**Task 2: Logistic Regression Analysis**

**Part I: Research Question**

“Disregarding Customer Demographics, what are the top 5 features that have the most impact on Customer Churn at our Telecom Company?”

Goals/Objectives:

1. Identify the top 5 features with the most impact. In my experience as an Analyst in Healthcare for a For-Profit Health System, it is better to provide a more granular approach and identify 3-5 “Low-Hanging Fruit” to focus our efforts on. These 3-5 features should be features that have the most impact on our desired outcome.
2. The F1\_Score for my adjusted/reduced model should be above 90%.
3. Recommend a potential course of action based on analysis.

**Part II: Method Justification**

Assumptions of a Logistic Regression Model:

For my analysis, I will be using binary logistic regression. In binary logistic regression the following assumptions/criteria must be met:

1. The dependent variable to be binary and nominal.
   1. Churn is a binary, nominal variable.
2. All observations in the dataset must be independent of each other.
   1. Each observation in my dataset represents an individual customer with no influence on the other.
3. There must also be little-to-no multicollinearity amongst the independent variables.
   1. After viewing my correlation matrix, I noticed that Tenure is highly correlated with Bandwidth\_GB\_Year. To avoid multicollinearity, I decided it would be best to remove Bandwidth\_GB\_Year feature from my analysis.
   2. When encoding/creating “dummy” columns, if you do not remove the 1st dummy variable, there will be multicollinearity. To avoid multicollinearity, I used the dummy\_cols() from the fastDummies library and set the remove\_first\_dummy parameter to TRUE. This removes the 1st dummy variable and avoids multicollinearity.
4. There must also be a sufficiently large enough sample size.
   1. In this case, I am using 10,000 observations.

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.

Logistic Regression is the appropriate technique as my research question is based around the binary categorical dependent variable of Churn (Churn is either “Yes” or “No”). My research question poses a binary classification problem which cannot be solved by multiple linear regression. When creating a binary dummy variable for Churn, Churn can either be 1 for Yes or 0 for no. In Multiple Linear Regression, the dependent variable is continuous without bounds and can extend past 0 or 1. Because of this, logistic 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. Perform MinMax scaling (normalization) on all features to remove bias in features with larger ranges and create a “level” playing field. This will ensure a workable range of 0 to 1 while normalizing the variables.
4. Remove variables with multicollinearity.
5. Rename any variables with spaces used for my adjusted model

Please see the following for the Outputs of my Summary Statistics

1. “Task2 – Dataset Summary Output.PNG”
2. “Task2 – Correlation Matrix.csv”
3. “Task2 – Correlation Heatmap”

Explanation of All Variables:

After normalizing all variables, I used summary() to get a summary of all variables and their distributions. Each variable has a min and max range of 0 and 1, respectively.

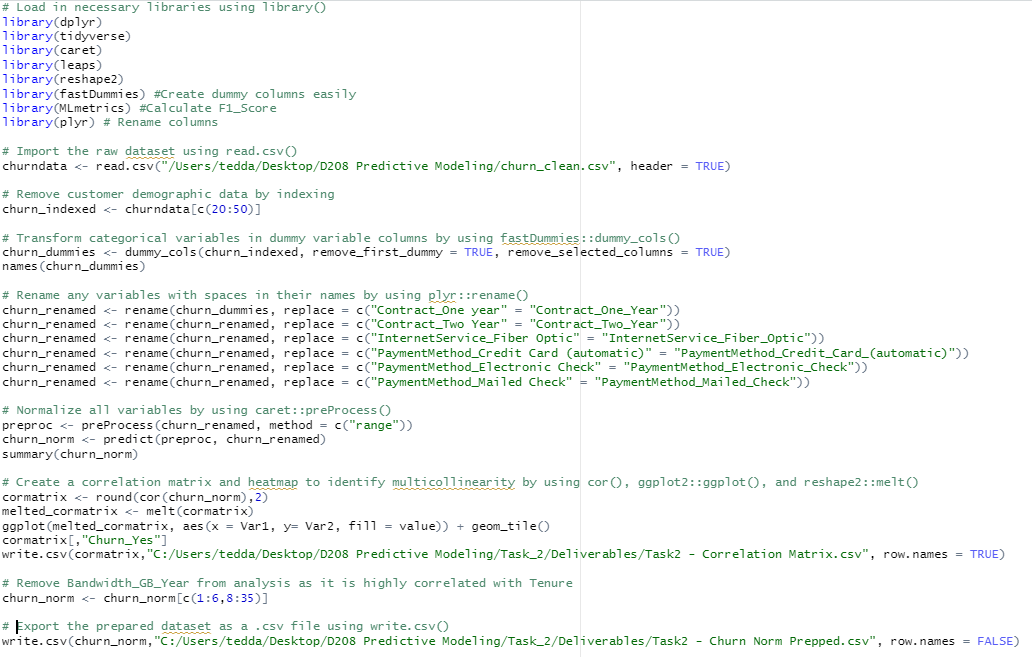
Explanation of Relationships of all 6 variables used in my reduced model:

