**Task 1: MLR Analysis**

**Part I: Research Question**

“Disregarding customer demographics, what are the top 3 features that impact how long a Month-to-month customer stays (tenure) with our telecom company?”

Goals:

1. Since the initial model has an Adjusted R-Squared value of .426, I am aiming for an Adjusted R-Squared value that is above .400 with my Reduced Model. This is because I am choosing only 3 variables to answer my research question. This means my input variables must explain at least 40% of the variation of the output variable.
2. Identify the top 3 variables that has the most impact on customer Tenure.
3. Since the initial model has a Residual Standard Error of 20.09, I am aiming for an Residual Standard Error of less than 21 with my Reduced Model.
4. Independent variables each have a p-value of less than 0.05.

**Part II: Method Justification**

Assumptions of a Multiple Linear Regression Model:

1. Residuals should be close to normally distributed (Multivariate Normality)
2. There should be little to No Multicollinearity
3. The variance should be constant and the same throughout (Homoscedasticity)
4. There should be a linear relationship between the dependent variable and independent variables.
5. There must be a large enough sample size (above 500).

I have chosen to use R for this analysis. The motto promoted in our text and videos is “R is for Regression”. These tasks do not require extensive computing power or the use of neural networks which would be better done in Python. Python would be overkill for this project. Also, with R, the code is much shorter especially with plotting and building multiple models with ease.

Linear Regression is the appropriate technique as my research question is based around the continuous, ordinal dependent variable of Tenure. My research question does not pose a binary classification problem which would be solved by Logistic Regression. The range for Tenure is infinite which cannot be solved using ogistic regression. In Multiple Linear Regression, the dependent variable is continuous without bounds and can extend past 0 or 1. Because of this, linear regression was chosen.

**Part III: Data Preparation**

My main goals in data preparation are the following:

1. Remove customer demographic data
2. Transform categorical variables into factors whether binary or tertiary for analysis
3. Remove variables with multicollinearity
4. Rename any variables with spaces used for my adjusted model

Please see the following for Outputs of my Summary Statistics.

1. “Task1 – Dataset Summary Output.PNG”
2. “Task1 – Correlation Matrix.csv”
3. “Task1 – Correlation Heatmap.pdf”

Explanation of all variables:

After prepping the dataset, I used summary() to get a summary of all variables and their distributions. Each categorical variable has a min and max range of 0 and 1, respectively. The dataset has not been standardized or normalized as I do not need power or logarithm transformation on my variables.

Explanation of the 4 variables used in my reduced model:

Tenure (Continuous; Dependent) – 50% of the data is within the range of 7.90 – 61.30 and the minimum and maximum values are 1.02 and 72.00, respectively. The minimum and maximum values are the shortest and longest amount of time a Month-to-Month customer has been a part of our company. The mean (34.24) and median (26.97) for Tenure are not extremely close to each other, but not too far apart. This indicates that the data is slightly skewed to the right. This can be seen in the Histogram Univariate Visualization for Tenure. Also seen in the Univariate Histogram Visualization for Tenure, the distribution is bimodal.

MonthlyCharge (Continuous; Independent) – The mean and median for MonthlyCharge are close together in the range which indicates that the data is not too skewed in any direction. 50% of the data is within the range of 139.98 – 200.12 and the minimum and maximum values are 79.98 and 290.16, respectively. The minimum and maximum values are the lowest and highest Monthly Charges a Month-to-Month customer has paid. The distribution of Monthly Charge is very close to a normal distribution. This can be seen in the Histogram Univariate Visualization for MonthlyCharge.

