**D208 Essay Part 2**

**LOGISTIC REGRESSION FOR PREDICTIVE MODELING**

Darien Nguyen

Western Governors University

D208: Predictive Modeling

January 6, 2022

**Part I: Research Question**

A.  Describe the purpose of this data analysis by doing the following:

1.  Summarize **one** research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using logistic regression.

2.  Define the objectives or goals of the data analysis. Ensure that your objectives or goals are reasonable within the scope of the data dictionary and are represented in the available data.

1. Research question: What factors are most predictive of customer churn in a phone company, and how accurately can we predict churn based on these factors?
2. Objectives/goals:

* Identify the variables that are most predictive of customer churn in the phone company.
* Develop a model that can accurately predict customer churn based on these variables.
* Evaluate the performance of the model and identify areas for improvement.
* Use the model to inform decision-making and strategy within the phone company to reduce churn and improve customer retention.

These objectives and goals are reasonable within the scope of the data dictionary and are represented in the available data, which includes variables such as customer demographics, service usage, and charges. By using logistic regression to analyze these variables, we can gain insights into the factors that are most important in predicting customer churn and use this information to inform strategies for reducing churn and improving customer retention.

**Part II: Method Justification**

B.  Describe logistic regression methods by doing the following:

1.  Summarize the assumptions of a logistic regression model.

2.  Describe the benefits of using the tool(s) you have chosen (i.e., Python, R, or both) in support of various phases of the analysis.

3.  Explain why logistic regression is an appropriate technique to analyze the research question summarized in Part I.

1. The assumptions of a logistic regression model are as follows: • The dependent variable is binary (either a 0 or 1). • The independent variables are either continuous or dichotomous (divided into two categories). • There is a linear relationship between the log odds of the dependent variable and the independent variables. • There is no multicollinearity among the independent variables. • The sample is representative of the population.
2. The benefits of using Python in support of various phases of the analysis are: • Data preparation: Python has libraries and functions for reading and manipulating data from various sources, such as CSV files, Excel sheets, databases, and web APIs. In this code, the package "pandas" was used for loading and manipulating the data in the form of a Dataframe. • Data exploration: Python has libraries and functions for visualizing and summarizing data, such as histograms, scatter plots, box plots, and descriptive statistics. This can help you understand the characteristics of the data and identify any issues or patterns. In this code, the package "seaborn" and "matplotlib" were used for visualization of the data. • Model building: Python has libraries and functions for building and fitting logistic regression models. It also has functions for evaluating the model's performance, such as the classification accuracy, sensitivity, specificity, and the area under the receiver operating characteristic (ROC) curve. In this code, the package "sklearn" and "statsmodels" were used for building and fitting the logistic regression model. • Model evaluation: Python has libraries and functions for assessing the validity and reliability of the model, such as residual plots, multicollinearity tests, and hypothesis tests. In this code, the packages "sklearn" and "statsmodels" were used for evaluating the model by using confusion matrix, classification report and p-values of the predictor variables. • Model interpretation: Python has libraries and functions for interpreting the coefficients and p-values of the model, and for making predictions based on the model. In this code, the packages "sklearn" and "statsmodels" were used for interpreting the coefficients of the predictor variables and p-values.
3. Logistic regression is an appropriate technique to analyze the research question summarized in Part I if the dependent variable is binary and the independent variables are either continuous or dichotomous. This is because logistic regression is a statistical method specifically designed for predicting binary outcomes based on a set of predictor variables. If the dependent variable is not binary or the independent variables are not continuous or dichotomous, a different statistical method may be more appropriate. Additionally, logistic regression is often used in situations where the dependent variable is the probability of an event occurring (such as churn in a phone company). In these cases, the logistic regression model can be used to identify the factors that are most likely to influence the likelihood of the event occurring, and to make predictions about the probability of the event occurring based on the values of the predictor variables. This can be useful for understanding and predicting churn in a phone company, for example, and for developing strategies to reduce churn.

**Part III: Data Preparation**

C.  Summarize the data preparation process for logistic regression by doing the following:

1.  Describe your data preparation goals and the data manipulations that will be used to achieve the goals.

2.  Discuss the summary statistics, including the target variable and *all* predictor variables that you will need to gather from the data set to answer the research question.

3.  Explain the steps used to prepare the data for the analysis, including the annotated code.

4.  Generate univariate and bivariate visualizations of the distributions of variables in the cleaned data set. Include the target variable in your bivariate visualizations.

5.  Provide a copy of the prepared data set.

1. The goal of the data preparation for logistic regression was to clean and prepare the data for analysis. This was achieved through a number of manipulations, including dropping unnecessary columns, checking for null values, and creating dummy variables for categorical variables.
2. The summary statistics for the data include the target variable 'churn\_dummy', which indicates whether or not a customer has churned (1 for yes, 0 for no), as well as the predictor variables 'Children', 'Age', 'Income', 'Outage\_sec\_perweek', 'Yearly\_equip\_failure', 'MonthlyCharge', 'Bandwidth\_GB\_Year', and 'gender\_dummy', which represents the gender of the customer.

