**D208 Predictive Modeling**

**Josh Rogers**

**Western Governors University**

**Contents**

**Part I: Research Question 3**

A1. Research question 3

A2. Objectives and Goals for Analysis 3

**Part II: Method Justification 4**

B1. Assumptions 4

B2.  Programming Language and Benefits 4

B3.  Justification of using Regression 5

**Part III: Data Preparation and Manipulation 5**

C1.  Data Cleaning 5

C2. Data Exploration (EDA) 6

Categorical Summary Statistic Visualizations 7

C3.  Univariate Visualizations 8

Bivariate Visualizations 11

C4.  Data Transformation (Data Wrangling) 13

C5.  Prepared Dataset 15

**Part IV: Model Comparison and Analysis 15**

D1. Initial Model 15

D2.  Model Reduction Method and Justification 15

D3. Reduced Model 16

E1.  Model Comparison 16

E2.  Residual Plot for Reduced Model 16

E3. Code 17

**Part IV: Data Summary and Implications 17**

F1.  Regression Equation 17

F2.  Recommendations 18

**Sources 19**

**Part I: Research Question**

**A1. State your research question**

Given that the telecommunications company data is centered around churn the research question posited for this logistic regression analysis is: what factors impact churn? The explanatory variables that will be included in this analysis are Children, Age, Income, Gender, Contract, PaymentMethod, Tenure, and MonthlyCharge.

**A2.  State Objectives and Goals for Analysis**

The purpose of this analysis is to identify the variables, if any, that may have a significant impact on customer churn. In doing so, the organization can build customer profiles for which retention strategies can be implemented to reduce churn. In addition to potentially highlighting at risk customer demographics the analysis may lend itself to changes in current business practices. For example, if contract type is found to be significantly impactful on churn the organization may opt to remove certain types or add options to better suit customer needs.

**Part II: Method Justification**

**B1. Assumptions**

The assumptions associated with logistic regression include firstly that the response variables can only take on two possible outcomes. In the following analysis we consider churn which yields the outcomes of either ‘Yes’ or ‘No’. Additionally, there is expected independence between observations in the data. Also, it is supposed that very little or no multicollinearity exists between the explanatory variables. Lastly, the relationship between log odds of the dependent variable and the independent variables should be linear (Logistic Regression in Machine Learning, 2017).

**B2.  Programming Language and Benefits**

The program language I have selected to utilize is python. While there are certainly many benefits to R, I have found that the consistent syntax among libraries in python to be highly beneficial. The readability of the language is another positive as it is like the English language (“R or python”*,* 2023). For this reason, it is easier to communicate and explain the code to others even those that do not potentially have a coding background.

The Python Package Index coupled with Anaconda house all the Python libraries. The aforementioned code readability makes it relatively easy to learn the functionality of new packages. This flexibility enables the analyst to do more than just data analysis (R or Python, 2023). The libraries that have been called for the logistic analysis to follow is shown below:

A screenshot of a computer program

Description automatically generated

The **pandas** library was utilized to call in the CSV file and create a pandas data frame which was cleaned and prepared for analysis. This includes the implementation of the describe() function for the initial summary statistics and the usage of the getdummies() function to convert the categorical variables to numerical values.

**Seaborn** and **matplotlib** were imported to assist in the univariate visualizations of the continuous variables in the form of histograms in addition to the residual bar chart found at the end of the analysis. It is important to be able to create effective visualizations as it not only assists the analyst in the work with the data it aids in communicating data to other stakeholders and decision makers.

The **logit** function was pulled in from **Statsmodels.formula.api** to build the logistic regression model once the categorical variables were one hot encoded and concatenated back with the original data set.

From the libraries **sklearn.feature\_selection, sklearn.linear\_model, sklearn.preprocessing and sklearn.metrics** the function of **RFE**, **LogisticRegression**, **StandardScalar, and accuracy\_score** were brought in respectively to perform the feature selection process for reducing the model. Initially, only the recursive feature elimination (**RFE**) and **LogisticRegression** functions were called. However, errors in the processing required some additional regularization which is why the **StandardScalar** function was included. The accuracy\_score function was called to perform the math on the confusion matrix and print out the accuracy percentage.

**B3.  Justification of using Regression**

The research question aims to determine what, if any, explanatory variables have a significant impact on customer churn. Given that customer churn is a categorical variable and not numerical, it is appropriate to utilize logistic regression.

**Part III: Data Preparation and Manipulation (Cleaning à Exploration à Wrangling)**

**C1.  Data Cleaning**

The research question asked what impacts customer churn? The variables being investigated are Children, Age, Income, Gender, Contract, PaymentMethod, Tenure, and MonthlyCharge.

Only columns being used for modeling were pulled into the pandas data frame which included the aforementioned variables. While the initial data set provided was relatively clean it was important to ensure that any null values were addressed and removed.

