**Logistic Regression to Assess Customer Churn**

Kelseyann Wright

Student ID: 00565814

College of Information Technology, Western Governors University

Predictive Modeling - D208

Dr. Eric Straw

April 18, 2023

**Logistic Regression to Assess Customer Churn**

**A1. Research Question**

The question that will be addressed is, "What variables are associated with customer churn?". This research question is important because understanding which factors determine if a customer “churns”, or cancels services, will allow the business to implement targeted changes to prevent customer churn.

**A2. Goals**

The goal of this analysis is to use a logistic regression to identify features that are related to customer churn to make predictions about how the related features will affect customer churn.

**B1. Summary of Assumptions**

The assumptions of a multiple linear regression are: 1) the outcome variable is binary, such as yes/no, 2) there is a linear relationship between each continuous variable and the log-odds of the outcome, 3) there are no strongly influential outliers, and 4) there is no multicollinearity.

**B2. Tool Benefits**

Python programming language was used to model the data. Python was chosen for its versatility, readability, and diverse functions. In addition, the syntax is consistent and simple to interpret, which is ideal when sharing your analysis with others who may not have coding experience. R does not have consistent syntax across libraries, making it more difficult to understand learn and understand.

Several python libraries were used, including pandas, numpy, matplotlib, scipy, sci-kit learn, and seaborn. Pandas is a common library used for working with data frames in python. It includes many tools for computations and analysis on data frames (McKinney, 2010). Numpy is used for many summary statistics throughout the analysis process (Harris et al., 2020; Larose & Larose, 2019). Scipy and sci-kit learn libraries work together to offer various statistical analyses (Pedregosa et al., 2011; Virtanen et al., 2020). Matplotlib and seaborn offer data visualization tools (Hunter, 2007; Waskom, 2021).

**B3. Appropriate Technique**

Logistic regression allows us to assess the relationship between a binary categorical variable, like customer churn, and multiple predictor variables of different types. This analysis will allow us to identify which variables have a significant relationship with customer churn and make predictions about customer churn based on other predictor variables.

**C1. Data Cleaning Goals**

First, data was checked for missing values and duplicates. No missing values or duplicates were found in the provided dataset. Second, the categorical variables with high cardinality, specifically more than five groups, were dropped from the dataset. The annotated code is attached.

**C2. Summary Statistics**

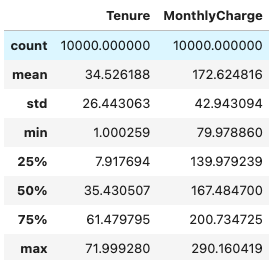
The dependent variable for the analysis is ‘Churn’. The independent variables chosen for the initial model are 'Techie', 'Phone', 'Multiple', 'OnlineBackup', 'DeviceProtection', 'StreamingTV', 'StreamingMovies', 'Tenure', 'MonthlyCharge', 'Gender’, 'Contract’ 'InternetService’, and 'PaymentMethod. These variables were identified using t-test statistics to identify potentially significant relationships between the variables and customer churn. In addition, a correlation heatmap was used to eliminate variables with multicollinearity.

Table : Summary Statistics of Relevant Quantitative Variables

The describe() function in python was used to calculate summary statistics (Table 1). The relevant categorical variables are not shown in the output as summary statistics are not calculated for categorical variables, only for quantitative variables.

The mean customer tenure is 34.5 months, which means that customers stay enrolled in services for 34.5 months on average. The standard deviation is 26.4 months, with a minimum of 1 month and a maximum of almost 72 months. This means that there is a wide range in customer tenures. The mean Diagram

Description automatically generated with low confidencefalls close to the middle of the range. The mean monthly charge is 172.6 GB per year, with a standard deviation of 42.9 and a range from 79.98 to 290.16. As with the tenure variable, the mean falls close the middle of the range.

**C3. Visualizations**

Univariate graphs for all variables are provided in figure 1. Bivariate graphs for all variables are provided in figure 2.

**C4. Data Transformation**

Categorical variables were transformed into dummy variables for the analysis. Dummy variables are binary values of 0 or 1, where 0 represents "no" or not being a member of that group and 1 represents "yes" or being a member of the group. Using Background pattern

Description automatically generateddummy variables prevents us from having to make multiple models to represent different groups (Garavaglia and Sharma, 1998).

Figure : Univariate graphs

Figure : Univariate and bivariate graphs for relevant variables

After creation of dummy variables, the data were scaled using the StandardScaler() function from the sci-kit learn library in python (Pedregosa et. al 2011). This function scales features so that they will have a Gaussian normal distribution with a mean of zero and a standard deviation of 1. Scaling data is important so that all features carry the same weight in the model and so that regression coefficients can be directly compared between features. This is also important for the recursive feature elimination used during model reduction.

Figure 2: Bivariate graphs

**C5. Prepared Dataset**

The final prepared dataset is attached.

**D1. Initial Model**

The initial model was run using the LogisticRegression() function in the sci-kit learn library. The regression output is provided below.

