­

**Predictive Modeling**

Sam Spencer

College of Information Technology,

Western Governors University

Dr. Keoiona Middleton

November, 25 2022

Table of Contents for Each Rubric

Part I: Research Question

[Describe Purpose, Summarize Research Question and Define Objectives:](#RubricA) 3

***Part II: Method Justification***

[Summarize Model Assumptions,](#RubricC1) Python Benefits and Why Choose Logistic Regression7

***Part III: Data Preparation***

[Describe](#RubricC4) Data Preparation, Summary Statistics, Visualizations and Code 8

Part IV: Model Comparison and Analysis

Compare Initial and Reduced Models, Justify Variable Selection 15

Provide Reduced Logistic Regression Models15

Analyze Data Using Reduced Mode, Explain with Residual Error15

Part V: Data Summary and Implications

Summarize 16

Part VI: Demonstration

[Video](#RubricF) 17

[Sources for Third-Party Code](#RubricG) 17

[Sources](#RubricH) 18

# Hospital Readmission Problem

Our chain of hospitals are facing a crucial problem with readmissions, so much that there are external organizations that have begun penalizing hospitals that have excessive readmissions within a 30-day time span. Many of the hospitals are underprepared for various reasons, but our goal is to assist in the mitigation of these charges by being proactive with our data.

For use to address the problem, we first need to get an understanding of our data and patient profiles, which patients are more likely to go to the hospital, are there specific demographics that are readmitted more than others? Our analysis will touch on these topics with the provided dataset.

## A1 - Research Question

Our research question is can we determine if various customer factors contribute to the reason for contacting the telecom company for tech support.

**A2 - Objectives and Goals**

There are several factors that contribute to customer churn, some of the primary causes are price and poor customer service. My objective with this analysis is to assist our chain of telecom companies in understanding if there is any correlation in service uptime, and how many times a customer has to contact them for customer support

B1 – Summary of Assumptions

There are 5 assumptions when it comes to multiple regression models:

1. “Linear relationship: There exists a linear relationship between each predictor variable and the response variable.
2. No Multicollinearity: None of the predictor variables are highly correlated with each other.
3. Independence: The observations are independent.
4. Homoscedasticity: The residuals have constant variance at every point in the linear model.
5. Multivariate Normality: The residuals of the model are normally distributed.”

(Zach,2020)

## B2 – Tool Benefits

Anaconda Navigator is a desktop GUI, its interface contains several useful applications for data analysis, but what is used for this project is the Jupyter Notebook application. Jupyter along with Python are my tools of choice for code. Python is a very useful language, within Python, I’ll use several libraries and packages for our analysis that allow us to visualize, clean, and manipulate the data. The libraries and packages that I’ll be using are:

* Pandas & NumPy –Pandas and NumPy are used in conjunction with each other, pandas is used to import and manipulate data, which is useful for our project, and NumPy, which is required to use Pandas, works with our numerical data
* MatPlotLib
* Seaborn
* Sklearn

**B3** – **Appropriate Technique**

Multiple regression is defined as, “A statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.” Multiple regression is the most obvious choice for our analysis because what it does great is determining how strong the relationship is with 2 or more independent variables and 1 dependent variable.

