|  |
| --- |
| Western Governors University |
| Predictive Modeling |
| D208 Task 2 |

|  |
| --- |
| Allison Casey  6-2-2024 |

**Part I: Research Question**

**A1: RESEARCH QUESTION**

What causes customers to churn?

**A2: GOALS**

The submission defines the goals of the data analysis, and the goals are relevant.

The objective of the analysis will be to increase insight into what factors from the data set correlate to churn for customers. Ideally this will help gain actionable insights to retain customers and prevent churn.

**Part II: Method Justification**

**B1: SUMMARY OF ASSUMPTIONS**

1. The dependent variable is binary.
2. The dependent variables are not too highly correlated with each other. (no multicollinearity)
3. The observations are independent and random.
4. The data which the independent variables are drawn from is big enough for the conclusions to be reliable.

**B2: TOOL BENEFITS**

The programming language used is python because it is very versatile and has lots of packages that can be easily used for each part of the analysis to simplify automating things such as the regression model, cleaning, or generating visuals. This can all be done without having to write a lot of code through using these packages and libraries. It is also an interpreted language which can make it quicker at iterative processes which in this case is useful for reducing the model that will be created.

**B3: APPROPRIATE TECHNIQUE**

Logistic regression is an appropriate technique to answer the research question because it is a tool used to model a relationship between a categorical dependent variable and independent variables both continuous and categorical. In this case the research question is looking to explore the relationship of churn, a binary categorical variable, with multiple other variables in the data set. As a result, logistic regression is a great option because it can be used to model this relationship.

**Part III: Data Preparation and Manipulation (Cleaning à Exploration à Wrangling)**

**C1: DATA CLEANING GOALS**

The submission describes the data cleaning goals and the steps used to clean the data to achieve the goals. The goals and steps align with each other and with multiple linear regression analysis. The annotated code is complete.

The data cleaning goals are to confirm that there are no missing values or outliers that need to be treated as well as remove columns that will not be helpful for the analysis. To accomplish this the following was done:

1. Info() was used to see the counts of each variable and the data types
   1. *df.info()*
2. The info call showed that there is only one column with missing values, but upon further reviewing the data set, this column doesn’t have any null values it just got read in as NULL rather than None indicating the customer doesn't have this service, so these were added back in so that the data set was accurate and not missing any values.
   1. *df['InternetService'].fillna('None',inplace=True)*
3. Columns not being used for analysis were removed. The columns chosen were columns that contained irrelevant data for the analysis such as the Customer\_id which wouldn’t have any effect on any variables. Demographic data and survey questions were also removed so that the analysis could focus on data that related more closely to services and performance of the telecommunications company.
   1. *df = df.drop(columns=['Customer\_id', 'Interaction', 'UID', 'City', 'State', 'County', 'Zip', 'Lat', 'Lng', 'TimeZone', 'Area', 'Job', 'Children', 'Age', 'Income', 'Marital', 'Gender', 'Techie', 'Item1', 'Item2', 'Item3', 'Item4', 'Item5', 'Item6', 'Item7', 'Item8'])*
4. The describe function was then called on the data frame to review the continuous variables for outliers and this was also used for the summary statistics below. Upon reviewing this it was decided that it would be best to not treat outliers to keep the model as accurate as possible since the outliers didn’t appear to be outrageous for any of the columns.
   1. *df.describe()*

**C2: SUMMARY STATISTICS\**

Continuous variables:

A screenshot of a computer

Description automatically generated

The continuous variables include the independent variables: 'Population', 'Outage\_sec\_perweek', 'Email', 'Contacts', 'Yearly\_equip\_failure', 'MonthlyCharge', and 'Bandwidth\_GB\_Year'. These are all shown in the image above with their summary statistics.

