Part I: Research Question

A1. Summarize one Research Question for Real-World Organizations

Can we predict which customers are more likely to Churn based on several predictor variables?

A2. Objective and Goals of Analysis

The goal and objective of this analysis are to use Logistic Regression to help us further understand why customers churn and what variables affect their decision to churn. By understanding these variables, we hope to be able to predict whether a customer will churn. For our analysis, we will be using several predictor variables such as “Sum_services”, “Area”, “Employment”, etc, and whether it is a good predictor of customer churn rate.

Part II: Method Justification

B1. Logistic Regression Summary

As stated on the website of IBM (What is Logistic Regression?, 2021), logistic regression is a type of analysis that can help predict the likelihood of an event happening or choices being made. Per Z. (2020, October 13), the assumptions of a logistic regression model are:

1. The response variable is Binary.
2. Observations are independent of each other and there is no duplication.
3. No multicollinearity among predictor variables.
4. No extreme outliers.
5. There is a linear relationship between predictor variables and the logit of the response variable.
6. Large enough sample size.

B2. Benefits of tools

For this analysis, we will be using Python as it has a wide array of libraries and packages that is suited for our purpose. Some of the packages and libraries necessary are numpy, pandas, statsmodels, scikit-learn, and seaborne among other libraries. The versatility of Python and its relationship to statistics and machine learning also make it a perfect fit for our use in this analysis (Talab, Z., 2021). Python will allow us to prepare our data during the data cleaning process using numpy and pandas. We can then use it to further plot our data during the exploratory analysis using libraries like seaborne to visualize our predictor and response variables and to detect outliers. Finally, Python with its diverse libraries will allow us to perform our necessary predictive analysis during our data modeling phase using statsmodel and scikit-learn.

B3. Why is Logistic Regression Appropriate?

Logistic Regression is appropriate for our use because it provides us with a perfect analysis for our binary "Churn" dependent variable against our independent variables "Sum_services", "Area", and "Employment" etc. Our objective is to understand whether we can predict a binary outcome "Churn", using predictor variables provided in our data. With logistic regression, it provides us with the necessary predictive modeling for our task.

Part III: Data Preparation

C1. Data Preparation Goals and Data Manipulation

For our data preparation goals, we want to make sure our variables are properly recoded, missing data imputed or removed, and that outliers have been reduced within our data. We have completed most of these tasks during our data cleaning step in the previous courses so we will be repurposing the code for this process. For reference, we will recode our binary categorical variable with a 1 or 0 so that it can be entered into our Logistic regression analysis. We will also recode predictor variables, "Contract",

"Area" and "Employment" into an ordinal variable for our univariate/bivariate analysis and then later into dummy variables for our Logistic regression analysis. Finally, we will update the survey question columns with their appropriate name "Timely_Response", "Timely_Fixes", "Timely_Replacement" etc. In the case of missing values, we will either impute a 0 for categorical variables or use central tendency values for continuous variables. For outliers, we will create a Z-Score for our continuous variable like "Sum_ services" column and remove any data with Z-Scores higher than 3 and less -3.

C2. Discuss the Summary Statistics

In our raw data, we have a total of 10,000 observation rows with 52 variable columns. After the data cleaning process, our dataset was reduced to 9978 observations with 1 response variable, 12 predictor variables of which were later expanded to 18 after creating dummy variables. For reference, our target variable "Churn" has a mean of 0.26 with a standard deviation of 0.44. Our predictor variables are "Area", "Employment", "Income", "Contract", "Sum_services", "Timely_Fixes", "Timely_Replacement", "Reliability", "Options", "Respectable_Response", "Courteous_Exchange", and "Evidence_of_active_listening". Before we do our initial analysis, we will create dummy variables for categorical variable "Area", "Employment, "Contract" and then remove one level to avoid dummy variable trap. These will be "Dummy_Urban", "Dummy_Suburban", "Dummy_Part Time", "Dummy_Retired", "Dummy_Student", "Dummy_Unemployed", "Dummy_One year", "Dummy_Two Year".

Part IV: Analysis

D1. Justify Model Reduction, Variable Selection Procedure, and Model Evaluation Metric

Based on our research question, we want to be sure we can statistically identify which variable will be a great predictor of customer churn and identify to see if our model is a good fit. For our variable selection, we will use a multicollinearity heatmap, VIF (variance inflation factor), and a variable's p-

value. To evaluate our model, we will use R-Square, verify the variables using the forward stepwise variable selection function in combination with the AUC score, and then plot it.

