50 variables and 10000 observations: to choose the best variables for the analysis a correlation Matrix was created. After creating the correlation heatmap and comparing the heatmap to other visualizations that were created; it was noted that some of the features did not have much significance and/or were strongly correlated. After running the first correlation matrix, I then ran the stats models logit method on the same dataset; this allowed me to take a closer look at my variables and their significance level, I removed any variables which had a p-value greater than 0.05. The following variables were removed because they did not have much significance and/or were strongly correlated: age, children, income, yearly\_equip\_failure, dmyTable, dmyStreamingTV, dmyPort\_Modem, dmyPaperessBilling', timelyresponse, fixes, dmyMultiple, replacements, reliability, bandwidth\_gb\_year. The following variables will be used in our logistic regression model: outage\_sec\_perweek, tenure, monthlycharge, dmyTechie, dmyInternetService, dmyPhone, dmyOnlineSecurity, dmyOnlineBackup, dmyDeviceProtection, dmyTechSupport, dmyChurn, options, respectfulness, courteous, and listening.

Chart

Description automatically generated

Figure : Reduced Logistic Regression Model Heatmap

## D3. Logistic Regression Model



Table

Description automatically generated

Figure 9: Classification Report

Table

Description automatically generated

Figure 10: Reduced Logistic Model Results

According to the classification report for our reduced model, 76% of the customers predicted to churn did so. The model also only correctly predicted this outcome for 63% of those customers. Our *f1-score* of 69% indicates that this reduced model did improve on predicting whether a customer will churn.

# E. Data Set Analyzation

Chart, treemap chart

Description automatically generated

Figure 11: Initial Confusion Matrix

Out of 2000 random observations our initial logistic model produced the following confusion matrix with the following results tn = 1363, fp = 122, fn=187, and tp=328.

Chart, treemap chart

Description automatically generated

Figure 12: Reduced Model Confusion Matrix

Out of 2000 random observations our reduced logistic model produced the following confusion matrix with the following results tn = 1318, fp = 167, fn=214, and tp=301.

## E3. Logistic Regression Model Code

*# Split the dataset into a training and testing set. Using an 80/20 testing/training split*x\_train, x\_test, y\_train, y\_test = train\_test\_split(df5.drop('Churn', axis=1), df5['Churn'], test\_size=0.2, random\_state=200)  
  
*# Model evaluation default parameters*LogReg = LogisticRegression(solver='liblinear')  
res = LogReg.fit(x\_train, y\_train)  
  
*# Classification report precision*y\_pred = LogReg.predict(x\_test)  
  
*# Print classification report*print(classification\_report(y\_test, y\_pred))  
  
*# Print predictions*print(y\_pred)  
  
*# Building and Printing the Logit Summary*log\_reg = sm.Logit(y\_train, x\_train).fit()  
  
print(log\_reg.summary())

*# Create reduced model*x\_train, x\_test, y\_train, y\_test = train\_test\_split (df6.drop('Churn', axis=1), df6['Churn'], test\_size=0.2, random\_state=200)

*# Model evaluation default parameters*LogReg2 = LogisticRegression(solver='liblinear')  
  
*# train model*LogReg2.fit (x\_train, y\_train)  
  
*# Classification report precision*y\_pred = LogReg2.predict(x\_test)

# F. Summary

After comparing both models the reduced model did improve the accuracy of predicting churn customers;the reduced model for customers who churned precision increased by 3%, the recall value decreased by 1%, and the *f1-score* increased by 1%. The initial logistic regression model yielded a precision score of 73%, which was accurately predicted 64% of the time. The reduced model yielded a precision score of 76%, which was accurately predicted 63% of the time.

**Regression formula**:

y = (-0.07)\*outage\_sec\_perweek + (-0.08)\*tenure + (0.04)\*monthlycharge + (0.65)\*dmytechie + (-1.63)\*dmyInternetService +(-0.80)\*dmyPhone +(-0.33)\* dmyOnlineSecurity +

(-0.35)\*dmyOnlineBackup + (-0.35)\*dmyDeviceProtection + (-0.40)\*dmyTechSupport + (-0.34)\*Options + (-0.19)\*Respectfulness + (-0.2)\*Courteous + (-014)\*Listening

**Interpretation of Coefficients:**

The coefficients represent a positive and/or negative multiplier; monthlycharge and dmytechie variables represent a positive predictor for churn, while outage sec perweek, tenure, dmyinternetservice, dmyphone, dmyonlinesecurity, dmyonlinebackup, dmydeviceprotection, dmytechsupport, options, respectfulness, courteous and listening represent a negative predicator for churn.

**Limitations of Analysis**:

There are some limitations to this analysis, this analysis does not cover those customers that signed up during a promotional period and ended their subscription after that period ended; this analysis is also a small representative of a larger dataset and covers a limited period of time.

## F2. Recommended Course of Action

Since the reduced model only increased our predictions slightly, I would recommend including some of the omitted variables to see if they improve the results of the model; or we can also reevaluate the variables which were included in our reduced model to see which or if any of them are having a negative impact on our reduced models. Other recommendations are to include or create incentives for customers to sign longer contracts (1 to 2 years) and include more services; the analysis showed the more services a customer has the less likely they are to churn. I would also recommend looking into the customers that signed up during a promotional period; see why they signed up and why they discontinued their services after the promotional period.

# G. Panopto video recording

[Video Link](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=07db45a5-2531-4e4a-9514-aece01386448)

References

Frankenfield, J. (2022, February 8). Churn rate. Investopedia. Retrieved May 13, 2022, from https://www.investopedia.com/terms/c/churnrate.asp

Lawton, G., Burns, E., & Rosencrance, L. (2022, January 20). What is logistic regression? - definition from Searchbusinessanalytics. SearchBusinessAnalytics. Retrieved July 4, 2022, from https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression

Narkhede, S. (2021, June 15). Understanding Confusion Matrix. Medium. Retrieved July 4, 2022, from https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

NBM2 TASK 2: Logistic Regression For Predictive Modeling. WGU Performance Assessment. (n.d.). Retrieved July 4, 2022, from https://tasks.wgu.edu/student/000194226/course/23300006/task/2788/overview