D208 Predictive Modeling Performance Assessment

Task 2: Logistic Regression

Author: Eric Colwell

Western Governors University

This report deals with the rubric for Western Governors University course D208 Predictive Modeling Task 2 and answers all the items in rubric order.

# **Part I:**

**A1:**

The research question for this analysis is “Which variables influence the probability of Churn”?

**A2:**

The objective of this data analysis is to determine which variables or factors have the greatest influence on the probability of Churn through Logistic Regression. For this analysis I have chosen Churn as the dependent or target variable. The independent or predictor variables are Population, Area, Children, Age, Income, Marital, Gender, Outage\_sec\_perweek, Email, Contacts, Yearly\_equip\_failure, Techie, Contract, Port\_modem, Tablet, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, PaperlessBilling, PaymentMethod, Tenure, MonthlyCharge, Bandwidth\_GB\_Year, Item1, Item2, Item3, Item4, Item5, Item6, Item7, and Item8. Variables not included in the initial model are CaseOrder, Customer\_id, Interaction(UID), City, State, County, Zip, Lat, Lng, TimeZone, and Job. These variables have high cardinalities and would not add predictive power to the model.

# **Part II:**

**B1:**

A logistic regression model assumes that the target variable is categorical. The observations are independent of each other. It requires little to no multicollinearity between the independent variables. It assumes the independent variables are linearly related to the log odds and it requires a large sample size. The model is based on the Bernoulli distribution because the target variable is binary. The predicted values are restricted to a range of nominal values like ‘Yes’ or ‘No. It predicts the probability of particular outcomes and is the logarithm of the odds of achieving 1 (Sewell, 2022).

**B2:**

The benefits of using the Python language for this analysis are that it is a capable tool for data science, statistics, data exploration, data analysis, and predictive analytics. Also, I am more familiar with Python. For these reasons I chose this language for the Logistic Regression Analysis. I imported pandas to import the dataset, data manipulation, and analysis. I imported numpy for numerical calculations. I imported matplotlib.pyplot for plotting data. I imported stats from scipy for statistical analysis and outlier detection. I imported plotnine for plotting data. I imported missingno to visually check for missing values. I imported seaborn for visualizations. I imported LogisticRegression for model creation. I imported train\_test\_split for model testing. I imported metrics for additional statistical output. I imported variance\_inflation\_factor from statsmodels to check VIF scores. I imported statsmodels.api for model creation.

**B3:**

Logistic regression is commonly used for predictive analysis. Logistic regression is a statistical approach to model a relationship between a categorical dependent (target) variable and a set of independent (predictor) variables. It is used to describe data and explain relationships between the target variable and one or more nominal, ordinal, interval, or ratio-level independent variables. Also, it is used to model the probability of a certain class or event existing such as pass/fail, win/lose, or churn yes/churn no. In this case, logistic regression is used to answer the research question posed in A1.

# **Part III:**

**C1:**

My data preparation goals follow the Data Analytics Life Cycle. The research question is the Business Understanding phase. Importing the .csv file into my Jupyter notebook is the Data Acquisition phase. For the Data Cleaning Phase I checked for duplicated rows or instances then checked for null values. Next, using z-scores, I checked for outliers and remedied them using proper techniques. In the Data Exploration phase, I identified categorical (nominal and ordinal) values, and encoded them properly. These steps are necessary for the Predictive Modeling phase to begin.

**C2:**

Below are the summary statistics for the target variable and all predictor variables including screenshots of output showing the count of instances (10,000), mean (average), standard deviation (amount of variation of a set of data), minimum (least amount), quartile ranges (measure of variability of the data), maximum (largest amount) for continuous variables. Unique values, and value counts for categorical variables.

**Churn (Target Variable):** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer discontinued service within the last month (Yes, No).

Graphical user interface, text

Description automatically generated

**Population:** Datatype is int64, quantitative, continuous, and has no null values. The value reflects the population within a mile radius of the customer, based on census data (range from 0 to 111,850).

Table

Description automatically generated

**Area:** Datatype is object, qualitative, categorical, and has no null values. The value reflects the area type (Rural, Urban, Suburban), based on census data.

