D208 Predictive Modeling: Performance Assessment

Task 2: Logistic Regression

Jill Rivera

Western Governors University

Table of Contents

[I. Overview 3](#_Toc105079868)

[II. Method Justification 4](#_Toc105079869)

[III. Data Preparation 6](#_Toc105079870)

[IV. Logistic Regression Models 24](#_Toc105079871)

[V. Data Summary and Implications 33](#_Toc105079872)

[VI. References 36](#_Toc105079873)

[VII. Code References 37](#_Toc105079874)

[VIII. Table of Figures 37](#_Toc105079875)

# **Overview**

1. This paper will analyze a given data set and associated dictionary using the logistic regression statistical method. This analysis will attempt to answer a beneficial business question for the company, so that improvement can be made in the future.
   1. **Research Question**: Customer churn is a problem that every company encounters. This analysis will attempt to answer which features are significant to the churn variable in the data set.
      1. **Objectives**: The data set contains 10,000 customer records from a telecommunications company, with 50 features describing each record. This analysis will use an iteration process to reduce the size of the data set and focus on answering the research question presented.
      2. **Data Preparation**: The data will first be prepared for the logistic regression calculation. The data will be cleaned to ensure it is complete and is free from outlying factors. Next, the data will be explored using descriptive statistics and univariate and bivariate visualizations. Finally, the data will be wrangled, and the categorical features will be re-expressed as numerical values. Lastly, the clean data frame will be extracted into a CSV file for submission. To answer our business question, the target variable will be Churn and the remaining variables will be predictor variables.
      3. **Model Analysis**: An initial model will be created using all the variables in the data set. Through the wrapping iteration process, the data will be reduced to eliminate those variables which do not meet the assumptions of the logistic regression method. That reduced model will be further refined in the same manner to create a final model for analysis. Each iteration will check for multicollinearity Residual plots will be created for each variable to check for normal distribution of the residuals.

# **Method Justification**

1. Logistic regression is used to infer how confidently the value of a target variable can be predicted using the predictor variables that follow the assumptions of the logistic method.
   1. **Assumptions**: The logistic regression method requires the variables to meet certain parameters. First, the target variable must be binary. Second, the variables should all be independent of each other. Third, there can be little to no multicollinearity among the independent variables. (Solutions, 2022) Next, the independent variables be linearly aligned with the log odds. Lastly, this analysis method requires a large sample size with a minimum of 10 independent variables.
   2. **Tools**: The tools selected for this model creation are Jupyter Notebook, Python, Pandas, Statsmodels, Matplotlib, and Seaborn. Jupyter Notebook is an open-source web application that allows you to document, and visualize your results in the same place, using multiple coding languages. Python is a powerful, open-source programming language that has a plethora of supporting libraries and is very easy to read and write. Pandas is a Python library that makes reading and manipulating the data frame easy. Statsmodels is a gathering of classes and functions for the use of estimating various statistical models. Lastly, Matplotlib and Seaborn are libraries that are used to plot graphical visualizations. These tools are imported as shown in the following figures:

**Text

Description automatically generated**

Figure : Standard Libraries

Graphical user interface, text

Description automatically generated

Figure : Configuring Matplotlib

Text

Description automatically generated

Figure : Configuring Sklearn

Text

Description automatically generated

Figure : Configure Seaborn

Text

Description automatically generated

Figure : Configure Pandas

* 1. **Method’s Appropriateness**: Logistic regression requires that the target variable is categorical. Since the target variable for this analysis is a binary categorical variable, the method fits our analysis method. (Solutions, 2022) This means that we will be performing binary logistic regression to answer what features influence churn for this company. The data set fits the method since it has more than 10 variables describing the data. This combination will result in an accurate logistic regression calculation.

# **Data Preparation**

1. The data will be prepared for the logistic regression calculation through the steps of cleaning, exploring, and wrangling the data frame.
   1. **Data Cleaning**: The initial data set is imported into the Jupyter Notebook using Pandas read\_csv method and the first five rows of data are shown.

