***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.*

Which customer variables are most strongly linked to churn?

*2.  Define the goals of the data analysis.*

The goal of this analysis is to determine which variables among the customer data correlate with customer churn.

***Part II: Method Justification***

*B.  Describe logistic regression methods by doing the following:*

*1.  Summarize****four****assumptions of a logistic regression model.*

Four assumptions made by logistic regression models are as follows:

1. The logit of the dependent variable and the independent variables have a linear relationship.
2. The dependent variable has a binary outcome.
3. Observations are selected independent and randomly within a population.
4. The independent variables are not substantially multicollinear.

*2.  Describe****two****benefits of using Python or R in support of various phases of the analysis.*

Python is an appropriate language for the analysis because for the following reasons:

1. Python has an abundance of libraries (statsmodels, matplotlib, seaborn, scipy, scikitlearn, pandas, numpy) for data science and data manipulation that make it well suited to both performing and visualizing a logistic regression.
2. Python libraries follow common syntax and are generally easy to implement for specific needs.

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

Logistic regression is an appropriate technique for this research question because it models the relationship of a binary variable to multiple influencing variables. The research question asks about the relationship of the binary variable Churn to other influencing variables.

***Part III: Data Preparation***

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

*1.  Describe your data cleaning goals and the steps used to clean the data to achieve the goals that align with your research question including the annotated code.*

The objective of the data cleaning process is to ensure that no columns contain missing or duplicate values. Missing values are detected using the missingno package and matrix function, which displays a chart of all missing values in the dataframe. Duplicate values are detected using .duplicated(). Outliers will be detected using scipy.stats.zscore() on continuous, ordinal columns and dropped if a record has a z score higher than 3 or lower than -3.

The result of the data cleaning process dropped 825 records, all of which were from outliers as the data had no duplicates or missing values.

The code is attached as a file titled “NCina D208 T2.ipynb”

*2.  Describe the dependent variable and*all*independent variables using summary statistics that are required to answer the research question, including a screenshot of the summary statistics output for each of these variables.*

The dependent variable in the dataset is the Churn column. Churn is a binary categorical that measures whether or not the given customer discontinued service in the last month. The independent variables selected were determined based on their correlation to Churn using two statistical tests. Categorical variables were tested for correlation to Churn using a Chi-square contingency test with an alpha of 0.05. Continuous variables were tested for correlation to Churn using a point-biserial R test with an alpha of 0.05. The list of independent variables found using these methods and their P-values are shown below:

A screenshot of a computer program

Description automatically generated

Below are screenshots of the summary statistics for all 20 independent variables found using these methods as well as the summary statistics for the dependent variable Churn. Categorical variables are represented as percent distributions. The three continuous variables are described with the following metrics:

count – Number of entries in the dataframe

mean – Average quantity of the distribution

std – Standard deviation of the distribution

min – Smallest value in the distribution

25% - Values below the lower quartile of the distribution

50% - The median value

75% - Values below the upper quartile of the distribution

max – Largest value in the distribution

A black screen with numbers and numbers

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA computer screen shot of a program