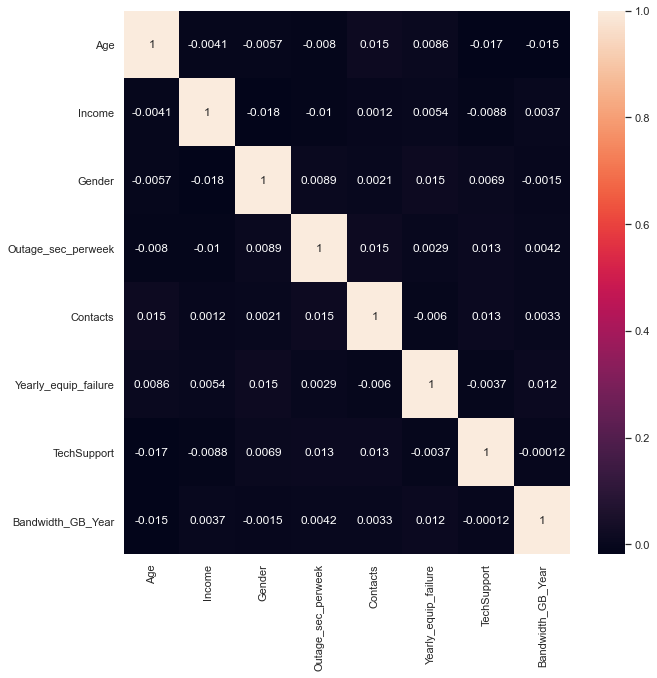


Figure 1 – Correlation heatmap

**C1** – **Data Goals**

Our original data set contained 10,000 rows with 50 columns many of which were not useful to us. We had to first understand the data, search for nulls and dupes. I converted categorical values to binary values if 0/1 for tech support and 1,2,3 for gender

Text

Description automatically generated

Figure 2 – Looking for null values

Next, dropped columns unrelated to the hypotheses and converted predictive categorical data to numbers to analyze.

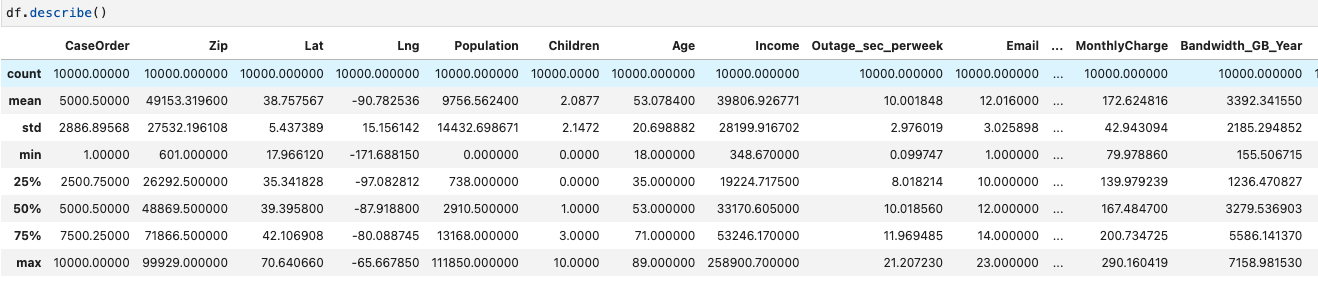


Figure 3 - Summary Statistics for the Data Frame

C2 – Summary Statistics

Using pandas describe, we can put together a profile of our average customers:

* Age ranges from 18-89 with an average customer age of 53
* Income ranges from $348-$258,900 with the average being around $40,000
* Outages per week (in seconds) ranges from 3-23 seconds with the average being 12 seconds of downtime per week
* The average user uses 3392 GB of data per year