The 50 features of the data frame are listed.Text

Description automatically generated

Figure : List of Features

First, the last eight features from the customer survey are renamed for easier reading.Text

Description automatically generated

Figure : Rename Columns

Then, missing values are detected using the isnull method along with the sum method. This verifies that there is no missing data in the data frame.

Table

Description automatically generated with medium confidence

Figure : Detect Missing Values

Next, duplicated columns and rows are detected using the duplicated and any methods. We find that no rows or columns are duplicated.

Graphical user interface, text, application, chat or text message

Description automatically generated

Figure : Detecting Duplicated Data

Then, features with a high degree of cardinality are removed from the data set so that they do not skew the calculations. There are now 30 features in the data set. Table

Description automatically generated with medium confidence

Figure : Dropping Features with High Cardinality

Finally, outliers are found in the data using their z-score and eliminated from the data set. We can see that the data went from 10,000 records to 9175, dropping the 825 customer records with outlying values.

Graphical user interface, text, application

Description automatically generated

Figure : Detect Outliers

* 1. **Variable Statistics**: Summary statistics of all variables (target and predictor) used to answer the research question are shown in the two tables below. The numerical table shows the label, range and mean of each variable.

|  |  |  |
| --- | --- | --- |
| **Numerical Data** | | |
| **Label** | **Range** | **Mean** |
| Population | 0 - 5267 | 8527.72 +/- 11802.99 |
| Children | 0 - 8 | 1.95 +/- 1.90 |
| Age | 18 - 89 | 53.08 +/- 20.65 |
| Income | 348.67 – 124025.1 | 38298.94 +/- 25056.13 |
| Outage\_sec\_perweek | 1.144796 – 18.85173 | 10.01 +/- 2.93 |
| Email | 3 - 21 | 12.02 +/- 3.01 |
| Contacts | 0 - 3 | 0.94 +/- 0.90 |
| Yearly\_equip\_failure | 0 - 2 | 0.37 +/- 0.58 |
| Tenure | 1.005104 – 71.99928 | 34.44 +/- 26.45 |
| Monthly Charge | 79.97886 – 290.160419 | 172.73 +/- 42.99 |
| Bandwidth\_GB\_Year | 155.5067148 – 7158.98153 | 3380.57 +/- 2186.23 |

Figure : Numerical Data Summary

The categorical table shows the label and values of each variable. Most of the categorical features have yes and no values, making encoding them later as binary easier. The categories of marital and payment method will need to be reduced in the number of values for them to be re-expressed as numerical data for the calculation.

|  |  |
| --- | --- |
| **Categorical Data** | |
| **Label** | **Values** |
| Area | Urban, Suburban, Rural |
| Marital | Widowed, Married, Separated, Never Married, Divorced |
| Gender | Male, Female, Nonbinary |
| Churn | Yes, No |
| Techie | Yes, No |
| Contract | One year, Month-to-month, Two Year |
| Port\_modem | Yes, No |
| Tablet | Yes, No |
| InternetService | Fiber Optic, DSL, None |
| Phone | Yes, No |
| Multiple | Yes, No |
| OnlineSecurity | Yes, No |
| OnlineBackup | Yes, No |
| DeviceProtection | Yes, No |
| TechSupport | Yes, No |
| StreamingTV | Yes, No |
| StreamingMovies | Yes, No |
| PaperlessBilling | Yes, No |
| PaymentMethod | Credit Card (automatic), Bank Transfer (automatic), Mailed Check, Electronic Check |

Figure : Categorical Data Summary

The target categorical variable of churn is explored deeper. The count for each value is displayed and visualized on a graph.