1. Churn\_Yes (Categorical; Dependent) – The median is 0 which indicates that most customers did not Churn. However, our mean is 0.265 which is not close to the median. This indicates that the distribution of Churn\_Yes is skewed to the right. This can be seen in the Univariate Histogram Visualization for Churn\_Yes.
2. Tenure (Continuous; Independent) – The mean and median for Tenure 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 0.097 – 0.852 (possible min and max are 0 and 1). When referencing the boxplot of Churn\_Yes vs. Tenure, churned customers had a lower median Tenure. This indicates that more churned customers left earlier on in their tenure. As seen in the Univariate Histogram Visualization for Tenure, the distribution of Tenure is bimodal.
3. 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 0.286 – 0.575 (possible min and max are 0 and 1). When referencing the boxplot of Churn\_Yes vs MonthlyCharge, churned customers had a higher median monthly charge. This can be seen by the median for churned customers (~0.58) and non-churned customers (~0.38).
4. Contract\_One\_Year (Categorical; Independent) – The median is 0 which indicates that most customers do not have a One-Year contract. In viewing the histogram distributions of our contract type (one year and two year), most customers are month-to-month customers. You can also see that the distribution is skewed to the right. This could be a potential growth opportunity for our telecom company by providing an incentive for customers entering a contract with us.
5. Contract\_Two\_Year (Categorical; Independent) – The median is 0 which indicates that most customers do not have a Two-Year contract. In viewing the histogram distributions of our contract type (one year and two year), most customers are month-to-month customers. You can also see that the distribution is skewed to the right. This could be a potential growth opportunity for our telecom company by providing an incentive for customers entering a contract with us.
6. InternetService\_Fiber\_Optic (Categorical; Independent) – The median is 0 which indicates that most 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 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.

Data Preparation Steps with Code

1. Load in necessary libraries using library()
2. Import the raw dataset using ***read.csv()***
3. Remove customer demographic data by indexing
4. Transform categorical variables in dummy variable columns by using ***fastDummies::dummy\_cols()***
   1. Remove 1st dummy variable to avoid multicollinearity
5. Rename any variables with spaces in their names by using ***plyr::rename()***
6. Normalize all variables by using ***caret::preProcess()***
7. Create a correlation matrix and heatmap to identify multicollinearity by using ***cor()****,* ***ggplot2::ggplot(), and reshape2::melt()***
8. Remove Bandwidth\_GB\_Year from analysis as it is highly correlated with Tenure
9. Export the prepared dataset as a .csv file using write.csv()

Annotated Code for Data Preparation:



Please see the following for my Univariate and Bivariate Visualizations:

1. “Task2 – Correlation Heatmap.pdf”
2. “Task2 – Univariate Distributions.pdf”
3. “Task2 – Bivariate Distributions.pdf”

Please see “Task2 – Prepped Data Set.csv” for a copy of my prepared dataset.

**Part IV: Model Comparison and Analysis**

Initial Model Output:

Please see “Task2 – Initial Model Summary Output.PNG”

Feature Selection Output:

Please see “Task2 – Regression Subset Summary Output.PNG".

Reduced Model Output:

Please see “Task2 – 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 5 variables which impact customer Churn. Regression Subsets allows me to find the best combination of 5 variables with the most impact by setting the parameter nvmax = 5. The result minimizes the AIC score to as low as possible with 5 variables selected. The 5 variables selected by regsubsets() were Tenure, MonthlyCharge, InternetService\_Fiber\_Optic, Contract\_One\_Year, and Contract\_Two\_Year. To test my logic, I also checked a model with 4 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 AIC of 5454.3.