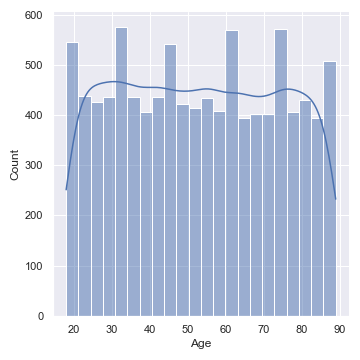
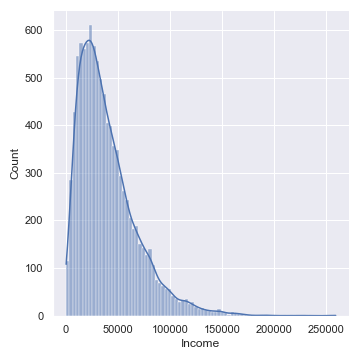
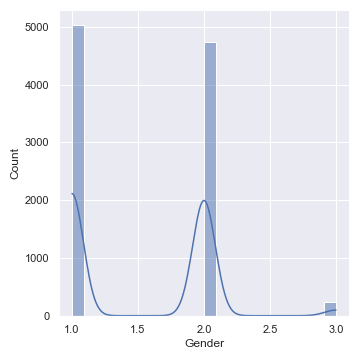
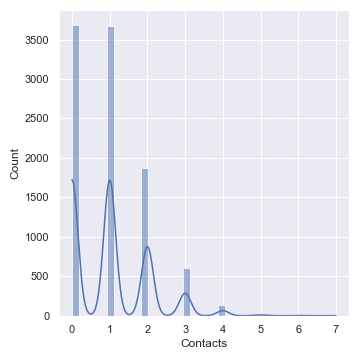
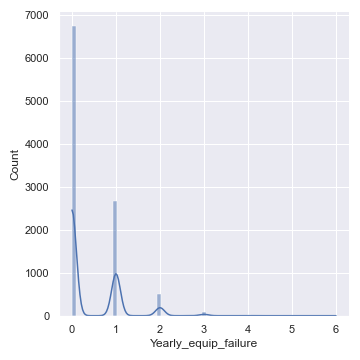
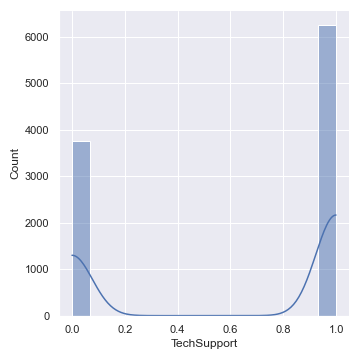
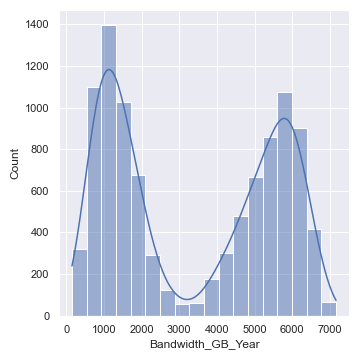
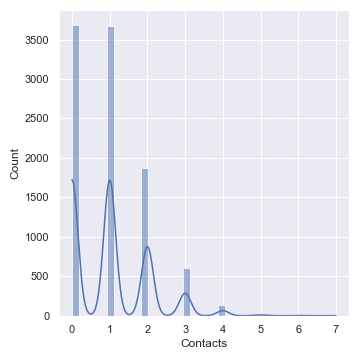
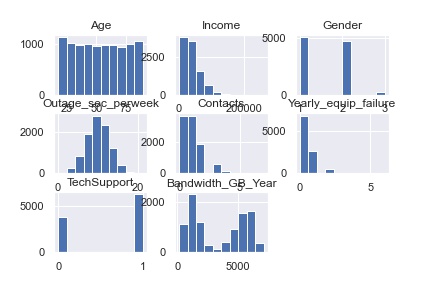
C3 – Steps to Prepare Data

1. Import libraries and packages
2. Import data using pandas
3. Check for null and dupes
   1. Remove and replace if null and dupes exist
4. Drop columns that weren’t necessary for our research
5. Convert categorical variables into binary values of 0/1
6. Export prepared dataset

