**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 determine why customers are high risk for churn or which features are a crucial influence on customer churn?

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.

Stakeholders in the company will benefit from this analysis because it will help them prepare, design, and add or remove services that will benefit the company by decreasing the churn rate.

**Part II: Method Justification**

B.  Describe logistic regression methods by doing the following:

1.  Summarize the assumptions of a logistic regression model.

* That the dependent variable is binary or ordinal, which the dependent variable is Churn so it is binary so will need binary logistic regression.
* Observations to be independent of each other and shod not come from repeated measurement or matched data.
* Little to no multicollinearity with the independent variables. These should not be highly correlated.
* Should have a large sample size with a minimum of 10 cases.
* Should have independent variables linearly related to the log odds

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.

Python is used for the programming language because of the number of resources, libraries, and support it provides. Also, Python language has easy to read syntax as well as writing, and very simple to get up and running. Furthermore, Google Colab for the programming environment where the Python code will be written and executed. Google Colab has a very simplistic interface that was design for data analysis, and strong support service for questions that may present itself during my development.

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

Since the dependent variable of Churn is a binomial “Yes or No”, it makes logistic regression an appropriate technique to use, to solve our initial question. Also, trying to figure out what is the chance that a customer will churn based on the independent variables that include, children, age, income, tenure, monthly charges, etc. This will help improve the probability of customer churn by increasing and decreasing different independent variables and figure out if they cause a positive or negative effect on the relationship to the dependent variable.

**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.

* Loading the csv file
* Check if data is cleaned
* Check columns of the data
* Fix all survey columns to appropriate names
* Check outliers
* Check for any missing values or nulls, and misspellings
* Create dummy variables for categorical columns
* Save copy to my google drive for backup purposes

After loading the data into the data frame, I will begin analyzing the data making sure that there are no missing values that I may have to take care of. Also, making sure all columns are properly named.

Some columns that were renamed were the discrete ordinal predictor variables Survey columns:

1. Item1: Timely response
2. Item2: Timely fixes
3. Item3: Timely replacements
4. Item4: Reliability
5. Item5: Options
6. Item6: Respectful response
7. Item7: Courteous exchange
8. Item8: Evidence of active listening

Furthermore, removing non-meaningful data columns from the dataset which included the columns:

1. CaseOrder
2. Customer\_id
3. Interaction
4. UID
5. City
6. State
7. County
8. Zip
9. Lat
10. Lng
11. Population
12. Area
13. TimeZone
14. Job
15. Marital
16. PaymentMethod

Next, have to change the categorical binary predictors to 1’s and 0’s which included the columns:

1. Gender
2. Techie
3. Contract
4. Port\_modem
5. Tablet
6. InternetService
7. Phone
8. Multiple
9. OnlineSecurity
10. OnlineBackup
11. DeviceProtection
12. TechSupport
13. StreamingTV
14. StreamingMovies
15. PaperlessBilling

Another thing is that we drop the first column because this is redundant since we can use 1 column to know if it is a yes or no “1 or 0”. Also, contract column consisted of 3 columns (month-month, one-year, two year) so even dropping that first column the other 2 columns will still allow us to figure which service the customer has. I also, dropped the column Gender\_Nonbinary because I noticed it just gave all 0’s which was not meaningful.

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 original data contained 10000 records and 50 columns, this model was then reduced by removing some of the categorical columns and records. Those columns removed were: CaseOrder, Customer\_id, Interaction, UID, City, State, County, Zip, Lat, Lng, Population, Area, TimeZone, Job, Marital and PaymentMethod. Once those columns were drop left with 34 categorical and numerical columns, but we then wanted all columns numerical to work with one set value of the data. So, focusing on binomial columns, these columns consisted of either “yes/no” or “male/female” values, which we converted to “1,0” values by using pandas get\_dummies function to handle this. Also, checking if all columns are fulfilled and the data did not contain any empty or null values, which there were not in this case all data seemed to be cleaned and complete. Next, I looked at boxplots, displots, and histograms for central tendency, this showed that some columns no longer contained outliers, for example Bandwidth\_GB\_Year, Age, Tenure, and MonthlyCharge. Furthermore, some columns still contained outliers in the clean data such as the Email, Contacts, Children, and Income columns.

When looking at linearity notice that Tenure and Bandwidth\_GB\_Year a strong linear relationship with each other on lmplot graph. Another thing was describing the data to look at averages and standard deviations of columns. This allowed to see the average age of the customers “53”, average children “2”, average income “39806”, and Tenure of “34” etc. Furthermore, the deviations of 20, 2, 28199, 26.