Categorical variables:

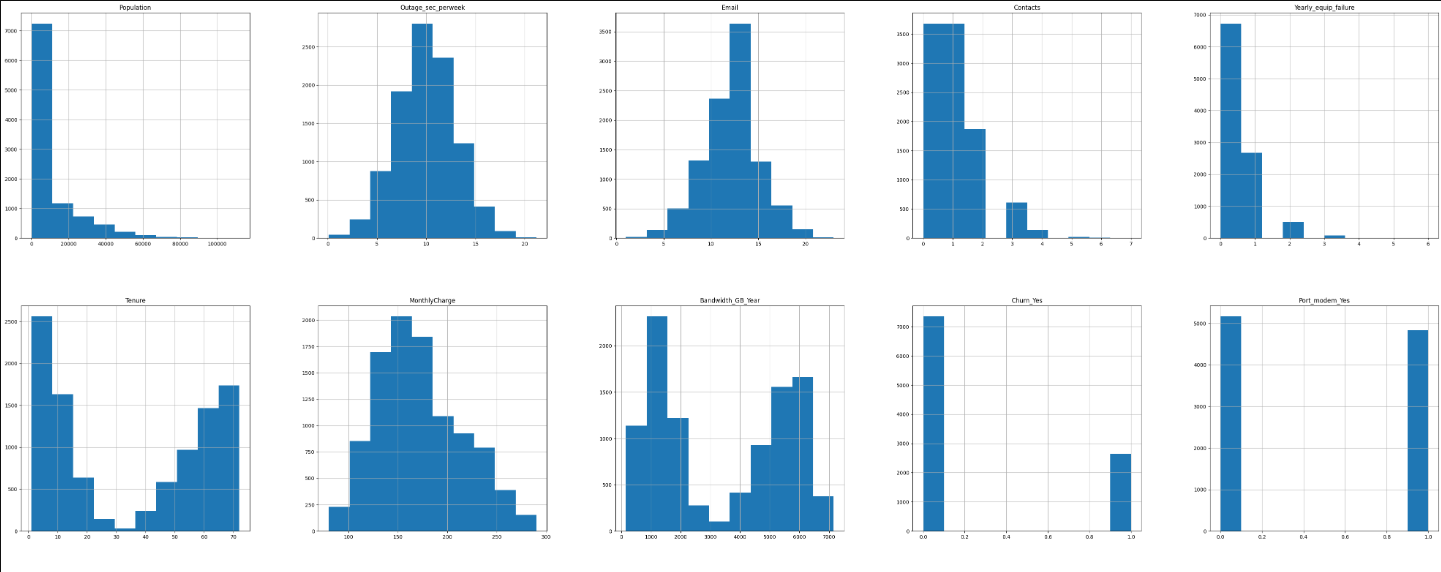
The categorical columns that will be used as independent variables include: 'Port\_modem', 'Tablet', 'Phone', 'Multiple', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'PaperlessBilling', 'Contract', 'InternetService', and 'PaymentMethod'. They also include the dependent variable ‘Churn’. These are all shown in the image to the left with their summary statistics.A screen shot of a computer

Description automatically generated

**C3: VISUALIZATIONS**

* A graph of a graph

  Description automatically generated with medium confidenceA group of blue and white bars

  Description automatically generatedUnivariate Visualizations
* **­**Bivariate Visualizations
  + A screenshot of a graph

    Description automatically generatedA blue and orange squares

    Description automatically generatedCategorical Variables

A graph showing a number of blue squares

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated­

* + A group of white rectangular objects with blue dots

    Description automatically generatedContinuous Variables

**C4: DATA TRANSFORMATION**

The goals for data transformation were to encode the categorical variables to numbers because statistical methods almost exclusively work with numeric data. The method used to do this was one hot encoding which for each feature a new column is created with a binary encoding of whether the row belongs to the category. This was done using the method get\_dummies() from the Panda’s library with the ‘drop\_first’ set to true to use number of columns -1 to mitigate some multicollinearity early on and then the columns created from this were renamed for clarity.

