**Part I: Research Question**

A. Describe the purpose of this data analysis.

1. Research Question:

What variables associate with Churn? Identifying these variables could allow the company to optimize their services to reduce churn. This question could also produce insights into what the customers find valuable , increase service metrics, and allow the company to develop more impactful products.

1. The goal of this analysis is to insight into what factors associate with customer churn.

**Part II: Method Justification**

B. Describe logistic regression methods.

1. Assumptions:

The first assumption is that the model will have a binary outcome. For example, two categories (1, 0) representing Yes and No, true/false etc.

Second, there is an assumption that there are no significant outliers Logistic regressions are particularly sensitive to outliers and can cause shifts in the model such that it loses its statistical significance. Outliers should be removed from the variables.

Third, no multicollinearity, if the dependent variables are too correlated it can impact the coefficients and make it hard to parse the individual impact of each dependent variable on the independent variable.

A fourth assumption is the assumption of independence of errors. Simply put, the difference between the predicted outcome and actual outcome of any input in the model should not be impacted by any other input in the model. Stock price data is a good example where the price each day might be correlated over time to itself and not the target variable.

Benefits of Python:

I will be using python for this project for a couple reasons. The first is the breadth of libraries such as Numpy, pandas, Scipy etc., These libraries will allow me to wrangle, visualize and perform the regressions on the data. Another reason is python is relatively easy to use and I have been using python for my previous assignment and gained more conform using it and its libraries.

Pandas: Managing the dataset

Numpy: Performing mathematical operations on arrays.

MatPlotlib: Graphing.

Seaborn: Graphing.

Sklearn: Statical modeling .

1. Justification for Logistic Regression:

Logistic regression allows us to look at how independent variables impact the probability of binary dependent variable outcomes. This means of predicting outcomes based on other factors we may be able to gain competitive advantages. It can more easily include categorical data, does not assume the dependent variables be normally distributed, and the model is more easily understood as the coefficients are probability and don’t require deeper understanding of the dependent variable like in linear regression.

**Part III: Data Preparation**

C. Summarize the data preparation process.

1. Data Cleaning:

I’m using the same data set that I cleaned in D206. I begin cleaning the data by creating an overview by using the .info() function. That gave an overview of the data and allowed me to validate the dtypes of each variable. Next, I identified duplicates by using the .duplicated() functions. Then I used the .isnull().sum() function to check for missing or null values of relevant data. Lastly, used the .boxplot() function to determine possible outliers in the quantitative data.

Here is an example of filling null values with median values to preserve the data.

A screen shot of a graph

Description automatically generated

1. Variable Summary Statistics:

I used df.info() to show me all the variables, their count, and data type. This provides a good overview of the data and allows me to make sure the data is in a type that makes sense for the variable and allows me to perform statistically significant analytics.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Bandwidth\_GB\_Year – Count indicates that there are 9549 entries in this variable. This variable has a mean (average) value of 3582.272427. The standard deviation between entries is 2143.621539. The minimum (smallest) value in the variable is 155.506714. The 25th, 50th, and 75th percentiles show how the data is segmented. Meaning 75% of the values are greater than 1301.393401 and 25% are between 155.506714 and 1301.393401. This example can be used for the other percentiles. The max (highest) value is 7138.309000.

A screenshot of a computer

Description automatically generated

Tenure– Count indicates that there are 9549 entries in this variable. This variable has a mean (average) value of 6.441438. The standard deviation between entries is 25.898210. The minimum (smallest) value in the variable is 1.005104. The 25th, 50th, and 75th percentiles show how the data is segmented. Meaning 75% of the values are greater than 8.691329 and 25% are between 1.005104 and 8.691329. This example can be used for the other percentiles. The max (highest) value is 71.999280.

A white screen with black numbers

Description automatically generated

Tablet – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .3 indicates 30% of the population surveyed have tablets. A standard deviation (std) of .45 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have a tablet. The min of 0 and max of 1 confirm the variable is binary.

A screenshot of a computer

Description automatically generatedA black text on a white background

Description automatically generated

Phone – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .91 indicates 91% of the population surveyed have tablets. A standard deviation (std) of .27 aligns with the 25th , 50th , and 75th percentiles being 1. This indicates at least 75% of the entries were 1. The min of 0 and max of 1 confirm the variable is binary.