A screenshot of a computer

Description automatically generated with medium confidence

**D2. Justification of Model Reduction**

Recursive feature elimination (RFE) was used to identify which features should be dropped from the model. RFE eliminates features by fitting the model multiple times while recursively eliminating the weakest features based on their importance. RFE is preferred over stepwise feature selection because stepwise feature selection can sometimes include features that do not directly impact the dependent variable but are correlated with another feature that does (Smith 2018). The RFE function from the sci-kit learn python library was used for this analysis (Pedregosa et al. 2011).

**D3. Reduced Linear Regression Model**

The reduced multiple linear regression model is:

*Tenure = b1* *Bandwidth\_GB\_Year + b2 Churn + b3 OnlineBackup + b4 DeviceProtection + b5 Contract\_MonthToMonth*

where and *bi (i= 1-5)* are the regression coefficients. The regression output is provided below.

Text

Description automatically generated with medium confidence

**E1. Model Comparison**

RFE removed five of the features from the dataset, resulting in a multiple regression model with five independent variables: Bandwidth\_GB\_Year, Churn, OnlineBackup, DeviceProtection, and Contract\_MonthToMonth. The R2 value increased slightly in the reduced model, but the change in R2 and RMSE is almost zero. Cross validation tests of the two models produced similar results, with the average R2 of the reduced model being slightly higher than the initial model.

*Cross validation results for initial model:*

**

*Cross validation results for reduced model:*

**

**E2. Output and Calculations**

The RMSE of the initial model was 0.011454. The RMSE of the reduced model was slightly lower at 0.011447. For both models, the RMSE difference between the training dataset and test dataset were small, indicating that the models were not overfit to the training data. The difference in RMSE was slightly higher for the reduced model, but the difference is so small it is negligible. Residual qqplots were created for each model, which yielded nearly identical results.

|  |  |
| --- | --- |
| **Initial Model** | **Reduced Model** |
|  |  |

**E3. Code**

The python code written in jupyter lab notebook is attached.

**F1. Results**

The reduced regression model equation with predicted coefficients is:

*Tenure = 0.965\*Bandwidth\_GB\_Year - 0.062\*Churn – 0.016\*OnlineBackup -0.016\*DeviceProtection + 0.019\*Contract\_MonthToMonth*

Based on the regression coefficients, Bandwidth\_GB\_Year and Contract\_MonthToMonth have positive relationships with customer tenure. The remaining variables, Churn, Onlinebackup, and DeviceProtection all have negative relationships with customer tenure. The correlation coefficient for Bandwidth\_GB\_Year is the largest number. Because the data were standardized before the regression, we can interpret this to mean that bandwidth has the largest effect on customer tenure.

A picture containing text, receipt

Description automatically generated Based on the high R2 values, both models are statistically significant. The reduced model statistics were checked using the OLS function in the statsmodels python library (Figure 2). The F-statistic is well above 4, indicating a statistically significant model, and all predicted coefficients are statistically significant based on the P-value being less than 0.01 (Figure 2).

Figure : Reduced model statistics from OLS function in statsmodels python library

The model is also practically significant, because it tells us there is a very strong positive relationship between the amount of bandwidth a customer uses and customer tenure. The more bandwidth a customer uses each year, typically the longer the tenure for that customer. However, one limitation of linear regressions is that correlation does not necessarily mean causation. So, just because bandwidth and tenure are very strongly correlated, that does not necessarily mean the increase in bandwidth causes the increase in customer tenure. There could be another reason the two variables are related that we do not have data for or that is not accounted for in our model.

**F2. Recommendations**

My recommendation for next steps would be to test a follow up research question, which is: “What causes increased bandwidth usage?”. This is an important question to answer to tease apart the relationship between bandwidth and tenure and assess if the correlation relationship is also causal. We could survey customers about how they use their internet (entertainment, work, etc.) to gain additional data about bandwidth use each year. Understanding what drives the increased bandwidth use would 1) give us more information to assess the causality of the relationship between bandwidth and tenure, and 2) allow us to better plan for ways to increase bandwidth use and therefore increase tenure if the relationship is causal.

**G. Panopto Video**

A Panopto video was attached and submitted with the performance task. The link can also be accessed here: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=5f5d2c2c-c3dd-4cdc-9830-afe700946534>

1. **Sources of Third-Party Code**

No additional web sources of data or third-party code were used.

**H. Sources**

Garavaglia, S., & Sharma, A. (1998). A smart guide to dummy variables: Four applications and a macro. In Proceedings of the northeast SAS users group conference (Vol. 43).

Harris, C.R., Millman, J.K., van der Walt, S.J, Gommers, R., Virtanen, P., Cournapeau, V., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., Fernandez del Rio, F., Wiebe, M., Peterson, P., Gerard-Marchant, P., Sheppard, K., … Oliphant, T.E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362.

Hunter, J.D., "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

Larose, C. D., & Larose, D. T. (2019). Data science using Python and R. Wiley.

McKinney, W. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Smith, G. (2018). Step away from stepwise. J Big Data 5, 32.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., … van Mulbregt,P. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17(3), 261-272.

Waskom, M. L. (2021). Seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.