*#One hot encoding*

*df = pd.get\_dummies(df, columns=categorical\_columns, drop\_first=True, dtype = int)*

*df = df.rename(columns = {'Contract\_One year':'Contract\_One\_Year', 'Contract\_Two Year':'Contract\_Two\_Year','InternetService\_Fiber Optic':'InternetService\_Fiber\_Optic', 'InternetService\_Fiber Optic':'InternetService\_Fiber\_Optic', 'PaymentMethod\_Credit Card (automatic)':'PaymentMethod\_CC', 'PaymentMethod\_Electronic Check':'PaymentMethod\_ECheck', 'PaymentMethod\_Mailed Check':'PaymentMethod\_MCheck'})*

*df.info()*

See D208\_Task1.ipynb for full annotated data transformation code.

**C5: PREPARED DATA SET**

PREPARED\_Task2\_churn\_clean\_data.csv

**Part IV: Model Comparison and Analysis**

**D1: INITIAL MODEL**

A screenshot of a computer screen

Description automatically generated

**D2: JUSTIFICATION OF MODEL REDUCTION**

The statistically based feature selection used for reduction was backward stepwise elimination where the model starts with all the features and then removes the least significant feature based on the p-value at each iteration until the model stops improving. For this reduction .05 was chosen as the alpha value that the p-value should not be greater than. This method was used because it reduces the number of features, contributes to reducing multicollinearity, and helps with preventing overfitting.

**D3: REDUCED LINEAR REGRESSION MODEL**

**A screenshot of a computer program

Description automatically generated**The image on the right displays the stepwise regression code and the columns removed at each iteration. The image on the right displays the resulting reduced model.

**A screenshot of a computer

Description automatically generated**

**E1: MODEL COMPARISON**

|  |  |
| --- | --- |
| Initial Pseudo R-squared | .6111 |
| Final Pseudo R-squared | .6094 |

The initial model was reduced using backwards stepwise regression based on the p-value. The pseudo r squared is on a scale of 0 to 1 where a better fit is indicated by higher values. In this case the values are very similar, but it would indicate that the initial model is better than the reduced model. However, the reduced model’s features all have p-values below .05 which means that all the features are statistically significant.

**E2: OUTPUT AND CALCULATIONS**

A confusion matrix and accuracy score was calculated on the final model using the sklearn functions confusion\_matrix() and accuracy\_score() which generated the below output:

A screen shot of a computer program

Description automatically generated

**E3: CODE**

D208\_Task2.ipynb

**Part V: Data Summary and Implications**

**F1: RESULTS**

A mathematical equation with numbers and letters

Description automatically generated

= -4.1038 - .3305(Phone\_Yes) + .4243(Multiple\_Yes) - .1861(OnlineSecurity\_Yes) - .1908(Tech\_Support\_Yes) + 1.3154(StreamingTV\_Yes) + 1.4445(StreamingMovies\_Yes) – 3.4039(Contract\_One\_Year) – 3.3911(Contract\_Two\_Year) – 2.0316(InternetService\_Fiber\_Optic) - .9329(InternetService\_None) + .5063(PaymentMethod\_ECheck) - .1114(Tenure) + .0360(MonthlyCharge)

For the context of this model:

* A 1 unit increase in Phone\_Yes will result in -0.3305 change in the log-odds of Churn\_Yes
* A 1 unit increase in Multiple\_Yes will result in 0.4243 change in the log-odds of Churn\_Yes
* A 1 unit increase in OnlineSecurity\_Yes will result in -0.1861 change in the log-odds of Churn\_Yes
* A 1 unit increase in Tech\_Support\_Yes will result in -0.1908 change in the log-odds of Churn\_Yes
* A 1 unit increase in StreamingTV\_Yes will result in 1.3154 change in the log-odds of Churn\_Yes
* A 1 unit increase in StreamingMovies\_Yes will result in 1.4445 change in the log-odds of Churn\_Yes
* A 1 unit increase in Contract\_One\_Year will result in -3.4039 change in the log-odds of Churn\_Yes
* A 1 unit increase in Contract\_Two\_Year will result in -3.3911 change in the log-odds of Churn\_Yes
* A 1 unit increase in InternetService\_Fiber\_Optic will result in -2.0316 change in the log-odds of Churn\_Yes
* A 1 unit increase in InternetService\_None will result in -0.9329 change in the log-odds of Churn\_Yes
* A 1 unit increase in PaymentMethod\_ECheck will result in 0.5063 change in the log-odds of Churn\_Yes
* A 1 unit increase in Tenure will result in -0.1114 change in the log-odds of Churn\_Yes
* A 1 unit increase in MonthlyCharge will result in 0.0360 change in the log-odds of Churn\_Yes