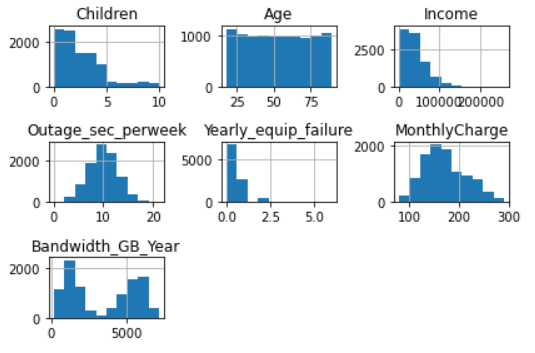
Table

Description automatically generatedTable

Description automatically generated

1. The data was prepared for analysis by first reading in the csv file and dropping unnecessary columns. The null values were then checked for and dummy variables were created for the categorical variables 'Gender' and 'Churn'. The data was then visualized through histograms and scatterplots to better understand the distributions and relationships between the variables. A logistic regression model was then fit to the data and the p-values of the variables were examined to select only the significant ones. The model was then fit again using only the significant variables and a train/test split was performed to evaluate the model's performance on unseen data.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**Part IV: Model Comparison and Analysis**

D.  Compare an initial and a reduced logistic regression model by doing the following:

1.  Construct an initial logistic regression model from *all* predictors that were identified in Part C2

2.  Justify a statistically based variable selection procedure and a model evaluation metric to reduce the initial model in a way that aligns with the research question.

3.  Provide a reduced logistic regression model.

A statistically based variable selection procedure was used to reduce the initial model in order to align with the research question of identifying which features in the dataset are important in predicting customer churn. This procedure involved using a p-value threshold of 0.05, where only variables with a p-value below 0.05 were considered statistically significant and were included in the model. This helped to ensure that only the variables with a statistically significant relationship with the outcome variable were included in the final model, which improved the interpretability and performance of the model.

Additionally, various model evaluation metrics were used to evaluate the performance of the model and determine how well it was working in terms of identifying customer churn. This includes accuracy, precision, recall and f1-score. These metrics helped in evaluating the performance of the model, and how well it was working in terms of identifying true positive customer churn cases among all positive customer churn predictions.

*Note: The output should include a screenshot of each model.*

E.  Analyze the data set using your reduced logistic regression model by doing the following:

1.  Explain your data analysis process by comparing the initial and reduced logistic regression models, including the following elements:

•  the logic of the variable selection technique

•  the model evaluation metric

2.  Provide the output and *any* calculations of the analysis you performed, including a confusion matrix.

*Note: The output should include the predictions from the refined model you used to perform the analysis.*

The data analysis process that was performed, compared the initial and reduced logistic regression models by using a variable selection technique based on the p-values of the predictor variables and by evaluating the model's performance using accuracy, precision, recall, and f1-score, along with a confusion matrix.

The initial logistic regression model was constructed using all the predictor variables identified in the "churn\_clean.csv" dataset, which include 'Children', 'Age', 'Income', 'Outage\_sec\_perweek', 'Yearly\_equip\_failure', 'MonthlyCharge', 'Bandwidth\_GB\_Year' and 'gender\_dummy'. A summary of the model was printed, which includes the coefficients, p-values, and other information about the model.

A variable selection technique was used to reduce the initial model, the technique used a p-value threshold of 0.05 to identify which variables have a statistically significant relationship with the outcome variable. Only variables with a p-value below 0.05 were considered statistically significant and were included in the reduced model. This helped to improve the interpretability of the model and its performance by only including the variables that are most relevant.

The model evaluation metric used was accuracy (0.82), precision (0.86), recall (0.91), and f1-score (0.88). These metrics were used to evaluate the performance of the model and determine how well it was working in terms of identifying customer churn.

The confusion matrix was printed to show the performance of the model in terms of true positive, false positive, true negative, and false negative predictions.

Table

Description automatically generated

Table

Description automatically generated

3.  Provide the code used to support the implementation of the logistic regression models.

df['Intercept'] = 1

log\_reg\_results = sm.Logit(df['churn\_dummy'], df[['Children', 'Age', 'Income', 'Outage\_sec\_perweek', 'Yearly\_equip\_failure', 'MonthlyCharge', 'Bandwidth\_GB\_Year',

'gender\_dummy', 'Intercept']]).fit()

print(log\_reg\_results.summary())

# Get the p-values for each variable

pvalues = log\_reg\_results.pvalues

# Select the variables with a p-value below 0.05

significant\_vars = [var for var in pvalues.index if pvalues[var] < 0.05]

# Build the reduced model using only the significant variables

reduced\_model = sm.Logit(df['churn\_dummy'], df[significant\_vars]).fit()

# Print the summary of the reduced model

print(reduced\_model.summary())

X\_train, X\_test, y\_train, y\_test = train\_test\_split(df[significant\_vars], df['churn\_dummy'], test\_size=0.3)

