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.**

   Can we predict the likelihood of customers to churn based on other data in the data set?

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.**

   The objectives of this analysis are:

   - Determine which, if any, variables are related to Churn.
   - Determine the business interest in any related variables.
   - Create a logistic regression model based on those variables.
   - Evaluate the model to determine if it can answer the business question.

Part II: Method Justification

B. Describe logistic regression methods by doing the following:

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

   Instead of predicting an outcome, logistic regression predicts the probability of class or event outcomes, such as pass/fail, yes/no, win/lose, leave/stay. It is the logarithm of the probability of reaching 1. Because the target variable is categorical, logistic regression is based on the Bernoulli distribution, not Gaussian. Predicted values must be a range of nominal values and the predictors cannot be highly correlated. (Sewell, 2021)

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.**

   I will be using python as my selected programing language. Libraries include:

   - Pandas to import the file and clean the data
   - Seaborn to create visualizations
   - Matplotlib to create visualizations
   - Statsmodels to create my regression model and summary
   - Numpy to call the model odds ratios
   - Sklearn to produce a confusion matrix and model accuracy score

   In addition to utilizing these libraries, I chose python because it is the tool my current employer prefers.

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

   Because we want to predict the likelihood a customer would churn based on other data, instead of predicting a specific value, and our target variable is categorical, logistic regression is appropriate.

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.**

   The goal of the data preparation is to create a data frame with only data that is relevant and usable in the logistic regression model. I will use the pandas library to change datatypes, manipulate actual text data into numeric data, create dummy variables, drop columns, and filter the data frame.

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.**

   The target variable is Churn_Yes, and has a mean of 0.27, a standard deviation of 0.44, a minimum of 0.00 and a maximum of 1.00. The predictor variable CaseOrder has a mean of 5000.50, a standard deviation of 2886.90, minimum of 1.00 and maximum of 10000.00. The predictor variable Tenure has a mean of 34.53, a standard deviation of 26.44, a minimum of 1.00 and a maximum of 72.00. The predictor variable MonthlyCharge has a mean of 172.62, a standard deviation of 42.94, a minimum of 79.98 and a maximum of 290.16. The predictor variable Bandwidth_GB_Year has a mean of 3392.34, a standard deviation of 2185.29, a minimum of 155.51, and a maximum of 7158.98. The predictor variable Multiple_Yes has a mean of 0.46, a standard deviation of 0.50, a minimum of 0.00, and a maximum of 1.00. The predictor variable StreamingTV_Yes has a mean of 0.49, a standard deviation of 0.50, a minimum of 0.00, and a maximum of 1.0. StreamingMovies_Yes has a mean of 0.50, a standard deviation of 0.50, a minimum of 0.00, and a maximum of 1.0.

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

   I will first import the original dataset. Then I will change object type categorical data to category codes, then create dummy variables for object type categorical data that have only “Yes” and “No” values. Then I will drop all the “No” dummy variables to prevent collinearity. Subsequently, I will drop variables that I deem unnecessary for business purposes. Using a Pearson Correlation heatmap, I will then choose the variables that are correlated to Churn_Yes for my initial regression data frame.

   For annotated code, see Keim D208 Task Two Code document.

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

   See Keim D208 Task Two Code document for univariate and bivariate visualizations of all variables included in the logistic regression model.

5. **Provide a copy of the prepared data set.**

   See Keim D208 Task Two Prepared Data document.

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**

   See Keim D208 Task Two Code document.

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.**

   I used a Pearson Correlation heatmap and pairwise correlation analysis to determine which variables were highly correlated with each other. I dropped variables that had $r > 0.5$ in the pairwise correlation analysis. I also considered business relevance when reducing variables. I used the accuracy of the fitted model to evaluate the model.

3. **Provide a reduced logistic regression model.**

   See Keim D208 Task Two Code document. The predictor variables chosen are MonthlyCharge, Bandwidth_GB_Year, Multiple_Yes and StreamingTV_Yes. The target variable for logistic regression must be categorical and is Churn_Yes.

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**

     For my initial variable selection, I used a Pearson Correlation heatmap to determine which variables correlated to the target, Churn_Yes. For the reduced model, I again used a Person Correlation heatmap to determine collinearity between predictor variables then continued to reduce based pairwise correlation analysis and business relevance

   - **the model evaluation metric**

     I used the accuracy of the fitted model as the model evaluation metric. The initial model had an accuracy of 0.8505. The reduced model had an accuracy score of 0.8314.

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

  The outputs, confusion matrices, and reduced model predictions can be found in the Keim D208 Task Two Code document.

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

   See Keim D208 Task Two Code document.

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**

     Log(0.01/1-0.01)=1.03+1.00x1+1.02x2+1.81x3

   - **an interpretation of coefficients of the statistically significant variables of the model**

     The variable MonthlyCharge has an odds ratio of 1.03. For one unit increase in MonthlyCharge, we would expect about a 1.03 increase in the odds of churn occurring. The variable Bandwidth_GB_Year has an odds ratio of 1.00. For one unit increase in Bandwidth_GB_Year, we would expect about a 1.00 increase in the odds of churn occurring. The variable Multiple_Yes has an odds ratio of 1.02. For one unit increase in Multiple_Yes, we would expect about a 1.02 increase in the odds of churn occurring. The variable StreamingTV_Yes has an odds ratio of 1.81. For one unit increase in StreamingTV_Yes, we would expect about a 1.81 increase in the odds of churn occurring.

   - **the statistical and practical significance of the model**

     Statistically, the model shows that the likelihood of churn occuring can be predicted by MonthlyCharge, Bandwidth_GB_Year, Multiple_Yes, and StreamingTV_Yes. Practically, this means that if we examine a customer's monthly bill, their average bandwidth per year, if they have multiple lines, and if they purchased streaming TV, we can predict the likelihood that they will or will not churn.

   - **the limitations of the data analysis**

     The limitation of this analysis is that correlation and likelihood don't tell us about causation or behavior of customers. Because there is an issue of multicollinearity, we also cannot include all variables available to us in a regression model, leaving some parts of the customer's picture out of the analysis.

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

   The analysis says that customers with a high monthly bill, a high yearly average bandwidth usage, more than one phone line, and an add on service for streaming TV, are more likely to churn than those who do not. I recommend targeting these customers with specials and incentives to stay with the company.

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

Citation for acquiring data: (Western Governor's University, 2021)**.** Citation for third party code: (Bedre, 2021)

Bibliography

Bedre, R. (2021, October 9). Logistic regression in Python (feature selection, model fitting, and prediction). Retrieved from https://www.reneshbedre.com/: https://www.reneshbedre.com/blog/logistic-regression.html

Sewell, D. W. (2021, October 8). D208 predictive modeling webinar.

Western Governor's University. (2021, 09 16). WGU Performance Assessment. Retrieved from tasks.wgu.edu: https://tasks.wgu.edu/student/001234696/course/20900013/task/2786/overview