The model is statistically significant because the p-values of the features are all below .05 and the pseudo r-squared value is closer to one than zero. The model also demonstrated an accuracy of 90%. This lends to the model being practically significant because it shows some statistical accuracy and provides insight into factors that influence churn which helps answer the research question. However, the model is limited firstly in terms of the fact that there may be other factors influencing churn from the columns that were dropped and there is not a huge amount of data to test and train the model with. Having more data sets would help provide more accurate insights and avoid overfitting.

**F2: RECOMMENDATIONS**

Based on the results of the analysis, it could be beneficial to really promote the fiber optic internet service since it appears to have a strong relationship negatively impacting the log-odds of the customer churning. There is also a strong negative relationship with the contracts which should be explored to see how those can prevent churn as well. It could also be worth further investigating whether offering a lot of the other services such as the streaming services are worth the resources to maintain since they had a positive relationship with churn. Further research could also be done regarding the columns that were dropped that contained more of the demographic and physical location which could also indicate areas for improvement. Overall, further analysis based on all the relationships discovered and explored would be a good course of action.

**Part VI: Demonstration**

**G: PANOPTO DEMONSTRATION**

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=acb4e8b1-f156-4a5b-b391-b1830179a6c5

**Sources**

“Accuracy\_score in Sklearn - Javatpoint.” *Www.Javatpoint.Com*, www.javatpoint.com/accuracy\_score-in-sklearn. Accessed 2 June 2024.

B, Harikrishnan N. “Confusion Matrix, Accuracy, Precision, Recall, F1 Score.” *Medium*, Analytics Vidhya, 1 June 2020, medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd.

Frost, Jim. “How to Interpret P-Values and Coefficients in Regression Analysis.” *Statistics By Jim*, 21 Mar. 2023, statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/.

Kwok, Ryan. “Stepwise Regression Tutorial in Python.” *Medium*, Towards Data Science, 9 Mar. 2021, towardsdatascience.com/stepwise-regression-tutorial-in-python-ebf7c782c922.

Middleton, Keiona. “Getting Started With D208 Part1.”

Middleton, Keiona. “Getting Started With D208 Part2.”

“Statsmodels.Tools.Tools.Add\_constant¶.” *Statsmodels.Tools.Tools.Add\_constant - Statsmodels 0.14.1*, www.statsmodels.org/stable/generated/statsmodels.tools.tools.add\_constant.html. Accessed 26 May 2024.

“Stepwise Regression in Python.” *GeeksforGeeks*, GeeksforGeeks, 23 May 2023, www.geeksforgeeks.org/stepwise-regression-in-python/.

Zach BobbittHey there. My name is Zach Bobbitt. I have a Masters of Science degree in Applied Statistics and I’ve worked on machine learning algorithms for professional businesses in both healthcare and retail. I’m passionate about statistics. “How to Create a Q-Q Plot in Python.” *Statology*, 11 June 2022, www.statology.org/q-q-plot-python/.

Zach BobbittHey there. My name is Zach Bobbitt. I have a Masters of Science degree in Applied Statistics and I’ve worked on machine learning algorithms for professional businesses in both healthcare and retail. I’m passionate about statistics. “How to Extract P-Values from Linear Regression in Statsmodels.” *Statology*, 5 Aug. 2022, www.statology.org/statsmodels-linear-regression-p-value/.