A picture containing chart

Description automatically generated

Figure : Churn Value Counts

Next, the ratio of positive (yes) and negative (no) values are calculated. We can see that the data is not evenly distributed between yes and no values. This will need to be balanced with predictor data for the model creation to maintain the assumptions of the logistic method. Text

Description automatically generated

Figure : Churn Ratio

Finally, the totals of positive and negative values for each remaining feature are calculated. From this, it can be determined that the customers who churn have a much lower tenure than those who stay with the provider. Those leaving are also paying a higher monthly charge, using significantly less bandwidth and have a slightly higher income.

Table

Description automatically generated

Figure : Target and Features Compared

* 1. **Data Exploration**: Univariate visualizations of the distribution of variables in cleaned data set are shown. Histograms of the numerical data show the distribution patterns. Even distribution can be seen in the age, outage\_sec\_perweek, email, gender\_male, port\_modem\_yes, multiple\_yes, streamingTV, and streamingmovies graphs. The remaining features have an uneven distribution that will be addressed in the model.

**Diagram

Description automatically generated**

Figure : Univariate Statistics

Bivariate visualizations of the explanatory variable distribution against the target variable are shown. When churn is compared to monthly charge, a right-skewed pattern can be seen. This indicates that the mean is below the average.

Chart, bar chart, histogram

Description automatically generated

Figure : Churn vs. Monthly Charge Histogram

Plotting churn against tenure produces a bimodal pattern. This indicates that there is more than one peak of data for these features.

Chart, bar chart

Description automatically generated

Figure : Churn vs. Tenure Histogram

Analyzing churn against bandwidth\_GB\_year again shows a bimodal pattern.

Chart, bar chart

Description automatically generated

Figure : Churn vs. Bandwidth\_GB\_Year Histogram

Plotting churn against age produces a graph showing an even distribution of data points.

Chart, bar chart, histogram

Description automatically generated

Figure : Churn vs. Age Histogram

When comparing churn to income, it produces a right-skewed graph showing a positive skew.

Chart, bar chart

Description automatically generated

Figure : Churn vs. Income Histogram

Outage\_sec\_perweek in relation to churn shows a normal distribution pattern.

Chart, bar chart

Description automatically generated

Figure : Churn vs Outage\_sec\_perweek Histogram

Email and churn also have a normal pattern.

Chart, bar chart

Description automatically generated

Figure : Churn vs. Email Histogram

The histogram for churn and contacts shows a positively skewed pattern with the mean and mode remaining close together.

Chart, bar chart

Description automatically generated

Figure : Churn vs. Contacts Histogram

Yearly\_equip\_failure, population and children also show a right-skewed pattern when compared to churn. These all indicate a positive skew of the data.

**Chart, bar chart

Description automatically generated**

Figure : Churn vs. Population Histogram

**Chart, bar chart, histogram

Description automatically generated**

Figure : Churn vs. Children Histogram

**Chart, bar chart

Description automatically generated**

Figure : Churn vs. Yearly\_equip\_failure Histogram

* 1. **Data Wrangling**: Categorical features must be addressed before creating the logistic model for this analysis. First, the categorical variables are visualized in plots.

**A picture containing chart

Description automatically generated**

Figure : Categorical Features Graphical Plots

The features with more than three values will be re-expressed to only contain two values each, so that the categorical data can be then converted into numerical data. The features of payment method and marital are re-expressed in this method. Payment method is changed to have the values of either automatic or check. Marital is reduced to either married or not married.

Text

Description automatically generated

Figure : Re-express Payment Method and Marital

Finally, the categorical data is converted into numerical data using the one-hot encoding method. This indicates a 1 if the category has a positive or yes value and a 0 if the value is a no, changing it to a binary numerical feature.

Text

Description automatically generated

Figure : One-Hot Encoding Categorical Features

* 1. **Data Extraction**: The prepared data set is extracted to a CSV file using Pandas to\_csv method. The extracted file is labeled clean\_task2.csv and is attached for submission.

