**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.

What factors contribute to customers leaving the company?

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.

Determining the factors of what influence customers to no longer require our services. The target variable is Churn with the predictor variables being Age, Income, Contacts, Yearly\_equip\_failure, Tenure, MonthlyCharge, Bandwidth\_GB\_Year, Item1 – Item 7, Gender, Contract, Techie, and TechSupport.

**Part II: Method Justification**

B.  Describe logistic regression methods by doing the following:

1.  Summarize the assumptions of a logistic regression model.

According to the Complete Dissertations by Statistics Solutions (*Assumptions of logistic regression* 2021), logistic regression has the following assumptions:

* Requires a large sample size
* Requires little to no multicollinearity for independent variables
* Response variable is binary

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.

Using Python for predictive model is beneficial in my case since this is the primary programming language I use from a day-to-day basis. There are multiple Python libraries that can be used in the different phases of analysis. For data manipulation/cleaning, Pandas can be used. For data visualization, matplotlib is a popular tool. Numpy is known for its various mathematical operations. Sci-kit learn is the machine learning library of choice that I will be using in this phase of the analysis.

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

Since we are predicting the likelihood of a customer leaving (0,1) based on several features, logistic regression is the better choice.

**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.

* Handling missing data
  + If there are any missing data detected, data will be imputed using either median, mean, or mode of the column.
* Outliers
  + If outliers are detected by utilizing creation of histograms, then they will be removed and replaced using imputation.
* Qualitative feature encoding
  + Categorical variables will be replaced with dummy variables.

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.

* Customers have an average of 2 children with the max being 10. (Children)
* The average age of a customer is 53 with the minimum being 18 and max 89. (Age)
* Reported income of customers has an average of 39,806. (Income)
* Customers have contacted technical support an average of .99 times with the max being 7. (Contacts)
* There has been an average of .398 yearly equipment failures with a standard deviation of .635 and a maximum occurrence of 6. (Yearly\_equipment\_failure)
* Customers have stayed with the company for an average of 34 months (Tenure)
* MonthlyCharge ranges from 79 dollars to 290 dollars. (MonthlyCharge)
* An average of 3392 GBs have been used by customers yearly. (Bandwidth\_GB\_Year)
* Item1 – Item7 have a scale of 1-8 with 1 being least important and 8 being most important on various topics. The average value selected for each of the survey questions is 3.5
* There are more Female customers than Male (5025). (Gender)
* Most customers (5456) are signing up for Month-to-month plans. (Contract)
* 6250 out of 10000 customers are declining tech support. (TechSupport)
* 8321 customers self-identify themselves as technically inclined. (Techie)

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

After importing the dataset, I checked for missing data by using the .isna() function. Dummy values were created for categorical variables in the Churn dataset that I considered to be features that a customer will sign up for. I removed a dummy column for each variable to prevent “dummy trap”. Histograms were then created for the initial model variables to pinpoint outliers, which I noted that for this dataset did not exist.

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

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Table

Description automatically generated

5.  Provide a copy of the prepared data set.

**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

Table

Description automatically generated

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.

To reduce the initial model, I will be utilizing the p-values and the correlation between the predictor variables and Churn. If the p-value is less than the alpha value (0.05) of a feature than that feature is statistically significant and should be included in the model (*Interpret the key results for binary logistic regression*). A correlation matrix was created and visualized to represent the relationship between several features and if it impacts the likelihood of a customer leaving the company or not(Upadhyay, *What is correlation?* 2020).

3.  Provide a reduced logistic regression model.

Table

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 and reduced logistic regression models, including the following elements:

•  the logic of the variable selection technique

•  the model evaluation metric

By using the correlation matrix and p-value feature selection method, I was able to reduce the initial model from 22 variables to 8. I removed all features that was not statistically significant (a p-value more than .05). The correlation matrix was a useful tool to visually determine which variables had the most positive/negative impact on the target variable, Churn. To compare the fit of the models, I created a confusion matrix and focused on the accuracy metric. The initial model had an accuracy of .8954 with 22 variables and the reduced model had accuracy rate of .8951 with only 8 variables.

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

Initial

Text, letter

Description automatically generated

Reduced

Text

Description automatically generated

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

**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

Y = -3.9627 + -3.1813(Contract\_OneYear) + -3.2944(Contract\_TwoYear) + 0.0029(Bandwidth\_GB\_Year) + -0.3411(Tenure) + 0.0143(MonthlyCharge) + 1.8897(StreamingMovies\_Yes) + 1.4883(StreamingTV\_Yes) + 0.8982(Multiple\_Yes)

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

* For every 1 unit, Contract\_OneYear will decrease 3.1813 units.
* For every 1 unit, Contract\_TwoYear will decrease 3.2944 units.
* For every 1 unit, Bandwidth\_GB\_Year will increase .00029 units.
* For every 1 unit, Tenure will decrease .3411 units.
* For every 1 unit, MonthlyCharge will increase by .0143 units.
* For every 1 unit, StreamingMovies\_Yes will increase by 1.8997 units
* For every 1 unit, StreamingTV\_Yes will increase by 1.4883 units.
* For every 1 unit, Multiple\_Yes will increase by .8982 units.

•  the statistical and practical significance of the model

The reduced model represents only 59.76% of the variance and has an accuracy of 89%. With these stats, a determination of practical significance is still unknown. I would not present these findings to the stakeholders.

•  the limitations of the data analysis

The amount of data that I have been presented with is a definite limitation. If this is a popular telecommunications company, then I would expect to have more than 10,000 rows of data. Especially, if my analysis is going to be presented to stakeholders.

2.  Recommend a course of action based on your results.  
I would consult with the data gathering team to inquire if there is more fields/features and data for me to work with to determine the key factors that impacts the customer’s decision to stay or not.

Sources/Third-Party Code

*Assumptions of logistic regression*. Statistics Solutions. (2021, August 11). Retrieved October 21, 2021, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/.

*Interpret the key results for binary logistic regression*. Minitab Express. (n.d.). Retrieved October 11, 2021, from https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/binary-logistic-regression/interpret-the-results/key-results/.

Szabo, B. (2020, May 26). *How to create a Seaborn Correlation Heatmap in python?* Medium. Retrieved October 11, 2021, from https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e.

Upadhyay, A. (2020, August 9). *What is correlation?* Medium. Retrieved October 11, 2021, from https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47.