Description automatically generated

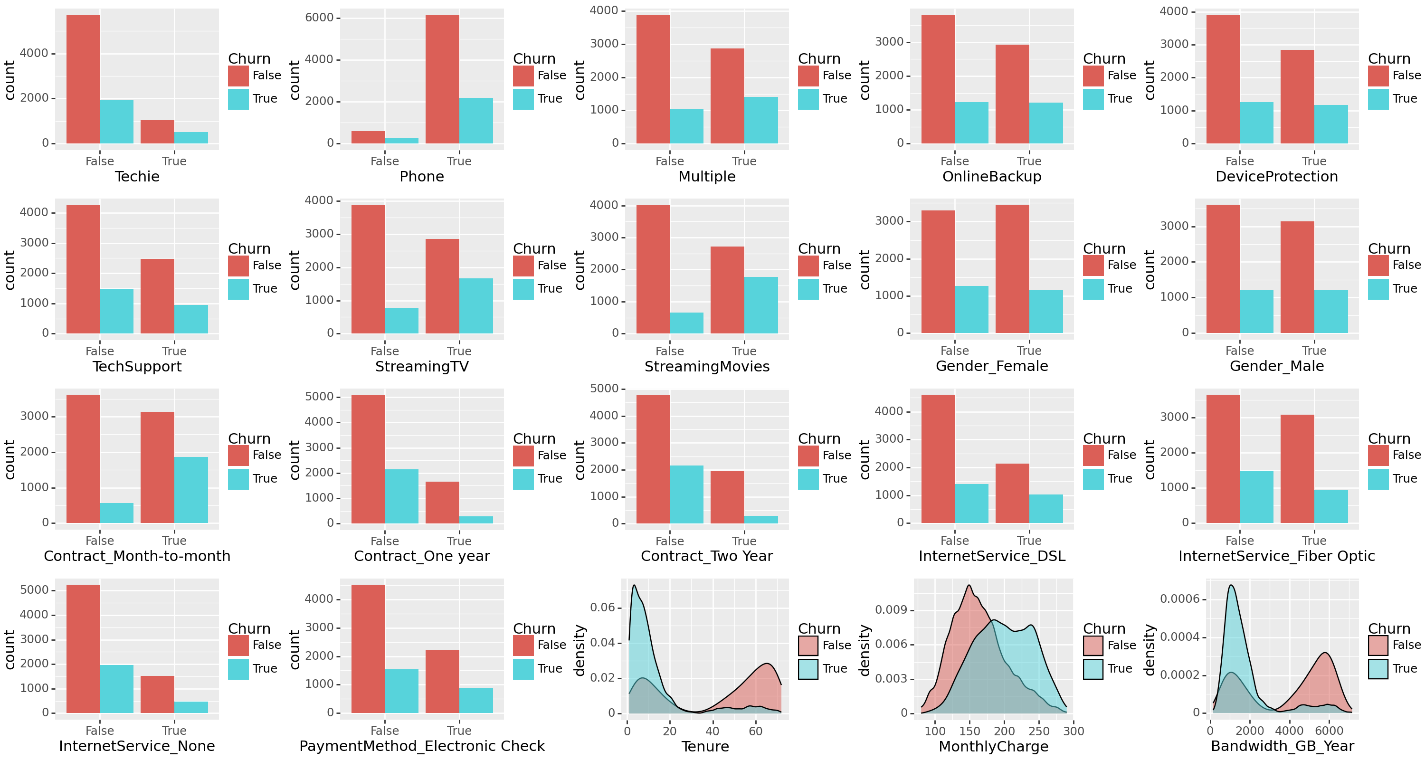
*3.  Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables, including the dependent variable in your bivariate visualizations.*

Univariate visualizations of the variables selected are below:

A screenshot of a graph

Description automatically generated

Bivariate visualizations of the variables selected are below:



*4.  Describe your data transformation goals that align with your research question and the steps used to transform the data to achieve the goals, including the annotated code.*

The goal of data transformation was to make the resulting dataframe easier to work with when performing logistic regression by converting the values from strings to integers. Categorical variables with True/False values were re-expressed on the same column as 0 = False, 1 = True. Categorical variables with multiple non-ordinal values were re-expressed using one-hot encoding to allow performing regression on individual categories within that column. Finally, the original columns that the one-hot encoded columns were created from were dropped from the set.

The code is attached as a file titled “NCina D208 T2.ipynb”

*5.  Provide the prepared data set as a CSV file.*

The prepared data is provided as “NCina D208 T2.csv”

***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*independent variables that were identified in part C2.*

Below is a screenshot of the creation of the logistic regression model and a summary of the model’s properties.

A screenshot of a computer

Description automatically generated

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

Variance inflation factor will be used first to reduce multicollinearity within the model. VIF is a ratio of how strongly the variance of a parameter affects the model when paired with other parameters compared to isolating the parameter. Parameters with high VIF (generally > 10) strongly imply multicollinearity. Eliminating a multicollinear variable reduces redundancy in the model.

To demonstrate the effect of removing multicollinear variables using VIF, the results of multiple VIF operations are shown below, with variables being removed until all VIF values fall below 10. The first image is the initial VIF with no variables removed.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer program

Description automatically generatedA screenshot of a computer program

Description automatically generated

The result of the VIF analysis saw the variables Gender\_Male, Contract\_Month\_to\_Month, InternetService\_DSL, MonthlyCharge, and Bandwidth\_GB\_Year removed.

The categorical variables in this set are due to one hot encoding producing multicollinear variables when all columns resulting from one column being encoded are included in a regression model. Any individual column can be explained by the remaining column’s values and is therefore multicollinear. Removing just a single column from each one hot encoded set remedies the issue.

To reduce dimensionality further, sequential forward selection was used as a feature selection method. Sequential forward selection examines the accuracy of a model to select the most relevant features that best predict the dependent variable. A total of 5 features were selected, as the prediction score stopped increasing by a significant amount with more features added.