The full code is attached but a screenshot of the code with annotations is below:

A screenshot of a computer program

Description automatically generated

**C2. Data Exploration (EDA)**

A screenshot of a computer program

Description automatically generated

The summary statics above visualize the numerical variables included in the analysis. *Summary statistics for categorical variables cannot be calculated in such a manner but are visualized below as percentages of the sample population*.

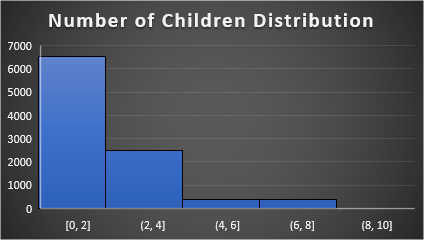
This summary provides an initial snapshot of some interesting measures of the sample data. The mean statistic calculates the average of each numerical category and provides a general understanding of the center values of the data. For example, the average number of children for the total population is 2.0877. Obviously there cannot be fractions of a child so we can estimate that the average customer has about 2 children.

The standard deviation describes the spread of the sample data from the mean. A lower standard deviation indicates that the data points are closer to the mean. Looking at the Income statistic we can see that most of the customers of this sample fall within the range of $39,806.93 plus or minus the standard deviation of $28,199.92. The max statistic for the Income variable of $258,900.70 may be seen as skewing the data due to its large deviation from the mean.

The 50% statistic, or the median, is also an important number to review. The median age for this sample group is about 53 years old. This may turn out to be an important explanatory variable when modeling the data and the potential influence of customer churn.

**Categorical Summary Statistic Visualizations:**

**C3.**  **Univariate** visualization



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Bivariate Visualizations**

**C4.  Data Transformation (Data Wrangling)**

Multiple actions were taken in wrangling the data to build the required models. First, when the CSV file was read in via Pandas, only the columns housing the selected explanatory variables along with the dependent variable (Churn) were pulled into the analysis. After this the data was filtered to remove null values by utilizing the dropna() function.

After these initial steps it was necessary to adjust the categorical variables to numerical values. First these columns were identified and stored in a pandas data frame called ‘categorical\_cols’. Then, by implementing the getdummies() function, the ‘categorical\_cols’ were passed through. Removal of the first output was set to ‘True’ and the datatype was set to output an integer for ease of calculations moving forward.

One the categorical values were adjusted those results were concatenated with the original filtered data and the original categorical column names were dropped to avoid confusion with those that were generated by the getdummies() function. Due to some processing errors in the model, the newly created columns were renamed to remove any spaces from their names.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

**C5.  Prepared Dataset**

Dataset attached: **D208\_T2\_Prepared\_Churn\_Data**.

**Part IV: Model Comparison and Analysis**

**D1.  Initial Model**

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer program

Description automatically generated

**D2.  Model Reduction Method and Justification**

The method utilized for feature selection was the iterative process of recursive feature elimination (RFE). RFE works to determine the level of impact of each feature and helps to remove those that rank lower than others (Filho, 2023). The analyst has the ability to determine how many variables will be eliminated and control the overall size of the reduced model. This method ensures that the most important features, relative to each other, are kept in the model.

Initially, the LosticRegression() function was passed through the RFE function with only the default parameters set which caused a processing error. It was determined that data regularization would be useful. The standardscaler() function was called to regularize the feature data and centering the data more appropriately. The L1 (Lasso) penalty was an additional regularization technique implemented in the feature selection process. This method helps to identify less important features and reduce their coefficients (Tewari, 2021). Lastly, with an iteration error still present, the **max\_inter** parameter in the LogisticRegression function was increased from the default 100 to 500.

**D3. Reduced Model**

A screenshot of a computer program

Description automatically generated

**E1.  Model Comparison**

As can be seen by comparing the two confusion matrices there is a high number of true positive (TP) and true negatives (TN) being shown in each model. However, it can bee seen that in the reduced model the number of TP/TN go up and the number of false positives (FP) and false negatives (FN) goes down. This indicates that the model has been improved with the performed reduction method.

While the confusion matrix provides an overview of the accuracy of the model the precision of its performance is another key metric to review. The calculation for precision is below:

A screenshot of a computer

Description automatically generated

Based on the confusion matrix calculated from the reduced model the precision of the model is 6816 / (6816 + 676) which comes out to approximately a 91% accuracy rate. This is to say that the model, with the selected features, accurately predicts the chances of a customer churning with a 91% precision rate.

Additionally, we can review the log-likelihood values which measure the goodness of fit for both models. The higher the log-likelihood value the better the model fits the data. In the original model the log-likelihood value was -4,593.8 compared to the much higher value for the reduced model of -2,708.8. Given this information, we can see that the reduced model fits the data better than the original.