Table

Description automatically generated

**Children:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the number of children in the customer’s household as reported in sign-up information (ranging from 0 to 10).

Table

Description automatically generated

**Age:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the age of the customer as reported in sign-up information (ranging from 18 to 89).

Table

Description automatically generated

**Income:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the annual income of the customer (or invoiced person) as reported at time of sign-up (ranging from roughly 348 to 258,900).

Text

Description automatically generated

**Marital:** Datatype is object, qualitative, categorical, and has no null values. The value reflects the marital status of the customer as reported in sign-up information (Divorced, Married, Never Married, Separated, Widowed).

Text

Description automatically generated

**Gender:** Datatype is object, qualitative, categorical, and has no null values. The value reflects the customers self-identification as male, female, or nonbinary.

Text

Description automatically generated

**Outage\_sec\_perweek:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the average number of seconds per week of system outages in the customer’s neighborhood (range from roughly 0.1 to 21.2).

Text

Description automatically generated

**Email:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the number of emails sent to the customer in the last year (range from 1 to 23).

Table

Description automatically generated with medium confidence

**Contacts:** Datatype is float64, quantitative, discrete, and has no null values. The value reflects the number of times the customer contacted technical support (range from 0 to 7).

Table

Description automatically generated

**Yearly\_equip\_failure:** Datatype is float64, quantitative, discrete, and has no null values. The value reflects the number of times customers equipment failed and had to be reset/replaced in the past year (range from 0 to 6).

Table

Description automatically generated

**Techie:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer considers themselves technically inclined (Yes, No).

Text

Description automatically generated

**Contract:** Datatype is object, qualitative, categorical, and has no null values. The value reflects the contract term of the customer (Month-to-month, One Year, Two Year).

Text

Description automatically generated

**Port\_modem:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has a portable modem (Yes, No).

**Text

Description automatically generated**

**Tablet:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer owns a tablet such as iPad, Surface, etc. (Yes, No).

**Graphical user interface, text

Description automatically generated**

**InternetService:** Datatype is object, qualitative, categorical, and has no null values. The value reflects the customers internet service provider (DSL, Fiber Optic, or None).

**Text

Description automatically generated**

**Phone:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has a phone service (Yes, No).

**Text

Description automatically generated**

**Multiple:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has multiple lines (Yes, No).

Text

Description automatically generated

**OnlineSecurity:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has an online security add-on (Yes, No).

**Text

Description automatically generated**

**DeviceProtection:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has device protection add-on (Yes, No).

**Text

Description automatically generated**

**TechSupport:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has a technical support add-on (Yes, No)

**Text

Description automatically generated**

**StreamingTV:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has streaming TV (Yes, No).

**Text

Description automatically generated**

**StreamingMovies:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has streaming movies (Yes, No).

**Graphical user interface, text

Description automatically generated**

**PaperlessBilling:** Datatype is object, qualitative, categorical, and has no null values. The value reflects whether the customer has paperless billing (Yes, No).

Text

Description automatically generated

**PaymentMethod:** Datatype is object, qualitative, categorical, and has no null values. The value reflects the customer’s payment Method (Bank Transfer, Credit Card, Electronic Check, Mailed Check).

Text

Description automatically generated

**Tenure:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the number of months the customer has stayed with the provider (range roughly from 1 to 72).

Table

Description automatically generated with medium confidence

**MonthlyCharge:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the amount charged to the customer monthly. This is an average for the customer. For brand new customers, this value is the average for other customers who fit the new customers profile (range from roughly 80 to 290).

Text

Description automatically generated

**Bandwidth\_GB\_Year:** Datatype is float64, quantitative, continuous, and has no null values. The value reflects the average amount of data used, in GB, in a year by the customer (range from roughly 155 to 7159).

Text

Description automatically generated

The following variables (Item1 - Item8) represent responses to an eight-question survey asking customers the importance of various factors on a scale of 1 to 8 where 1 is most important and 8 is least important. The datatypes are float64, quantitative, continuous, and have no null values.