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

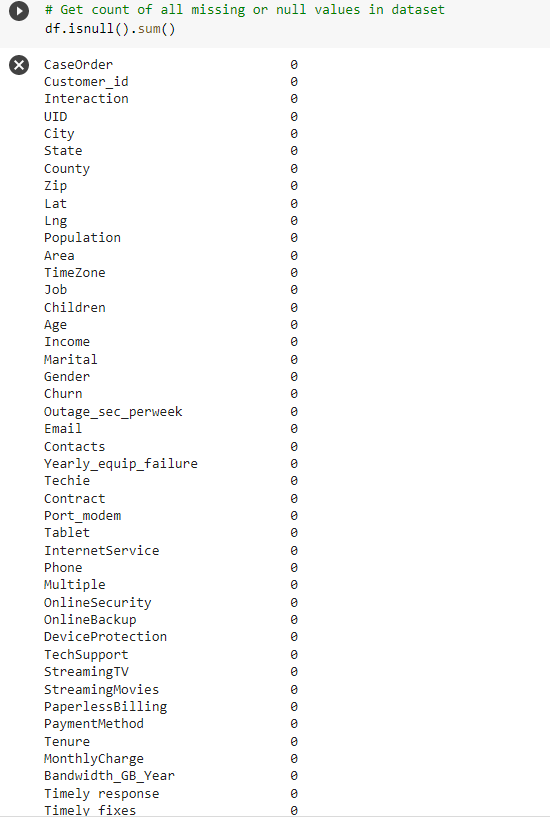
* Import libraries
* Read csv and store in a data frame
* Get an idea of the columns and rows
* Give appropriate names for survey columns
* Describe the data for stats information
* Drop columns non-meaningful
* Check for empty values or null
* Create graphs for columns using displots, lmplots, boxplots, histograms
* Create dummy variables for the columns
* Univariate and bivariate visualizations

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated



Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

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

Diagram

Description automatically generated with low confidence

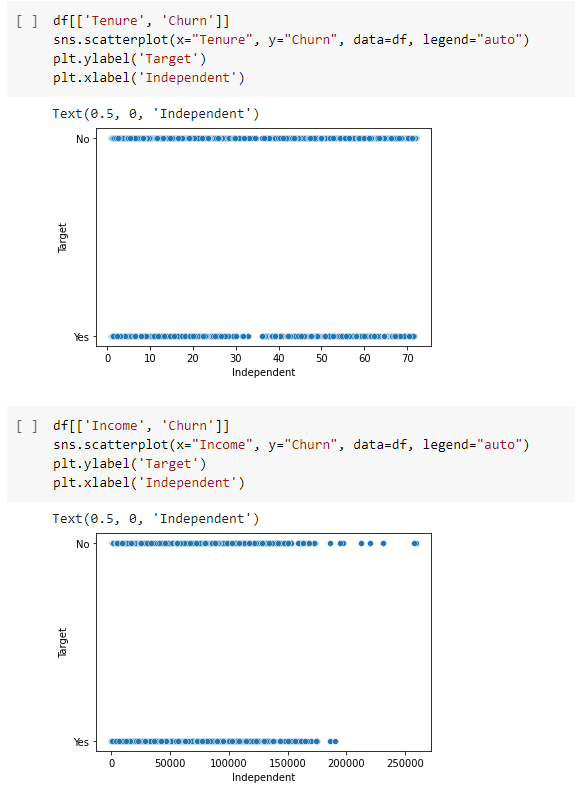
Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated





5.  Provide a copy of the prepared data set.

Diagram

Description automatically generated with low confidence

**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

A picture containing chart

Description automatically generated

Text

Description automatically generated

A picture containing text

Description automatically generated

Text, table

Description automatically generated  


Table

Description automatically generated

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.

The way I justified the variable selection for the reduced model is in the form of a heatmap. Looking at the heatmap I can see a strong correlation with Churn and MonthlyCharge at 0.37, and on the other hand nothing else correlated with Churn. Next, I ran a feature importance technique “Sequential Feature Selector” for top 10 feature important variables, which were:

* Tenure
* MonthlyCharge
* Techie\_Yes
* Contract\_One year
* Contract\_Two Year
* InternetService\_Fiber Optic
* Multiple\_Yes
* OnlineBackup\_Yes
* DeviceProtection\_Yes
* TechSupport\_Yes

Graphical user interface, text

Description automatically generated with medium confidence

Chart

Description automatically generated

Chart

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated





Text

Description automatically generated

3.  Provide a reduced logistic regression model.

Table

Description automatically generated

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

For reducing the model, I looked at the correlation map to see which features were highly correlated with Churn, and the heatmap showed only 1 MonthlyCharge. Next, I used a Sequential Feature Selector to determine the top 10 important features to Churn, which produced the results in a list of “True/False” for each column. I then compared those results to the column names and extracted each column that read True, those variables were:

* Tenure
* MonthlyCharge
* Techie\_Yes
* Contract\_One year
* Contract\_Two Year
* InternetService\_Fiber Optic
* Multiple\_Yes
* OnlineBackup\_Yes
* DeviceProtection\_Yes
* TechSupport\_Yes

This reduced model with 10 variables still produced a Psuedo R2-Value of 0.6113, from the initial model value of 0.6204 Psuedo R2-Value that contained all variables including dummy variables.