**E2.  Reduced Model Confusion Matrix and Accuracy Calculation**

1. Confusion Matrix

A screenshot of a computer program

Description automatically generated

1. accuracy calculation

A screenshot of a computer program

Description automatically generated

**E3. Code**

Please find the link to my replit.com page below for the executable code. Additionally, a Word document has been attached containing this same code.

[Josh Rogers D208 Python Code](https://replit.com/@jrog248/Python#main.py)

**Part V: Data Summary and Implications**

**F1.   Regression Equation**

x1 = Tenure

x2 = PaymentMethod\_E\_Check

x3 = Contract\_1y

x4 = Contract\_2y

x5 = Monthlycharge

A screenshot of a computer

Description automatically generated

(Logistic regression calculator, n.d.)

1. **Interpretation**

Each of the coefficients represents the average change in log odds of the response variable, in this case churn, in relation to a one unit increase of the feature variable. For example, the coefficient for the variable PaymentMethod\_E\_Check is a positive 0.3725. This positive number indicates that a customer with this payment method increases the chances of a customer churning being true.

For the continuous variables we can interpret the coefficients a little different. Using the negative coefficient for tenure we can conclude that customers with a longer tenure are less likely to be involved in customer churn.

1. **Statistical and Practical Significance**

The pseudo R-squared value for the reduced model was calculated as 0.5315 which shows an improvement from the original model of a pseudo R-squared value of 0.2055. This metric helps explain variability and goodness of fit for the model. We can infer that the selected features explain 53% or the variance of customer churn and thus deem the model statistically significant at least compared to the original.

This model does also provide some practical significance. Businesses try to identify what models can help predict which customers are likely to churn. The high precision score calculated for this model (91%) suggests that there is a low number of false positives. This allows for more accurate resource allocation towards retention efforts with more confidence that positive results can be achieved.

Decisions regarding which variables to include may have impacted model performance. Including more variables from the beginning could have increased the strength of the reduced model after the iterative reduction elimination method was utilized.

**F2.  Recommendations**

It is recommended that the company utilize this model due to the measured accuracy and precision of its predictive ability. Specifically, leadership should look at ways of addressing their contract type offerings. This model shows that both the one and two year contract types decreases the probability that a customer will churn. Therefore, it can be inferred that the month-to-month contract type is more prevalent amongst churning customers.

Eliminating the month-to-month contract altogether may have negative impacts on sales and overall customer population. However, more investigation regarding alternative contract types would be a suggestion I would offer given the model results.

Additionally, customers using the electronic check payment method increases the probability of customer churn given its coefficient of a positive 0.3725. Again, it may be prudent to either eliminate this payment method to decrease churn or provide more incentives for utilizing other payment methods.

**Sources**

**H. Code Sources**

*How to create a dictionary in Python*. (2023, July 4). GeeksforGeeks. https://www.geeksforgeeks.org/how-to-create-a-dictionary-in-python/

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

MLNerds (2021) *Recursive feature elimination for feature selection*, *Machine Learning Interviews*. Available at: https://machinelearninginterview.com/topics/machine-learning/recursive-feature-elimination-for-feature-selection/#:~:text=Recursive%20Feature%20Elimination%20or%20RFE%20can%20be%20readily,from%20sklearn.feature\_selection%20import%20RFE%20from%20sklearn.linear\_model%20import%20LogisticRegression (Accessed: 13 March 2024).

Rithp (2023) *Logistic regression for feature selection: Selecting the right features for your model*, *Medium*. Available at: https://medium.com/@rithpansanga/logistic-regression-for-feature-selection-selecting-the-right-features-for-your-model-410ca093c5e0 (Accessed: 13 March 2024).

**I. Content Sources**

Filho, M. (2023) *How to solve logistic regression not converging in Scikit-learn*, *Forecastegy (Alt + H)*. Available at: https://forecastegy.com/posts/how-to-solve-logistic-regression-not-converging-in-scikit-learn/ (Accessed: 13 March 2024).

*Logistic regression calculator*. (n.d.). Online Statistics Suite. <https://stats.blue/Stats_Suite/logistic_regression_calculator.html>

*Logistic Regression in Machine Learning*. (2017, May 9). GeeksforGeeks. <https://www.geeksforgeeks.org/understanding-logistic-regression/#assumptions-of-logistic-regression>

R or Python. (2023, July 7). Retrieved from https://www.wgu.edu/online-it-degrees/programming-languages/r-or-python.html

Tewari, U. (2021, November 9). Regularization — Understanding L1 and L2 regularization for Deep Learning. Medium.com. Retrieved March 14, 2024, from https://medium.com/analytics-vidhya/regularization-understanding-l1-and-l2-regularization-for-deep-learning-a7b9e4a409bf

Tim (2020) *Regularization: Simple definition, L1 & L2 penalties*, *Statistics How To*. Available at: https://www.statisticshowto.com/regularization/ (Accessed: 13 March 2024).