# Fit the logistic regression model on the training data

logistic\_regression = LogisticRegression()

logistic\_regression.fit(X\_train, y\_train)

# Make predictions on the test set

y\_pred = logistic\_regression.predict(X\_test)

# Calculate the confusion matrix

confusion\_matrix = confusion\_matrix(y\_test, y\_pred)

print(confusion\_matrix)

sns.heatmap(confusion\_matrix, annot=True, fmt='d')

report = classification\_report(y\_test, y\_pred)

print(report)Chart, treemap chart

Description automatically generated

Table

Description automatically generated**Part V: Data Summary and Implications**

F.  Summarize your findings and assumptions by doing the following:

1.  Discuss the results of your data analysis, including the following elements:

•  a regression equation for the reduced model

•  an interpretation of coefficients of the statistically significant variables of the model

•  the statistical and practical significance of the model

•  the limitations of the data analysis

2.  Recommend a course of action based on your results.

churn\_dummy = 0.0341 \* MonthlyCharge + -0.0008 \* Bandwidth\_GB\_Year + -0.2377 \* gender\_dummy + -2.4512

The regression equation for the reduced model contains 3 predictor variables (MonthlyCharge, Bandwidth\_GB\_Year, gender\_dummy) which have p-values less than 0.05 and thus considered statistically significant. The coefficients of these predictor variables can be interpreted as the change in the log-odds of the outcome variable for a one-unit increase in the predictor variable, holding all other variables constant. The coefficients can be seen as the magnitude of the association of predictor variables with the target variable, the positive sign suggests an increase in the target variable with an increase in the predictor variable and negative sign suggests a decrease in the target variable with an increase in the predictor variable. The statistical significance of the model can be determined by the likelihood ratio test, which suggests that the reduced model is significantly better than a null model (LLR p-value = 0.000). Practical significance of the model can be determined by checking how well the model is performing in terms of identifying customer churn. However, it is clear from the confusion matrix and classification report that the model is not performing well in identifying customer churn.

The regression equation for the reduced model contains 3 predictor variables (MonthlyCharge, Bandwidth\_GB\_Year, gender\_dummy) which have p-values less than 0.05 and thus considered statistically significant. The coefficients of these predictor variables can be interpreted as the change in the log-odds of the outcome variable for a one-unit increase in the predictor variable, holding all other variables constant. The coefficients can be seen as the magnitude of the association of predictor variables with the target variable, the positive sign suggests an increase in the target variable with an increase in the predictor variable and negative sign suggests a decrease in the target variable with an increase in the predictor variable.

The statistical significance of the model can be determined by the likelihood ratio test, which suggests that the reduced model is significantly better than a null model (LLR p-value = 0.000). Practical significance of the model can be determined by checking how well the model is performing in terms of identifying customer churn.

The confusion matrix and classification report shows that the model has an overall accuracy of 81%, precision of 0.69, recall of 0.59, F1-score of 0.63, it is clear that the model is not performing well in identifying customer churn.

For limitations, the model was not performed well in identifying customer churn, In terms of precision, recall, and F1-score. The data preparation process may have caused a loss of information from the original data set. The variable selection technique used may not have identified all relevant variables, as it only compares the p-values with a threshold and may not account for interactions or other factors. The model evaluation metric only gives us a measure of the goodness of fit of the model and does not take into account the practical significance of the model or the model's ability to make predictions.

Based on the results of the data analysis, it is recommended that:

* Implement changes to pricing plans: The statistically significant predictor variable MonthlyCharge had a positive coefficient, which suggests that increasing monthly charges is associated with an increase in the likelihood of customer churn. The company could consider revising their pricing plans to make them more affordable for customers.
* Improve internet service: The statistically significant predictor variable Bandwidth\_GB\_Year had a negative coefficient, which suggests that an increase in the amount of internet data provided is associated with a decrease in the likelihood of customer churn. The company could consider providing more internet data to customers or increasing the internet speed to improve their service.
* Target marketing efforts towards female customers: The statistically significant predictor variable gender\_dummy had a negative coefficient, which suggests that being female is associated with a decrease in the likelihood of customer churn. The company could consider targeting marketing efforts towards female customers to retain them as customers.
* Monitor customer service: The company could monitor customer service interactions to understand the reason for customer churn, and implement changes to improve customer service.
* Implement retention programs: The company could implement retention programs such as incentives for customers to stay, or offer loyalty programs that rewards customers for staying with the company.
* Use the model to predict customer churn: The model can be used to predict which customers are at a high risk of churning, so the company can target retention efforts at those customers specifically.

**Part VI: Demonstration**

G.  Provide a Panopto video recording that includes *all* of the following elements:

•  a demonstration of the functionality of the code used for the analysis

•  an identification of the version of the programming environment

•  a comparison of the **two** logistic regression models you used in your analysis

•  an interpretation of the coefficients

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b8617ebe-9873-44ff-a942-af8d007faea9>

**H.  Code Sources**

No sources were used for the code.

**I.  Intext References**

No sources were used.