**Item1:** Timely response

Text

Description automatically generated with medium confidence

**Item2:** Timely fixes

Table

Description automatically generated

**Item3:** Timely replacements

Text

Description automatically generated with medium confidence

**Item4:** Reliability

Table

Description automatically generated

**Item5:** Options

Table

Description automatically generated

**Item6:** Respectful response

Table

Description automatically generated

**Item7:** Courteous exchange

Table

Description automatically generated

**Item8:** Evidence of active listening

Table

Description automatically generated

**C3:**

I imported the churn.csv file using pandas. Next, I checked the dataset shape and information. There are zero null values, 10,000 rows, and 50 columns. I checked for duplicated rows and found there were none. I made a copy of the original to preserve it. I made a new dataset that included the variables or columns that I used for the analysis. I checked the uniqueness of each of those chosen variables for validity. I used z-scores to identify outliers in the continuous variables, set the outliers equal to nan, then used median or mean imputation for replacement. I made a new dataframe (df2) that had been cleaned and includes only the variables being utilized for the Logistic Regression. In the Data Exploration phase I created univariate visualizations, and bivariate visualizations that are shown in C4. For the data wrangling phase I used ordinal encoding for the variables with Yes/No values. Then I used pd.get\_dummies with k-1 to encode the nominal variables. The dataset is now prepared for Logistic Regression.

The Jupyter notebook file (.ipynb format) containing the complete annotated code for this project will be uploaded with the submission.

**C4:**

Below are the univariate and bivariate visualizations for the cleaned dataset.