In choosing our variables for analysis, we will first plot out a correlation heatmap to identify highly correlated variables and then use that in combination with VIF to indicate variables with high multicollinearity above 5 and remove them. Next, we remove any variables whose p-values are greater than 0.05 as we only want statistically significant variables. In our initial model, we can see that there was no observed multicollinearity but there were variables with high p-values. Since we only want statistically significant predictor variables with p-values less than 0.05 (so that we can reject the null hypothesis and indicate there is a difference between the response and predictor variable), all variables with p-values higher than alpha were removed. Those predictor variables were “Income”, “Timely_Response”, “Timely_Fixes”, “Timely_Replacement”, “Reliability”, ”Options”, “Respectable_Response “, “Courteous_Exchange”, “Employment”, “Evidence_of_active_listening”, “Dummy_Suburban”, “Dummy_Urban”, “Dummy_Part Time”, “Dummy_Retired”, ‘Dummy_Student”, and “Dummy_Unemployed”

During our model evaluation, we will calculate R-square to determine if our predictor variable can explain the variance in our response variable “Churn”. We then do further evaluations on all variables using the forward stepwise variable selection function which helps us evaluate their AUC score and verify our variable selection process. Then by plotting an AUC model, we can further reduce the number of variables needed. Finally, we can then determine our best fit predictor variables for our analysis.

D2. Comparison of the Initial, Reduced Model and Model Metric Evaluation

The initial model had 18 independent variables and one response variable “Churn”. We then removed the variable “Income”, “Timely_Response”, “Timely_Fixes”, “Timely_Replacement”, “Reliability”, ”Options”, “Respectable_Response “, “Courteous_Exchange”,

"Evidence_of_active_listening", "Dummy_Suburban", "Dummy_Urban", "Dummy_Part Time", "Dummy_Retired", 'Dummy_Student", and "Dummy_Unemployed" using a combination of forward stepwise variable selection and eliminating variables with p-values greater than 0.05. We were left with "Sum_services", "Dummy_One year", and "Dummy_Two Year". There was no observed multicollinearity among the variables in both models. The initial model has an AUC score of 0.66 and the reduced model has an AUC score of 0.75 indicating a better model. After establishing the AUC score, we used a forward stepwise variable selection in verifying the three statistically significant variables ($p<.05$). We first set the forward stepwise variable selection to 5 and after performing the initial variable selection reduction, we received an AUC score of 0.75. We later expanded upon the variable selection by plotting the AUC model where we were able to identify that with just 3 variables, "Sum_services", "Dummy_One year", and "Dummy_Two Year", we can achieve the best possible score with the least number of variables. With these three variables, the model showed a 75% accuracy in predicting Churn among customers which is considered "acceptable discrimination" by Hosmer and Lemeshow in Applied Logistic Regression (H, 2013). The R-Squared was low on both the initial model (r-squ=0.143) and reduce model (r-squ=0.139) though this may be irrelevant due to the nature of this analysis (Z, 2020).

Part V: Data Summary and Implications

E1. Results of Analysis

- The Regression equation of the reduce model is y = (-2.89 + (0.50 * Sum_services) + (-1.39 * Dummy_One year) + (-1.65 * Dummy_Two Year)).
- The coefficient for "Sum_services" is 0.50 which indicates that as the number of services goes up by 1 unit, the odds of a customer churn increases by 64%. The coefficient of "Dummy_One year" is -1.39 which indicates that a customer who signs a contract for one year has a 75% decrease odds of churning when compared to a month-to-month customer. The coefficient of

Dummy_Two year is -1.65 which indicates that a customer who signs a contract for two years has an 80% decrease odds of churning when compared to a month-to-month customer.