A screenshot of a computer

Description automatically generatedA number of text on a white background

Description automatically generated

OnlineSecurity – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .35 indicates 35% of the population surveyed have tablets. A standard deviation (std) of .47 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have a tablet. The min of 0 and max of 1 confirm the variable is binary.

A screen shot of a computer code

Description automatically generatedA black text on a white background

Description automatically generated

OnlineBackup – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .45 indicates 45% of the population surveyed have the service. A standard deviation (std) of .49 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have the service. The min of 0 and max of 1 confirm the variable is binary.

A screenshot of a computer

Description automatically generatedA white background with black text

Description automatically generated

DeviceProtection – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .43 indicates 43% of the population surveyed have device protection. A standard deviation (std) of .49 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have a protection. The min of 0 and max of 1 confirm the variable is binary.

A screen shot of a computer code

Description automatically generatedA black text on a white background

Description automatically generated

TechSupport – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .33 indicates 33% of the population surveyed have the service. A standard deviation (std) of .47 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have a tech support. The min of 0 and max of 1 confirm the variable is binary.

A computer screen shot of a number

Description automatically generatedA close-up of a computer screen

Description automatically generated

StreamingTV – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .49 indicates 49% of the population surveyed have the service. A standard deviation (std) of .49 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have a StreamingTV. The min of 0 and max of 1 confirm the variable is binary.

A screen shot of a computer

Description automatically generatedA black text on a white background

Description automatically generated

StreamingMovies – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .48 indicates 48% of the population surveyed have the service. A standard deviation (std) of .49 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have the service. The min of 0 and max of 1 confirm the variable is binary.

A screen shot of a computer

Description automatically generatedA white background with black text

Description automatically generated

PaperlessBilling – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .58 indicates 58% of the population surveyed have the service. A standard deviation (std) of .49 aligns with the 75th and 50th percentiles being 1. This indicates at least half the entries were 1. The 25th percentile being one indicates at least 25% don’t have the service. The min of 0 and max of 1 confirm the variable is binary.

A computer screen shot of numbers

Description automatically generatedA white background with black text

Description automatically generated

InternetService\_DSL – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .34 indicates 34% of the population surveyed have the service. A standard deviation (std) of .47 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have the service. The min of 0 and max of 1 confirm the variable is binary.

A computer screen shot of a number

Description automatically generatedA number of numbers on a white background

Description automatically generated

InternetService\_Fiber Optic – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .44 indicates 44% of the population surveyed have the service. A standard deviation (std) of .49 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have the service. The min of 0 and max of 1 confirm the variable is binary.

A computer screen shot of a computer program

Description automatically generatedA number of numbers on a white background

Description automatically generated

Categorical Variables Summary Statistics.

None of these are used in the regression but apparently required. These display the percentage of the total that each unique entry equals. For example, “K409198” = 0.01% of the customer IDs in the ‘Custimer\_id’ variable because there are 9549 entries and each one is unique but equally useless in the logistic regression. The logic can be applied to each variable in the screen shots below.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A number of numbers and symbols

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated

A screenshot of a document

Description automatically generated

A screenshot of a computer

Description automatically generated

Univariate Visualizations:

A group of graphs showing different types of data

Description automatically generated with medium confidence

A graph of different types of graphs

Description automatically generated with medium confidence

Bivariate Visualizations:

A screenshot of a graph

Description automatically generated

A group of white rectangular shapes

Description automatically generated

1. Data Transformation:

First , when thinking about the question of association with MonthlyCharge, I decided the variables covering service would be important to include. The issue is they were “Yes” or “No” which is nominal categorical data. So, I converted these to binary data where Yes = 1 and No = 0. Allowing me to include these variables in the initial regression.

A screen shot of a computer program

Description automatically generated

The second thing I wanted to do was include the internet service category. However, unlike the previous variables the data is not binary. There are three options: DSL, Fiber Optics, and None. To be able to include these I needed to create dummy variables which splits them out into their own variables with binary entries for Yes, they have the service = 1 and No they didn’t have the service = 2.