**Churn (Target variable):**

Chart, bar chart

Description automatically generated

**Population (univariate):**

Chart, histogram

Description automatically generated

**Population/Churn (bivariate):**

Chart

Description automatically generated

**Area (univariate):**

Chart, bar chart

Description automatically generated

**Area/Churn (bivariate):**

Chart, bar chart

Description automatically generated

**Children (univariate):**

**Chart, histogram

Description automatically generated**

**Children/Churn (bivariate):**

**Chart, histogram

Description automatically generated**

**Age (univariate):**

Chart, histogram

Description automatically generated

**Age/Churn (bivariate):**

Chart

Description automatically generated

**Income (univariate):**

Chart, histogram

Description automatically generated

**Income/Churn (bivariate):**

Chart, histogram

Description automatically generated

**Marital (univariate):**

**Chart, bar chart

Description automatically generated**

**Marital/Churn (bivariate):**

**Chart, bar chart

Description automatically generated**

**Gender (univariate):**

Chart, bar chart

Description automatically generated

**Gender/Churn (bivariate):**

Chart, bar chart

Description automatically generated

**Outage\_sec\_perweek (univariate):**

Chart, histogram

Description automatically generated

**Outage\_sec\_perweek/Churn (bivariate):**

A picture containing histogram

Description automatically generated

**Email (univariate):**

**Chart, histogram

Description automatically generated**

**Email/Churn (bivariate):**

**Chart, histogram

Description automatically generated**

**Contacts (univariate):**

Chart, histogram

Description automatically generated

**Contacts/Churn (bivariate):**

A picture containing chart

Description automatically generated

**Yearly\_equip\_failure (univariate):**

Chart, histogram

Description automatically generated

**Yearly\_equip\_failure/Churn (bivariate):**

Shape

Description automatically generated with medium confidence

**Techie (univariate):**

**Chart, bar chart

Description automatically generated**

**Techie/Churn (bivariate):**

**Chart, bar chart

Description automatically generated**

**Contract (univariate):**

Chart, bar chart

Description automatically generated

**Contract/Churn (bivariate):**

Chart, bar chart

Description automatically generated

**Port\_modem (univariate):**

**Chart, bar chart

Description automatically generated**

**Port\_modem/Churn (bivariate):**

**Chart, bar chart

Description automatically generated**

**Tablet (univariate):**

**Chart, bar chart

Description automatically generated**

**Tablet/Churn (bivariate):**

**Chart, bar chart

Description automatically generated**

**InternetService (univariate):**

**Chart, bar chart

Description automatically generated**

**InternetService/Churn (bivariate):**

**Chart, bar chart

Description automatically generated**

**Phone (univariate):**

**Chart, bar chart

Description automatically generated**

**Phone/Churn (bivariate):**

**Chart, bar chart

Description automatically generated**

**Multiple (univariate):**

Chart, bar chart

Description automatically generated

**Multiple/Churn (bivariate):**

Chart, bar chart

Description automatically generated

**OnlineSecurity (univariate):**

**Chart, bar chart

Description automatically generated**

**OnlineSecurity/Churn (bivariate):**

**Chart, bar chart

Description automatically generated**

**OnlineBackup (univariate):**

**Chart, bar chart

Description automatically generated**

**OnlineBackup/Churn (bivariate):**

**Chart, bar chart

Description automatically generated**

**DeviceProtection (univariate):**

**Chart, bar chart

Description automatically generated**

**DeviceProtection/Churn (bivariate):**

**Chart, bar chart

Description automatically generated**

**TechSupport (univariate):**

**Chart, bar chart

Description automatically generated**

**TechSupport/Churn (bivariate):**

**Chart, bar chart

Description automatically generated**

**StreamingTV (univariate):**

**Chart, bar chart

Description automatically generated**

**StreamingTV/Churn (bivariate):**

**Chart, bar chart

Description automatically generated**

**StreamingMovies (univariate):**

**Chart, bar chart

Description automatically generated**

**StreamingMovies/Churn (bivariate):**

**Chart, bar chart

Description automatically generated**

**PaperlessBilling (univariate):**

Chart, bar chart

Description automatically generated

**PaperlessBilling/Churn (bivariate):**

**Chart, bar chart

Description automatically generated**

**PaymentMethod (univariate):**

**Chart, bar chart

Description automatically generated**

**PaymentMethod (bivariate):**

**Chart, bar chart

Description automatically generated**

**Tenure (univariate):**

**Chart, histogram

Description automatically generated**

**Tenure (bivariate):**

A picture containing histogram

Description automatically generated

**MonthlyCharge (univariate):**

Chart, histogram

Description automatically generated

**MonthlyCharge/Churn (bivariate):**

Chart, histogram

Description automatically generated

**Bandwidth\_GB\_Year (univariate):**

Chart, histogram

Description automatically generated

**Bandwidth\_GB\_Year/Churn (bivariate):**

Histogram

Description automatically generated with low confidence

**Item1 (univariate):**

**Chart, histogram

Description automatically generated**

**Item1/Churn (bivariate):**

**Chart

Description automatically generated**

**Item2 (univariate):**

**Chart, histogram

Description automatically generated**

**Item2/Churn (bivariate):**

**Chart

Description automatically generated**

**Item3 (univariate):**

**Chart, histogram

Description automatically generated**

**Item3/Churn (bivariate):**

**A picture containing chart

Description automatically generated**

**Item4 (univariate):**

**Chart, histogram

Description automatically generated**

**Item4/Churn (bivariate):**

**Chart

Description automatically generated**

**Item5 (univariate):**

**Chart, histogram

Description automatically generated**

**Item5/Churn (bivariate):**

**Chart

Description automatically generated with medium confidence**

**Item6 (univariate):**

**Chart, histogram

Description automatically generated**

**Item6/Churn (bivariate):**

**Chart

Description automatically generated with medium confidence**

**Item7 (univariate):**

**Chart, histogram

Description automatically generated**

**Item7/Churn (bivariate):**

**Chart

Description automatically generated with medium confidence**

**Item8 (univariate):**

**Chart, histogram

Description automatically generated**

**Item8/Churn (bivariate):**

**Chart

Description automatically generated**

**C5:**

A copy of the cleaned and prepared dataset in .csv format will be uploaded with the submission.

# **Part IV:**

**D1:**

Initial Logistic Regression model with y-intercept below:

Text, letter

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

**D2:**

The method used for variable selection is the wrapper method backward stepwise elimination and the pseudo r-squared for the metric. The pseudo r-squared value of the initial model was 0.625. First, I looked for high correlation between the independent variables. Tenure and Bandwidth\_GB\_Year had a very high correlation at 0.991495 (Max is 1.0). I removed Tenure because it had a higher correlation to Churn. Next, I checked VIF (Variance Inflation Factor) values and found there were 13 variables with a VIF greater than 10. The variables were MonthlyCharge, Outage\_sec\_perweek, Email, and Item1 through Item8. Then, using the backward stepwise elimination variable selection technique, I removed the least significant variable with the highest p-value, re-fit the model, and ran Logistic Regression again. I iterated through this process until there were no variables with a p-value greater than 0.05 and all VIF values were less than 10. Once I achieved these goals, the reduced model was complete.