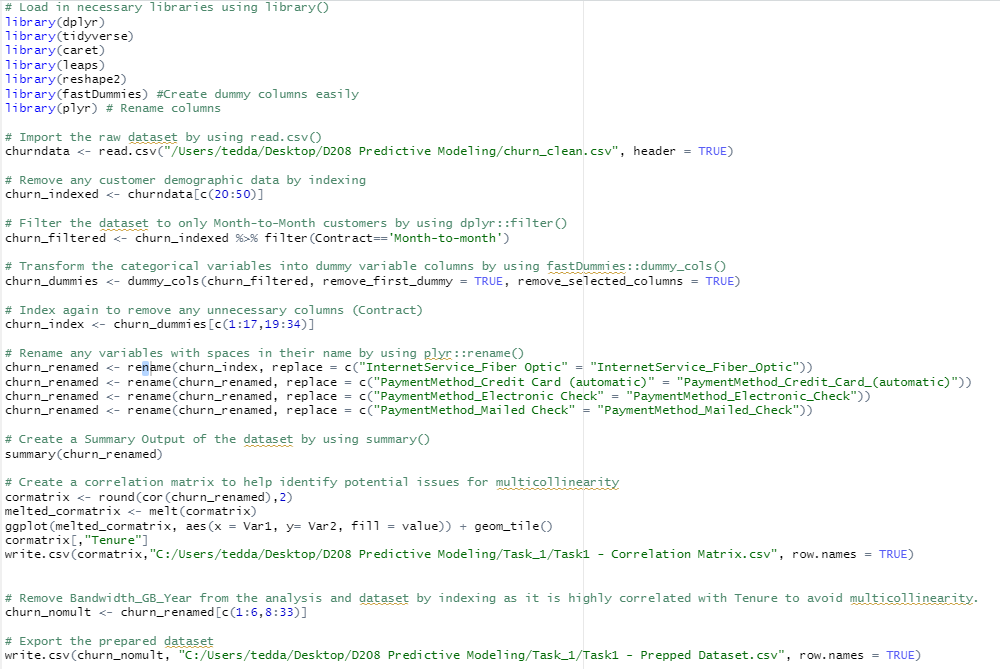
InternetService\_Fiber\_Optic (Categorical; Independent) – The median is 0 which indicates that most Month-to-Month customers do not have the Fiber Optic Internet Service. The mean of 0.44 indicates that there is a decent amount of Fiber Optic Internet Service, Month-to-Month customers, and the data is skewed to the right. This could be a potential growth opportunity for our telecom company to promote our internet service. This skew can be seen in the Histogram Univariate Visualization for InternetService\_Fiber\_Optic.

Churn\_Yes (Categorical; Independent) – The median is 0 which indicates that more customers did not churn than those that did (This is good). The mean of 0.3728 indicates that there is a decent number of customers which did churn, and the data is skewed to the right. This skew can be seen in the Histogram Univariate Visualization for Churn\_Yes.

Data Preparation Steps:

1. Load in necessary libraries using library()
2. Import the raw dataset by using read.csv()
3. Remove any customer demographic data by indexing
4. Filter the dataset to only Month-to-Month customers by using dplyr::filter()
5. Transform the categorical variables into dummy variable columns by using fastDummies::dummy\_cols()
   1. Remove the 1st dummy column of each variable by setting the remove\_first\_dummy = TRUE to avoid multicollinearity
   2. Remove the original categorical variable column by setting the remove\_selected\_columns = TRUE to avoid multicollinearity
6. Rename any variables with spaces in their name by using plyr::rename()
7. Create a correlation matrix to help identify potential issues for multicollinearity
8. Remove Bandwidth\_GB\_Year from the analysis and dataset by indexing as it is highly correlated with Tenure to avoid multicollinearity.
9. Export the prepared dataset

Annotated Code for Prepping the Dataset:



Please see the following for my Univariate and Bivariate Visualizations:

1. “Task1 – Univariate Visualizations.PNG”
2. “Task1 – Bivariate Visualizations.PNG”

Please see “Task1 – Prepped Dataset” for the copy of my prepared dataset.

**Part IV: Model Comparison and Analysis**

Initial Model Output:

Please see “Task1 – Initial Model without Multicollinearity Summary Output.PNG”

Feature Selection Output:

Please see “Task1 – Regression Subsets Summary Output.PNG”

Reduced Model Output:

Please see “Task1 – Adjusted Model Summary Output.PNG”

For my variable selection technique, I chose Regression Subsets by using leaps::regsubsets(). My reasoning for this is because my research question asks for the top 3 variables which impact Month-to-Month customer Tenure. Regression Subsets allows me to find the best combination of 3 variables with the most impact by setting the parameter nvmax = 3. The result maximizes the Adjusted R-Square to the highest possible with 3 variables selected. The 3 variables selected by regsubsets() were MonthlyCharge, InternetService\_Fiber\_Optic, Churn\_Yes.

