|  |
| --- |
| Western Governors University |
| Predictive Modeling |
| D208 Task 1 |

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

**Part I: Research Question**

**A1: RESEARCH QUESTION**

What causes a longer tenure for customers?

**A2: GOALS**

The objective of the analysis will be to increase insight into what factors from the data set correlate to a longer tenure for customers. Ideally this will help gain actionable insights to continue to grow customer tenure.

**Part II: Method Justification**

**B1: SUMMARY OF ASSUMPTIONS**

1. There is a linear relationship between the independent and dependent variables.
2. The dependent variables are not too highly correlated with each other. (no multicollinearity)
3. The observations are independent and random.
4. The residuals of the model are normally distributed.

**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**

Multiple linear regression is an appropriate technique to answer the research question because it is a tool used to model a relationship between a continuous dependent variable and continuous variables both continuous and categorical. In this case the research question is looking to explore the relationship of tenure, a continuous variable, with multiple other variables in the data set. As a result, multiple linear 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**

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', 'Churn', '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()*

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

**C2: SUMMARY STATISTICS**

Continuous variables:

A screenshot of a computer

Description automatically generated

The continuous variables include the dependent variable ‘Tenure’ as well as 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:

A black screen with white text

Description automatically generatedThe 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'. These are all shown in the image to the left with their summary statistics.

**C3: VISUALIZATIONS**

* **­**Univariate Visualizations

A group of blue graph

Description automatically generated

A group of blue and white graphs

Description automatically generated

* Bivariate Visualizations
  + Categorical Variables

A chart of blue and orange rectangular shapes

Description automatically generated

* + Continuous Variables

A group of blue dots

Description automatically generated

Please note that visualizations were created after the data transformation for ease of visualizing categorical variables.

**C4: DATA TRANSORMATION**

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\_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. Prior to performing the backwards stepwise elimination, the model was also tested for multicollinearity to remove any explanatory variables that are too highly correlated with one another. This was done by calculating the variance inflation factor (VIF) and then removing any features with a value greater than 5 which would indicate a severe correlation.

**D3: REDUCED LOGISTIC REGRESSION MODEL**

First the VIF was calculated and the columns were removed with a VIF greater than 5 which were StreamingTV\_Yes, StreamingMovies\_Yes, and MonthlyCharge.

A screenshot of a computer screen

Description automatically generatedA screenshot of a black screen

Description automatically generated

After backward stepwise elimination we see that all the p-values are all now below .05.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**E1: MODEL COMPARISON**

|  |  |
| --- | --- |
| Initial R squared | .998 |
| Final R squared | .993 |

The model was first reduced by removing columns with a VIF higher than 5 then using backward stepwise elimination by P value. The adjusted R squared value helps determine how well the independent variables explain the dependent variable with 1 being a perfect fit. Both the initial and the final models have very good R-squared values indicating a good fit. What is interesting to note is that there was a decrease in this metric indicating a worse fit. This is likely due to overfitting of the data in the initial model whereas there might be different results when comparing the two models against different data sets in which case the reduced model could potentially be a better fit despite that not being the case in this scenario.

**E2: OUTPUT AND CALCULATIONS**

QQ Plot

**A graph with a red line

Description automatically generated**

Residual Standard Error

Calculated using the statsmodels library by applying the standard deviation method with the degree of freedom equal to the number of predictors (p) + 1 as below:

A screen shot of a computer

Description automatically generated

**E3: CODE**

D208\_Task1.ipynb

**Part V: Data Summary and Implications**

**F1: RESULTS**

Regression Equation:

y = -8.1347 - .1067x - .8702x - .9945x - 1.0999x - 1.0998x + 5.0472x + 5.0252x + .0121x

* For every 1 value increase, Tablet\_Yes has a - .1067 relationship on tenure
* For every 1 value increase, Multiple\_Yes has a -0.8702 relationship on tenure
* For every 1 value increase OnlineSecurity\_Yes has a -0.9945 relationship on tenure
* For every 1 value increase OnlineBackup\_Yes has a -1.0999 relationship on tenure
* For every 1 value increase DeviceProtection\_Yes has a -1.0998 relationship on tenure
* For every 1 value increase InternetService\_Fiber\_Optic has a 5.0472 relationship on tenure
* For every 1 value increase InternetService\_None has a has a 5.0252 relationship on tenure
* For every 1 value increase Bandwidth\_GB\_Year has a 0.0121 relationship on tenure

The high F-statistic indicates that the model is statistically significant though the R-squared decrease from the initial model and the p-values cast some doubt as to the practicality of the model. As far as limitations for the model go, it was only trained using one data set and could benefit from training against multiple data sets to avoid overfitting and improve the reduction process. Overall, the model is useful in helping identify what causes tenure based on the relationships discovered through the model.

**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 positive linear relationship with tenure for both new and existing customers. It may also be worth targeting customers that are heavy internet bandwidth users since that also appears to have a positive correlation with tenure. It could also be worth further investigating whether offering a lot of the other services such as security and backup is worth the resources to maintain since they had a negative effect on tenure. 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=37d534a3-14bd-4260-8390-b1830176048f

**Sources**

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/.