****

Figure : Exporting the Clean Data Set

# **Logistic Regression Models**

1. **Model Creation**: The initial model will be created using all the variables in the data set.
   1. **Initial Model**: Before the initial model can be created, the data must be rebalanced. With the data being imbalanced, there are too few examples of the minority class for a model to effectively learn the decision boundary. (Brownlee, SMOTE for Imbalanced Classification with Python, 2021) The data is rebalanced using the Synthetic Minority Oversampling Technique (SMOTE). This technique creates synthetic samples that are close to the feature so that the majority and minority data are balanced in quantity.

**Graphical user interface, text, application, email

Description automatically generated**

Figure : Rebalance using SMOTE (Lemaitre, Nogueira, & Aridas, 2022)

The Recursive Feature Elimination (RFE) process is now used to rank the features. The features are ranked based on how relevant they are at predicting the target variable. (Brownlee, Recursive Feature Elimination (RFE) for Feature Selection in Python, 2020) The feature selection and linear model libraries are imported, then a loop is created to perform the RFE.

Graphical user interface, text, application

Description automatically generated

Figure : Recursive Feature Elimination (Pedregosa, Varoquaux, & Gramfort, 2022)

The features that are ranked the highest are selected. The summary shows that the RFE went through seven iterations with 20 features. This shows a few variables with p-values higher than 0.05, which will need to be addressed before running the next model. We can see marital\_not\_married, device protection and techie are the features with the highest p-values.

Table

Description automatically generated

Figure : RFE Ranking Summary

A confusion matrix is created for the initial model to describe the performance of the model. This plots the expected and predicted data against each other. In this graph, we can see a total of 2806 predictions were made by the model. The model predicted 1474 positive and 1332 negative values, with the positive values being customers who churn. In actual numbers, 1414 customers churned and 1392 did not.

Graphical user interface, text, application

Description automatically generated

Graphical user interface

Description automatically generated

Figure : Initial Model Confusion Matrix

The number of predictions and the ratio of each are calculated for the model. The model created 85% correct predictions. Next, the model will be reduced to increase that correct value on the next model.

Text

Description automatically generated with medium confidence

Figure : Prediction Data

Finally, predictor pairs with high correlation are found by using a function so that the process can be re-run again later. The pairs with highest correlation are area with urban and suburban values, and internet service with fiber optic and none values. One value from each pair must be eliminated to prevent skewing or overfitting the data.

Table

Description automatically generated

Figure : Highly Correlated Feature Pair Detection

* 1. **Reduced Model**: The reduced and final model is created by first eliminating the features with high p-values and splitting up highly correlated pair of features. The model now has 15 explanatory variables to run through the RFE process. The variables for the model are reset and the model is run again. After seven iterations, the summary is shown. All p-values now are below our metric of 0.05.

Table

Description automatically generated

Figure : Model Feature Reduction

A confusion matrix for the reduced model is created. This time, a total of 2796 predictions were created. Of those, 1432 were positive and 1364 were negative. Actual numbers were 1434, while 1362 were predicted values.

Graphical user interface, application

Description automatically generated

Figure : Reduced Model Confusion Matrix

1. **Final Logistic Model Analysis**: The final model consisted of the target variable of churn, along with the following 15 features:
   * Yearly\_equip\_failure
   * Tenure
   * Area\_Suburban
   * Gender\_Nonbinary
   * Contract\_One year
   * Contract\_Two Year
   * Port\_modem\_Yes
   * Tablet\_Yes
   * InternetService\_Fiber Optic
   * Phone\_Yes
   * Multiple\_Yes
   * OnlineSecurity\_Yes
   * OnlineBackup\_Yes
   * TechSupport\_Yes
   * PaymentMethod\_Check
   1. **Model Comparison**: The initial model was created using all 20 variables in the data set. Through refinement, 5 variables were removed using variable selection techniques, and the final model was created with 15 explanatory variables. The evaluation metric was used to evaluate the model’s accuracy and applicableness.
      1. **Variable Selection**: The assumptions of logistic regression discussed earlier guided the variable selection. Those variables with a high p-value or multicollinearity with another variable were eliminated from the data set.
      2. **Evaluation Metric**: The Receiver Operating Characteristic (ROC) curve plot is constructed to evaluate the model, along with a classification report. The classification report shows that the model has an accuracy of 84% in predicting the data.