Model evaluation criteria: (AIC, Confusion Matrix, F1\_Score)

I used three model evaluation metrics to measure the effectiveness of my model:

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Initial Model** | **Reduced Model (5 Variables)** | **Reduced Model (Only 4 variables)** |
| **AIC** | 4430.5 | 4749.7 | 5454.3 |
| **ConfusionMatrix\_Accuracy** | 0.9036 | 0.8942 | 0.8775 |
| **ConfusionMatrix\_Recall(Pos Pred Value)** | 0.9302 | 0.9223 | 0.9081 |
| **ConfusionMatrix\_Precision(Sensitivity)** | 0.9393 | 0.9348 | 0.9272 |
| **F1\_Score** | 0.9347 | 0.9285 | 0.9175 |

Compared to the initial model, the reduced model was slightly less accurate based on the above statistics (A higher AIC and lower Confusion Matrix and F1\_Scores). However, this is an incredibly good thing considering the original model consisted of 34 variables while the reduced model consisted of only 5 variables. I have included the Model Evaluation Criteria for a reduced model with only 4 variables. The comparison of the initial model against one with only 4 variables indicates the reduced model with 4 variables is much less accurate than one with 5 variables.

Please see the following for my Outputs and Calculations:

1. “Task2 – Initial Model Confusion Matrix Summary Output.PNG”
2. “Task2 – Adjusted Model Confusion Matrix Summary Output.PNG”
3. “Task2 – 4\_Variable Model Confusion Matrix Summary Output.PNG”
4. “Task2 – F1\_Scores of All 3 Models”

Please see “Task2 – All Code.R” for my entire code.

**Part V: Data Summary and Implications**

Reduced Model Regression Equation:

ln(y) = -2.07 – 7.56x1 + 11.56x2 – 3.11x3 – 3.19x4 – 2.11x5

ln(Churn\_Yes) = -2.07 - 7.56Tenure + 11.56MonthlyCharge - 3.11Contract\_One\_Year - 3.19Contract\_Two\_Year – 2.11InternetService\_Fiber\_Optic

Interpretation of Coefficients:

Please see “Task2 – Coefficients Summary Output.PNG”

1. Intercept – The log-odds of Churn\_Yes is -2.07 when all other variables are equal to 0.
2. Tenure – For every unit increase in Tenure, the log-odds of a customer churning (Churn\_Yes) decreases by 7.52 when all other variables remain the same
3. MonthlyCharge – For every unit increase in MonthlyCharge, the log-odds of a customer churning (Churn\_Yes) increases by 11.56 when all other variables remain the same.
4. Contract\_One\_Year – If a customer has a one-year contract with our telecom company, the log-odds of the customer churning (Churn\_Yes) decreases by 3.11 when all other variables remain the same.
5. Contract\_Two\_Year – If a customer has a two-year contract with our telecom company, the log-odds of the customer churning (Churn\_Yes) decreases by 3.19 when all other variables remain the same.
6. InternetService\_Fiber\_Optic – If a customer has the fiber optic internet service with our telecom company, the log-odds of the customer churning (Churn\_Yes) decreases by 2.11 when all other variables remain the same.

Statistical and Practical Significance of Model:

My Reduced Model has met all objectives/goals I initially set out for this analysis. It is statistically significant with a p-value of less than 2.2 x e-16. Also, the accuracy of the model is about 89.4% at the 95% confident interval. The model also has a F1\_Score of 92.85% which is above 90%. This signifies that my reduced model does a great job of representing the dataset.

Please note that this dataset is purely fictitious and may not represent an accurate depiction of a real-world telecom company’s data. I would use this reduced model in this specific, fictitious business/practical setting since it represents the data well based on our model evaluation metrics.

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.

Recommendations:

1. Provide incentives to customers in tiers based on how long they’ve used our services that reduce the customer’s monthly charge amount by a small, competitive percentage after their first one or two-year contracts. The longer the contract, the greater the percentage since Two-Year has a slightly greater coefficient. This percentage should be comparable to current market charge rates.
2. Offer a slightly reduced price for Fiber Optic Internet Service to existing customers that are still using DSL or have no internet included in their plan.

**Part VI: Demonstration**

Please see Panopto Presentation Link.