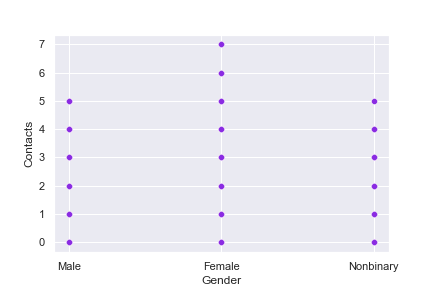
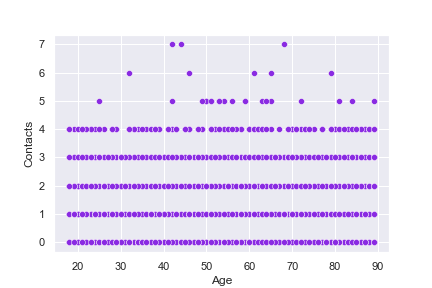
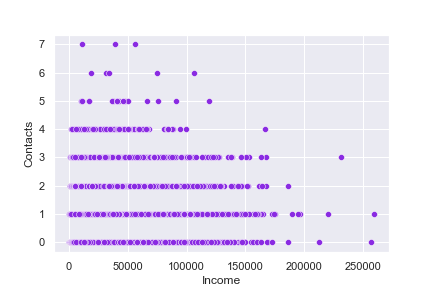
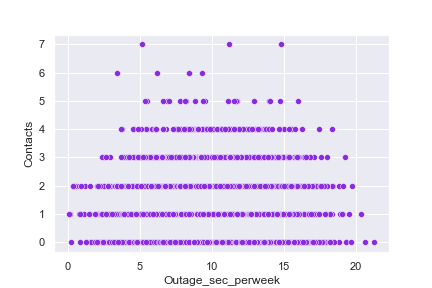
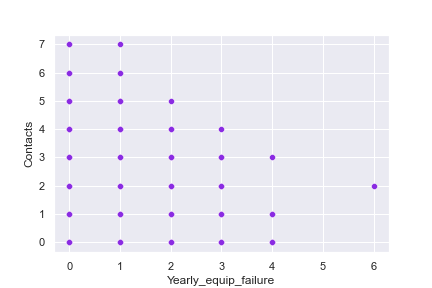
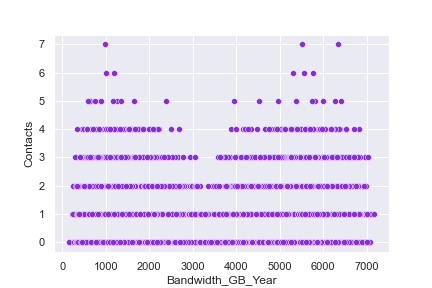
C4 – Visualizations

The following univariate and bivariate visualizations are from the cleaned and reduced data set. The bivariate visualization includes the target variable.

Univariate:



Bivariate:



**C5** – **Prepared data set**

Prepared data set is attached with submission

**D1** – **Initial multiple regression model from *all* predictors that were identified in Part C2**

**A picture containing text, receipt, screenshot

Description automatically generated**

X\_train2 = sm.add\_constant(X\_train)

model = sm.OLS(y\_train,X\_train2)

print(model.fit().summary())

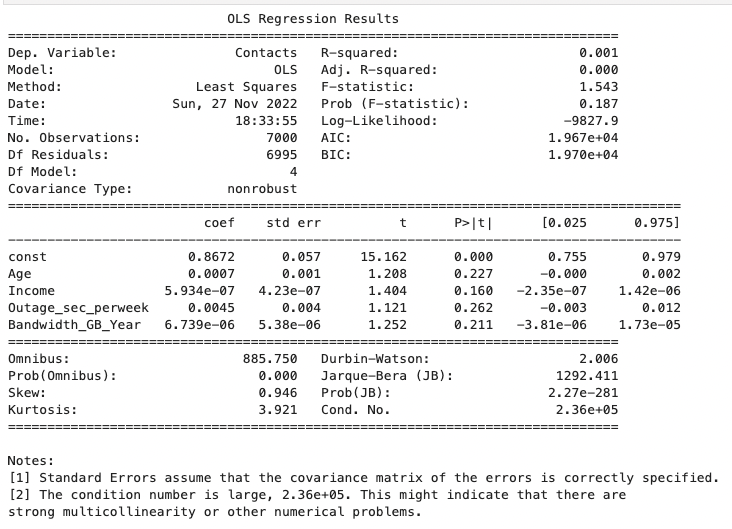
p\_vals = dict(model.fit().pvalues[1:])