Table

Description automatically generated

Figure 41: Final Model Classification Report

The ROC curve is shown. The curve in the plot visualizes the ability of a random classifier as its discrimination threshold is varied. This tells how much the model can distinguish between the true positive rate vs. the false positive rate. (Hoo, Candlish, & Teare, 2017) The higher the accuracy, the more the curved line moves away from the dotted line of Y = 1. The area under this curve measures 0.84.

Chart, line chart

Description automatically generated

Figure 42: Receiver Operating Characteristic Plot

* 1. **Model Analysis**: The final model tested with an accuracy of 84%, which means that we have a fairly accurate model with the correct features. The ROC plot shows that a random classifier’s ability to distinguish between the true and false positives in the model, was calculated at 0.84.

# **Data Summary and Implications**

1. **Findings and Assumptions**: The final model is evaluated and recommendations for the business are the following.
   1. **Results**: The final model was evaluated using a regression equation to obtain coefficients for each feature in the model. The coefficients are plotted, and the model’s applicableness is evaluated. Finally, recommendations for the business moving forward are given.
      1. **Regression Equation**: A regression equation was constructed for the final model to obtain the coefficients of the features.

Text

Description automatically generated

Figure : Model Regression Equation

* + 1. **Coefficients**: The final model summary produced a coefficient for each explanatory variable. They are listed in the figure shown. The coefficients are the log odds of the customer’s likelihood of churning from the company. The constant of 4.0647 along with the coefficient of the explanatory variable can be added or subtracted together to obtain a value equal to the likelihood of that customer churning from the company. Positive coefficients indicate a higher likelihood of a customer churning, while the negative coefficients indicate that the customer will be less likely to churn. As an example, considering the predictor variable of multiple\_yes with a coefficient of +0.708. Calculating 4.023 + 0.708 = 4.731, we can see that if a customer has multiple services with the provider, they are more than four times likely to churn from the company. On the other hand, when viewing the feature of fiber optic service with a coefficient of -1.032, we see a different story. Performing the calculation of 4.023 – 1.032 = 2.991 shows that a customer who has fiber optic is less likely to churn.

Text

Description automatically generated

Figure 44: Final Model Coefficients

Coefficients are plotted in a logarithmic plot as well to show the range of values.

Chart, line chart

Description automatically generated

Figure : Final Model Coefficients

* + 1. **Model Significance**: The model created has proved to be accurate to an 84% rate. This is a practical model in determining the likelihood of a customer churning from the provider based on the presence of certain features. This model can be used to predict the most optimal combinations of services to offer customers, with the benefit of retaining the customer for a longer term.
    2. **Limitations**: A limitation of this analysis is when dealing with categorical features. One-hot encoding was used to re-express the categorical data as numerical. During this process, the drop\_first method in Pandas is marked as True. If this is not done, or marked as False, the data will be redundant since it is binary. All the information needed is contained in just one column, so the second one is eliminated.
  1. **Recommendations**: Based on the figures from the calculations, the provider would benefit from having customers sign up for multiple services. Offering bundle packages at a reduced cost to the customer would increase their likelihood of remaining at the provider for a longer period. Also, extending customer’s contracts will reduce the likelihood a customer will churn.

# **References**

Brownlee, J. (2020, August 28). *Recursive Feature Elimination (RFE) for Feature Selection in Python*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/rfe-feature-selection-in-python/

Brownlee, J. (2021, March 17). *SMOTE for Imbalanced Classification with Python*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve? *The Royal College of Emergency Medicine*, 34:357-359.