Initial regression equation of 35 variables:

Y= -4.6179 + 9.196e-05(Children) + 0.0020(Age) + 5.6e -07(Income) – 0.0022 (Outages\_sec\_perweek) -0.0079 (Email) + 0.0005 (Bandwidth\_GB\_Year) – 0.0560 (Contacts) -0.0348 (Yearly\_equip\_failure) -0.1541(Tenure) +0.0389 (MonthlyCharge) -0.0296 (Timely response) + 0.0032 (timely fixes) + 0.0280 (Timely replacements) -0.0262 (Reliability) -0.0303 (Options) -0.0229 (Respectful response) + 0.0042 (Courteous exchange) -0.0107 (Evidence of active listening) + 1.0932 (Techie\_Yes) -3.3860 (Contract\_One Year) -3.4653 (Contract\_Two Year) + 0.1357 (Port\_modem\_Yes) -0.0483 (Tablet\_Yes) -1.9546 (InternetService\_Fiber Optic) -0.7461 (InternetService\_None) -0.2833 (Phone\_Yes) + 0.3532(Multiple\_Yes) -0.2731 (OnlineSecurity\_Yes) -0.1113 (OnlineBackup\_Yes) -0.0951 (DeviceProtection\_Yes) -0.2163 (TechSupport\_Yes)+ 1.1334 (StreamingTv\_Yes) +1.2910 (StreamingMovies\_Yes) + 0.1645 (PaperlessBilling\_Yes) + 0.2404 (Gender\_Male)

Graphical user interface, application

Description automatically generated

(Tenure)

Text

Description automatically generated

Text, letter

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Chart

Description automatically generated

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

All calculations shown above.

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

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

All calculations shown above

**Part V: Data Summary and Implications**

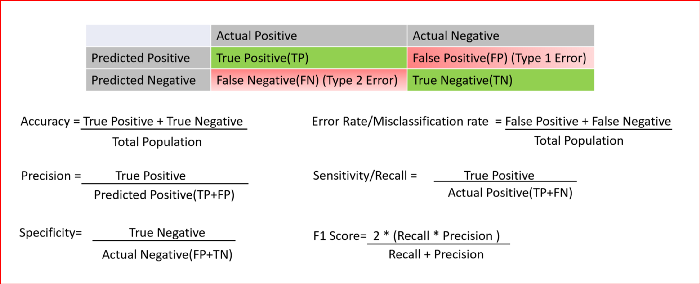
F.  Summarize your findings and assumptions by doing the following:

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

•  a regression equation for the reduced model

Reduced regression equation of 10 variables:

Y= -6.8436 -0.1143 (Tenure) + 0.0635 (MonthlyCharge) -3.2977 (Contract\_One Year) -3.3687 (Contract\_Two Year) -2.3783 (InternetService\_Fiber Optic) -0.4467 (Multiple\_Yes) -0.6511 (OnlineBackup\_Yes) -0.3648 (DeviceProtection\_Yes) -0.5482 (TechSupport\_Yes) + 1.0485(Techie\_Yes)



1. The regression model consisted of 10 independent variables
2. Tenure
3. MonthlyCharge
4. Techie\_Yes
5. Contract\_One year
6. Contract\_Two Year
7. InternetService\_Fiber Optic
8. Multiple\_Yes
9. OnlineBackup\_Yes
10. DeviceProtection\_Yes
11. TechSupport\_Yes

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

The coefficients show for every 1 unit

* Tenure-Churn: coefficient - 0.1143
* MonthlyCharge-Churn: coefficient – 0.0635
* Contract\_One year-Churn: coefficient - -3.2977
* Contract\_Two Year-Churn: coefficient - -3.3687
* Techie\_Yes -Churn: coefficient – 1.0485
* Multiple\_Yes – Churn: coefficient - -0.4467
* OnlineBackup\_Yes – Churn: coefficient - -0.6511
* DeviceProtection\_Yes – Churn: coefficient - -0.3648
* TechSupport\_Yes-Churn: coefficient - -0.5482

•  the statistical and practical significance of the model

P-values for Tenure, MonthlyCharge, Contract\_One year, Contract\_Two Year, Techie\_Yes, Multiple\_Yes, OnlineBackup\_Yes, DeviceProtection\_Yes, and TechSupport\_Yes were 0.000 “statistically significant”.

•  the limitations of the data analysis

The limitations of the data analysis are that the dataset given may need to gather more data overtime since the data is relatively small.

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

1. Based on the results the positive relationship with these variables, the stakeholders should focus on increasing the value of Techie\_Yes, Multiple\_Yes, OnlineBackup\_Yes, DeviceProtection\_Yes, TechSupport\_Yes to continue to retain customers. Also, can focus promotions and marketing to these individuals with could drive more attention to customer that fit this category. Furthermore, promotions or discounts to reduce monthly charges, that may be packaged for different contract types can improve customer tenure and reduce customer churn.

**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

Note: The audiovisual recording should feature you visibly presenting the material (i.e., not in voiceover or embedded video) and should simultaneously capture both you and your multimedia presentation.

Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the “WGU” option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto’s website.

To submit your recording, upload it to the Panopto drop box titled “Logistic Regression Modeling – NBM2 | D208.” Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.

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.

None

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

None were used

J.  Demonstrate professional communication in the content and presentation of your submission.