df

**D2** – **Justification of Model Reduction**

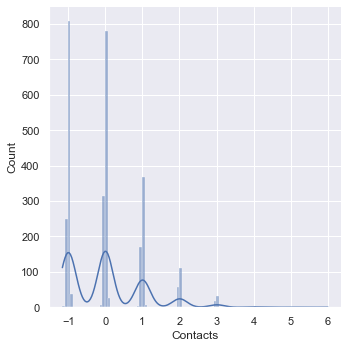
The variables we were reduced using their p-values. A p value of .3 or lower returned 'Age', 'Income', 'Outage\_sec\_perweek', 'Bandwidth\_GB\_Year'.

**D3** –Reduced Multiple Regression Model



**E1 – Model Comparison**

To arrive to our conclusion, we used stepwise regression to help in deciding which variables we should omit from the model. According to (XXX), Stepwise regression is achieved by “Trying out one independent variable at a time and including it in the regression model if it is statistically significant or by including all potential independent variables in the model and eliminating those that are not statistically significant. Some use a combination of both methods and therefore there are three approaches to stepwise regression:”. In addition, I used P-values to determine which coefficients were significant enough to remain in our model with a .3 threshold. Below is our residual plot and evaluation methods



* MSE states that the total number of contacts will be roughly 0.7346414355983919
* With a negative R2 score our model lacks confidence.

**E3 – Code**

Code will be attached

**F1 – Results**

**H – Sources for Third-Party Code**

* Help using Markdown: <https://www.markdownguide.org/basic-syntax/>
* How to use Dummies: [How to use Pandas get\_dummies to Create Dummy Variables in Python](https://www.marsja.se/how-to-use-pandas-get_dummies-to-create-dummy-variables-in-python/)
* Univariate Plots: [12 Univariate Data Visualizations With Illustrations in Python](https://www.analyticsvidhya.com/blog/2020/07/univariate-analysis-visualization-with-illustrations-in-python/)
* Pandas Help: <https://pandas.pydata.org/docs/user_guide/index.html#user-guide>
* Numpy Help: <https://numpy.org/doc/stable/>
* Seaborn Help: <https://seaborn.pydata.org/>
* Matplotlib Help: <https://matplotlib.org/>
* Sklearn Help: <https://scikit-learn.org/>

## References

Boggs, W. (2015, April 8). *Depression and Anxiety Linked to Higher Hospital Readmission Rates*. Hmpgloballearningnetwork.com. Retrieved October 28, 2022, from https://www.hmpgloballearningnetwork.com/site/pcn/article/depression-and-anxiety-linked-higher-hospital-readmission-rates

Centers for Disease Control and Prevention. (2022, September 8). *Health and economic costs of chronic diseases*. Centers for Disease Control and Prevention. Retrieved November 25, 2022, from https://www.cdc.gov/chronicdisease/about/costs/index.htm

Hayes, A. (2022, October 8). *Multiple Linear Regression (MLR) Definition, Formula, and Example*. Investopedia. Retrieved October 30, 2022, from https://www.investopedia.com/terms/m/mlr.asp

Waskom, M. L. (2021). *Seaborn Pairplot*. seaborn.pairplot - seaborn 0.12.1 documentation. Retrieved November 1, 2022, from https://seaborn.pydata.org/generated/seaborn.pairplot.html

Weiss, A. J. (2021, July 21). *Overview of Clinical Conditions With Frequent and Costly Hospital Readmissions by Payer, 2018*. Overview of Clinical Conditions With Frequent and Costly Hospital Readmissions by Payer, 2018 #278. Retrieved October 28, 2022, from https://www.hcup-us.ahrq.gov/reports/statbriefs/sb278-Conditions-Frequent-Readmissions-By-Payer-2018.jsp#second

Zach. (2021, November 16). *The five assumptions of multiple linear regression*. Statology. Retrieved November 25, 2022, from https://www.statology.org/multiple-linear-regression-assumptions/