Solutions, S. (2022, May 26). *Assumptions of Logistic Regression*. Retrieved from Complete Dissertation: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

# **Code References**

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2022, May 31). *SMOTE*. Retrieved from Imbalanced Learn: https://imbalanced-learn.org/stable/references/generated/imblearn.over\_sampling.SMOTE.html

Pedregosa, F., Varoquaux, G., & Gramfort, A. (2022, May 31). *sklearn.feature\_selection.RFE*. Retrieved from Scikit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.feature\_selection.RFE.html

# **Table of Figures**

[Figure 1: Standard Libraries 5](#_Toc105078055)

[Figure 2: Configuring Matplotlib 5](#_Toc105078056)

[Figure 3: Configuring Sklearn 5](#_Toc105078057)

[Figure 4: Configure Seaborn 5](#_Toc105078058)

[Figure 5: Configure Pandas 5](#_Toc105078059)

[Figure 6: List of Features 6](#_Toc105078060)

[Figure 7: Rename Columns 7](#_Toc105078061)

[Figure 8: Detect Missing Values 8](#_Toc105078062)

[Figure 9: Detecting Duplicated Data 9](#_Toc105078063)

[Figure 10: Dropping Features with High Cardinality 9](#_Toc105078064)

[Figure 11: Detect Outliers 10](#_Toc105078065)

[Figure 12: Numerical Data Summary 10](#_Toc105078066)

[Figure 13: Categorical Data Summary 11](#_Toc105078067)

[Figure 14: Churn Value Counts 12](#_Toc105078068)

[Figure 15: Churn Ratio 12](#_Toc105078069)

[Figure 16: Target and Features Compared 13](#_Toc105078070)

[Figure 17: Univariate Statistics 14](#_Toc105078071)

[Figure 18: Churn vs. Monthly Charge Histogram 15](#_Toc105078072)

[Figure 19: Churn vs. Tenure Histogram 15](#_Toc105078073)

[Figure 20: Churn vs. Bandwidth\_GB\_Year Histogram 16](#_Toc105078074)

[Figure 21: Churn vs. Age Histogram 16](#_Toc105078075)

[Figure 22: Churn vs. Income Histogram 17](#_Toc105078076)

[Figure 23: Churn vs Outage\_sec\_perweek Histogram 18](#_Toc105078077)

[Figure 24: Churn vs. Email Histogram 18](#_Toc105078078)

[Figure 25: Churn vs. Contacts Histogram 19](#_Toc105078079)

[Figure 26: Churn vs. Population Histogram 20](#_Toc105078080)

[Figure 27: Churn vs. Children Histogram 20](#_Toc105078081)

[Figure 28: Churn vs. Yearly\_equip\_failure Histogram 21](#_Toc105078082)

[Figure 29: Categorical Features Graphical Plots 22](#_Toc105078083)

[Figure 30: Re-express Payment Method and Marital 23](#_Toc105078084)

[Figure 31: One-Hot Encoding Categorical Features 23](#_Toc105078085)

[Figure 32: Exporting the Clean Data Set 24](#_Toc105078086)

[Figure 33: Rebalance using SMOTE 24](#_Toc105078087)

[Figure 34: Recursive Feature Elimination 25](#_Toc105078088)

[Figure 35: RFE Ranking Summary 26](#_Toc105078089)

[Figure 36: Initial Model Confusion Matrix 27](#_Toc105078090)

[Figure 37: Prediction Data 27](#_Toc105078091)

[Figure 38: Highly Correlated Feature Pair Detection 28](#_Toc105078092)

[Figure 39: Model Feature Reduction 29](#_Toc105078093)

[Figure 40: Reduced Model Confusion Matrix 30](#_Toc105078094)

[Figure 41: Final Model Classification Report 31](#_Toc105078095)

[Figure 42: Receiver Operating Characteristic Plot 32](#_Toc105078096)

[Figure 43: Model Regression Equation 33](#_Toc105078097)

[Figure 44: Final Model Coefficients 34](#_Toc105078098)

[Figure 45: Final Model Coefficients 35](#_Toc105078099)