Table

Description automatically generated

**D3:**

Reduced model shown below:

**Text

Description automatically generated**

**Text

Description automatically generated with low confidence**

**Table

Description automatically generated**

**E1:**

The logic of the backward stepwise elimination variable selection technique is to remove the non-significant features iteratively according to the highest p-value. The p-values indicate whether there is a statistically significant relationship between the independent variable and the dependent variable. By removing the independent variables with p-value greater than 0.05 I created a model that is statistically significant.

I used pseudo r-squared for the model evaluation metric. This metric shows the percentage of the dependent variable variation the model explains. The pseudo r-squared value is between 0 and 100 percent. Zero percent means the model does not explain any of the variation in the dependent variable and 100 percent means the model explains all the variation in the dependent variable. For the reduced model, the pseudo r-squared value is 0.614, meaning 61 percent of the variation in the dependent variable can be explained.

**E2:**

**Confusion matrix below:**

Graphical user interface, text

Description automatically generated

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chart, treemap chart

Description automatically generated

**Predictions below:**

**Chart, histogram

Description automatically generated**

**E3:**

A Jupyter .ipynb file containing the full code for this project will be uploaded along with the Performance Assessment.

# **Part V: Data Summary and Implications**

**F1:**

The regression equation for the reduced model is:

y (Churn) = -0.83 + (0.06 (Children) + 1.10 (Techie) -0.31 (Phone) + 1.75 (Multiple) + 0.93 (OnlineBackup) + 0.54 (DeviceProtection) + 0.29 (TechSupport) +3.27 (StreamingTV) +3.77 (StreamingMovies) + 0.16 (PaperlessBilling) – 0.00 (Bandwidth\_GB\_Year) + 0.35 (Gender\_Male) -3.30 (Contract\_One year) -3.40 (Contract\_Two Year) -1.94 (InternetService\_Fiber Optic) -2.01 (InternetService\_None) +0.46 (PaymentMethod\_Electronic Check)).

The coefficient of each independent variable gives the size of the effect that variable has on the dependent variable. For example, keeping all things constant, a one unit increase in Children with change the log odds of Churn by + 0.06. For the categorical variable coefficients, StreamingMovies for instance, if the customer has the StreamingMovies add-on there is a change in the log odds of Churn by + 3.77. The sign (+ or -) gives you the direction of that effect. The coefficient for the constant (-0.83) is the average expected value for the dependent variable when all the independent variables are equal to zero.

As explained above, there is statistical significance for this model. The model has 18 statistically significant variables due to p-values being less than 0.05 and a pseudo r-squared value of 61 percent. This means there can be conclusions drawn from the model that may have practical implications on the dependent variable. The practical significance of the model is that, given the coefficients of the independent variables, efforts could be made to retain customers by offering add-ons at a reduced price or free.

The limitations of the analysis are that overfitting can occur because of limited data or too many independent variables. When there are independent variables that have high correlation with the other independent variables multicollinearity can occur. This can cause p-values and coefficients to be unreliable (Sewell, 2022).

**F2:**

The analysis of the reduced multiple linear regression model shows insight to some independent variables having a significant impact on Churn. For example, StreamingMovies has a coefficient of + 3.77. My recommendation is for further Logistic analysis using different variable selection techniques and metrics to see if these same independent variables continue to show statistical significance. Using different techniques and returning similar results would support the significance of the independent variables. If these variables show significance in different models, then I would recommend a course of action where the company could potentially retain customers at a higher rate with bonus offers. One example would be to offer free StreamingMovies add-on for customers who have completed a one- or two-year term.

**F:**

The Panopto video recording URL will be uploaded as part of the submission.

References

Sewell, W. (2022). D208 Predictive Modeling Webinar. Retrieved from <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b7ead95b-c392-4973-aa9c-ad1901031ab1>

Stojiljkovic, M. (n.d.). Logistic regression in python. <https://realpython.com/logistic-regression-python/>