A computer screen shot of a program

Description automatically generated

Third I chose a list of variables to drop. These variables I decided were not likely to give valuable insight, risked the statistical significance of the model by adding unnecessary complexity. For example, Items 1-8. Each have 8 possible responses which would need dummy variables created for them. I also dropped InternetService\_None to try to avoid the issue of multicollinearity when creating dummy values.

A screen shot of a computer

Description automatically generated

1. Data Set: *See attached.*



**Part IV: Model Comparison and Analysis**

D. Compare initial and reduced linear regression model.

1. Initial Model:

A screenshot of a computer

Description automatically generated

1. Feature Selection or Model Evaluation:

I chose to reduce the model by eliminating variables with the highest p value over 0.05 and then rerunning the regression. This is called Backward Stepwise Elimination. Below that indicates a strong statistical significance and are variables I deemed to be important. Therefore eliminating: Tablet, OnlineBackup, DeviceProtection, TechSupport and Paperless Billing.

Removing Device Protection because it has the highest p-value

A screenshot of a computer

Description automatically generated

Removing Online Backup because it has the highest p value.

A screenshot of a computer

Description automatically generated

Removing Tablet because it has the highest p value.

A screenshot of a computer

Description automatically generated

Removing Paperless Billing because it has the highest p value over 0.05.

A screenshot of a computer

Description automatically generated

Removing Tech Support because it has the highest p value over 0.05.

A screenshot of a computer

Description automatically generated

1. Reduced Model: Can’t be improved.

A screenshot of a computer

Description automatically generated

E. Analyze the data set using the reduced model.

1. Data Analysis Process:

The initial regression contained 14 variables which had a varying degree of applicability. First, we can look at the p values in the output for the initial regression. This is an indicator of the statical significance of the variables. Generally, a p value of < .05 is acceptable. We can see there are five variables with p values over that threshold. These include Tablet, OnlineBackup, DeviceProtection, TechSupport and Paperless Billing. The existence of these variables in the regression could call into question it’s statical significance. Looking at the reduced regression we see our remaining variables have p values of 0.000 or 0.003 meaning they are statically significant. A Pseudo R-squared of 0.4164 indicates a moderate fit and a LLR p-value of essentially 0 means the regression is statistically significant.

1. Output and Calculations:

A screenshot of a computer program

Description automatically generated

1. Code: See *Attached*

**Part V: Data Summary and Implications**

F. Summarize findings and implications.

1. Results Discussion:

Converting coefficients to odds.

A screenshot of a computer program

Description automatically generated

The coefficients show log odds. I converted them to odds to make them easier to understand. For example, holding all other variables constant, adding Online Security would increase the chance of churn 2.44 times. The accuracy score of ~84% is solid, and means the model was able to predict churn with that accuracy. The confusion matrix shows the model accurately predicted 1933 instances of non-churn cases and predicted false positives for churn 162 times. The model predicted a no churn situation when there was churn 300 times and correctly predicted churn 470 times. One limitation of the model is that is stronger at predicting instances of no churn rather than when a customer will churn.

Recommendations:

I think logically the model passes the sniff test. It shows having fiber optic service as well as online security would decrease the likelihood of customer churn. On the other hand, adding tv or movie streaming increase the risk of churn. These services could increase the monthly charge and make customers more likely to churn. To reduce churn, based on this model, I would suggest upgrading customers to fiber optic service internet.

**Part VI: Demonstration**

G. Provide a Panopto video recording. – *See Attached*

[Panopto Video](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=95d432b5-530c-4ab2-a978-b122004cdc71)

H. List of Web Sources

Scikit-Learn Developers. (n.d.). *Logistic regression in scikit-learn*. Scikit-Learn Documentation. Retrieved March 11, 2024, from

McLeod, S. A. (2019). What a p-value tells you about statistical significance. Simply Psychology.

StatSoft. (n.d.). Logistic Regression. Electronic Statistics Textbook. Tulsa, OK: StatSoft.

<https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704-ep713_multivariablemethods/BS704-EP713_MultivariableMethods5.html>

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704\_multivariable/bs704\_multivariable8.html