Please see the following for the model evaluation metrics:

1. “Task1 – Initial Model without Multicollinearity Summary Output.PNG”
2. “Task1 – Adjusted Model Summary Output.PNG”
3. “Task1 – Reduced Model with only 2 Variables Summary Output.PNG”

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Initial Model** | **Adjusted Model** | **Adjusted Model w/ only 2 Variables** |
| **Adjusted R-Squared** | 0.4227 | 0.4134 | 0.3936 |
| **Multiple R-Squared** | 0.426 | 0.4137 | 0.3939 |
| **Residual Standard Error** | 20.09 | 20.25 | 20.58 |

The reduced model with only 3 variables resulted in an Adjusted R-Squared of 0.4134 which was about 0.01 less than initial model with 34 variables (0.4227). The Adjusted Model’s adjusted R-Square value explains 41.34% of the variance in Tenure. To further test my logic, I also checked a model with 2 variables (All variables above except for InternetService\_Fiber\_Optic) just in case I could get better results with less variables, but it resulted in an Adjusted R-Squared of 0.3936 which didn't meet my objective of an adjusted R-Squared of above 0.40.

The reduced model with only 3 variables resulted in a Residual Standard Error of 20.25 which is very close to the Residual Standard Error of the Initial Model of 20.09. There is a small difference of 0.16.

Please see the following for my residual plots:

1. “Task1 – Initial Model Residual Plots.pdf”
2. “Task1 – Adjusted Model Residual Plots.pdf”

The adjusted model residual plots are extremely like the initial model residual plots. With the adjusted model Residuals vs. Fitted plot, you can note that the residuals are starting to pivot more towards the x-axis (0,0).

Please see “Task1 – All Code.R” for my entire code.

**Part V: Data Summary and Implications**

Reduced Model Regression Equation(s):

Y = 18.944 – 38.918x1 + 0.194x2 – 7.999x3

Tenure = 18.944 – 38.918Churn\_Yes + 0.194MonthlyCharge - 7.999InternetService\_Fiber\_Optic

Interpretation of Adjusted Model Coefficients:

Please see “Task1 – Adjusted Model Coefficients”

Intercept – When all variables are 0, the intercept is 18.944.

Churn\_Yes – If a Month-to-Month customer churns (Churn\_Yes), the length of Tenure decreases by 38.918 when all other variables remain the same.

MonthlyCharge – For every unit (dollar) increase in MonthlyCharge, the length of Tenure increases by 0.194 when all other variables remain the same.

InternetService\_Fiber\_Optic – If a Month-to-Month customer has the Fiber Optic Internet Service, the length of Tenure decreases by 7.999 when all other variables remain the same.

Model’s Statistical and Practical Significance:

Although my model is statistically significant with a p-value of less than 2.2 x e-16, the Adjusted R-Squared for both the initial and reduced models are only about 0.42. This means that each model only accounts for about 42% of the variation of Tenure. With the removal of Bandwidth\_GB\_Year (multicollinearity), almost 60% of the variation of Tenure is unexplained by my models. With that amount of variation unexplained, I would not use this model in a practical, business setting. However, it does warrant the need for further analysis within the company (please see my recommendations). Further analysis would be practical and beneficial for our company.

Also, please note that this dataset is purely fictitious and may not represent an accurate depiction of a real-world telecom company’s data.

Limitations:

1. This analysis does not take Age, Gender, and Income (Customer Demographics) into consideration which can potentially have an influence on customer churn.
2. This analysis does not include customers from other telecom companies. If the data from other telecom companies was available, it could potentially provide differing results.
3. With the removal of Bandwidth\_GB\_Year due to multicollinearity, the Initial Model’s adjusted R-Squared value was reduced from 0.99 to 0.42. This means that the Initial Model only explains 42% of the variance in Tenure.

Recommendations:

1. Based on my analysis, I would recommend taking a deeper dive into the reasons of why a Month-to-Month Customer churns. This variable has the largest impact (based on coefficients) on Tenure and the strongest correlation of -0.58 with Tenure compared to the other two independent variables.
2. Based on my analysis, I would also recommend taking a deeper dive into the quality of the Fiber Optic Internet Service. Considering that Fiber Optic is the more advanced option for Internet Service that our telecom company provides, it is odd that it has a negative impact on Month-to-Month customer Tenure. There could be a potential quality issue that we are overlooking.
3. Based on my analysis, I would keep our price point for the Monthly Charges the same as our current rates. The current rates for Monthly Charges have a positive impact on Month-to-Month customer Tenure.

**Part VI: Demonstration**

Please see Panopto Presentation Link.