- The p-values of our variables, “Sum_services”, “Dummy_One year”, “Dummy_Two Year” for our reduced model was statistically significant ($p<0.05$). Therefore, we reject the null hypothesis indicating that there is a difference between the predictor variables and the response variable. The model also has some practical significance in that we can expect the longer a customer signs up for a contract, they will be locked into the contract and therefore are less likely to not churn. Since practical significance is subjective, we can say that the more services a customer signs up for, the higher their bill is, and therefore more likely for them to churn to reduce their total overall personal bills.
- One of the limitations of this Data Analysis is that with our low R-square (r-squ=0.139) our predictions may not be as accurate though according to Z. 2020 (April 14), a low r-square may not be as impactful due to the nature of our analysis as human decisions are understandably hard to predict. Another limitation is that we couldn’t find more variables that can statistically predict customer churn so reiteration with other different variables may be necessary. We started with eighteen possible predictor variables and after our variable selection, we were left with 3 predictor variables. Also, since some data were imputed with a central tendency value, it may have affected our predictions. Other limitations include data that may have been improperly collected or stored.

E2. Recommended Course of Action

Based on our results, we can understand that there is a linear relationship between customer churn and the sum of their services along with their contract duration (one-year, two-year). With that, we can expect to decrease the odds of a customer churning, by reducing the number of services a customer signs up for and increasing the duration of each customer’s contract (when compared to a month-to-

month customer). To accomplish these tasks, we may have to investigate a combination of both variables including bundling up services together into longer contracts to decrease the odds of customer churn and at the same time, sell more services and increase our profits. As with most analyses, reiteration with more variables or reclassifying some variables may be needed for a more thorough understanding of customer churn.

Part VI. Supporting Documents

G. Video

H. Sources for Third-Party Code

Aruchamy, V. (2021, September 29). *How To Plot Confusion Matrix In Python And Why You Need To?* Stack Vidhya. https://www.stackvidhya.com/plot-confusion-matrix-in-python-and-why/#:%7E:text=To%20plot%20the%20confusion%20matrix%20with%20percentages%2C%20first%2C,can%20sum%20all%20values%20in%20the%20confusion%20matrix.

Building the AUC curves | Python. (2021). DataCamp. https://campus.datacamp.com/courses/introduction-to-predictive-analytics-in-python/forward-stepwise-variable-selection-for-logistic-regression?ex=12

D. (2020a, May 17). *Example of Logistic Regression in Python*. Data to Fish. https://datatofish.com/logistic-regression-python/

Intellipaat, P. (2021, September 1). *Introduction to Confusion Matrix in Python Sklearn*. Intellipaat Blog. https://intellipaat.com/blog/confusion-matrix-python/#:%7E:text=%20Implementing%20Confusion%20Matrix%20in%20Python%20Sklearn%20%E2%80%93,Create%20and%20train%20the%20model.%20%20More%20

Python Logistic Regression with Sklearn & Scikit. (2021). DataCamp Community. https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python

sklearn.metrics.roc_auc_score. (2021). Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

Split Train Test - Python Tutorial. (2021). Python Tutorial. https://pythonbasics.org/split-train-test/

I. Sources

Chelliah, I. (2021, September 8). *Univariate and Bivariate Analysis using Seaborn*. SheCanCode. https://shecancode.io/blog/univariate-and-bivariate-analysis-usingseaborn

H. (2013). *Applied Logistic Regression, 3rd Edition* (3rd ed.). Wiley.

Interpret Logistic Regression Coefficients [For Beginners]. (2021). Quantifying Health. https://quantifyinghealth.com/interpret-logistic-regression-coefficients/

Talab, Z. (2021, November 12). *Benefits of Python for AI*. Developer.Com. https://www.developer.com/languages/benefits-of-python-for-ai/#:%7E:text=In%20addition%20to%20its%20versatility%2C%20Python%20is%20a,platforms%3B%20Windows%2C%20MacOS%2C%20Linux%2C%20Unix%2C%20and%20many%20more.

What is Logistic regression? | IBM. (2021). IBM. https://www.ibm.com/topics/logistic-regression

Z. (2020, April 24). *What is a Good R-squared Value?* Statology. https://www.statology.org/good-r-squared-value/#:%7E:text=How%20high%20an%20R-

squared%20value%20needs%20to%20be,if%20there%20is%20extreme%20variability%20in%20the%20dataset.

Z. (2020, October 13). *The 6 Assumptions of Logistic Regression (With Examples)*. Statology. https://www.statology.org/assumptions-of-logistic-regression/

Z. (2021, September 9). *What is Considered a Good AUC Score?* Statology. https://www.statology.org/what-is-a-good-auc-score/