A printout of the sequential forward selection features and prediction score can be found below:



*3.  Provide a reduced logistic regression model that follows the feature selection or model evaluation process in part D2, including a screenshot of the output for each model.*

Below is a screenshot of the creation of the reduced logistic regression model and a summary of the model’s properties.

*A screenshot of a computer

Description automatically generated*

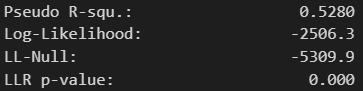
*E.  Analyze the data set using your reduced logistic regression model by doing the following:*

*1.  Explain your data analysis process by comparing the initial logistic regression model and reduced logistic regression model, including the following element:*

*•   a model evaluation metric*

Screenshots of the model evaluation metrics that will be used for the model comparison are shown below. The reduced model is on the right.

*A black screen with white text

Description automatically generated*

Generally, high pseudo R-squared values and lower Log-likelihoods equate to better models. However, the change in each of the two statistics is acceptable. The reduced model drops 15 variables and has comparable statistics. Models with fewer independent variables that produce similar results will always be better models.

*2.  Provide the output and*all*calculations of the analysis you performed, including the following elements for your reduced logistic regression model:*

*•   confusion matrix*

*•   accuracy calculation*

The confusion matrix and accuracy calculation for the reduced model is listed below:

*A black screen with white text

Description automatically generated*

*3.  Provide an executable error-free copy of the code used to support the implementation of the logistic regression models using a Python or R file.*

Code included as “NCina D208 T2.ipynb”

***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 the coefficients of the reduced model*

*•   the statistical and practical significance of the reduced model*

*•   the limitations of the data analysis*

The regression equation for the reduced logistic model is as follows:

logit(p) = -0.7585 + StreamingTV \* 2.3631 + StreamingMovies \* 2.7961 – Contract\_One\_Year \* 2.5945 – Contract\_Two\_Year \* 2.6767 – Tenure \* 0.0886

The intercept in this case is not meaningful, as it is negative and probability ranges from 0 to 1. The coefficient for each variable translates to the percentage chance that customer will churn, with positive values increasing the chance and negative values decreasing the chance.

According to the model, the variables most associated with customers churning are StreamingMovies, and StreamingTV. StreamingMovies has a slightly more pronounced impact that StreamingTV, with a coefficient of 2.7961 compared to 2.3631. This seems to suggest that customers who use the service to stream movies and TV will switch to a different service more frequently than customers who do not. This could be caused by the quality of the streaming service or other streaming services offering better deals or better selections for those customers.

The variables most associated with customers *not* churning are Tenure, Contract\_One\_Year, and Contract\_Two\_Year. Tenure is a continuous variable, so for each month of tenure the chance to churn decreases by 0.0886. By the same logic, the contract variables both indicate that the longer a customer’s contract is, the less likely they are to churn. Contract\_One\_Year and Contract\_Two\_Year have similar effects on churn, with coefficients -2.5945 and -2.6767 respectively. In summary, customers with higher Tenure are less likely to churn, which is a reasonable assumption. Additionally, customers with longer contracts are less likely to churn.

The model is believed to be statistically significant, as the confusion matrix shows the model is fairly accurate and the p values of the individual variables are far below the acceptable alpha level (0.05). The variables chosen do not appear to be multicollinear according to the VIF analysis.

The practical utility of the model is discovering which customers are more or less likely to churn, and guiding decision making on how to increase customer retention. Customers who sign longer contracts and have longer tenures tend to not churn, whereas customers who stream movies and TV have a higher chance to churn.

The main limitation with this model is the quantity of records that were dropped from the original set. The data preparation process removed nearly 10% of the original data set, and this could have skewed the model. However, all removed records were due to trimming outliers.

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

The model suggests that customers who stream movies and television have a significant disposition toward churning. It would be beneficial to investigate the cause of this trend, whether that be the quality of the service the company currently offers not being enough for customers that prioritize streaming, the current streaming market offering more competitive price models, the selection the company offers not being expansive or up to date enough for clients that prioritize streaming, or any other cause not accounted for here.

***Part VI: Demonstration***

*G.  Provide a Panopto video recording that includes the presenter and a vocalized demonstration of the functionality of the code used for the analysis of the programming environment, including the following elements:*

*•   an identification of the version of the programming environment*

*•   a comparison of the initial logistic regression model you used and the reduced logistic regression model you used in your analysis*

*•   an interpretation of the coefficients of the reduced model*

Panopto Video Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=649ff68a-1067-4b0d-a3de-b187002baae2>

*H.  List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable.*

No web sources used.

*I.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.*

No text sources used.