Shay Snyder

April 30, 2020

CSCI 4957 Data Analytics

**Capstone Project**

***Introduction / Description***

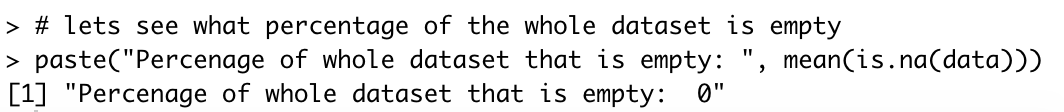
A health care organization believes they suffer from over $30 million in losses annually due to the readmission of patients who are discharged from the hospital too soon. Keeping all patients in the hospital longer, regardless of their condition, is a costly measure and inconvenient to patients.

I was recently hired by the hospital to use my data analytics skills to develop a model that predicts the readmission risk. Doctors will use my model to decide whether to discharge a patient without delay. The hospital shared a dataset of 10,000 patients in a csv file named 10kPatients.csv.

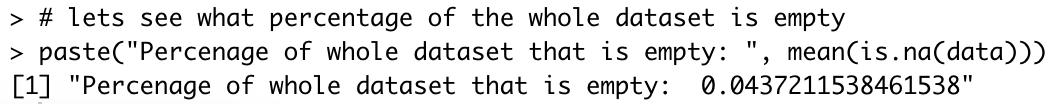
To help doctors understand and trust my approach, I will thoroughly explains my thought processes and various visualizations.

***Data Preprocessing***

First and foremost, I wanted to calculate the percentage of the dataset that is empty. If a majority of the data turns out to be empty, the integrity of the data comes into question. Below is the function that I used to calculate the percentage of errors.



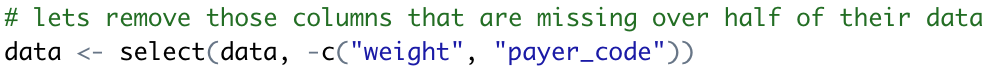
Since the function returns 0, this implies that 0% of the dataset is empty when that is not reality. The data set must contain empty Strings or other variables to represent empty values. Upon further investigation, we see that many empty values are accounted for by empty Strings (“”), question marks (“?”), “Not Available”, and “Not Mapped”.



Now my calculations show that approximately four percent of the dataset is empty. This value is more realistic, and we can now assume our data is ready to be processed further. If any of the columns are missing over half of their observations, we will drop them as they provide little data. Below is a list of the columns that had missing data and the percentage of missing values in each.

|  |  |
| --- | --- |
| Column 2 | 2.221% |
| Column 5 | 95.92% |
| Column 6 | 13.73% |
| Column 7 | 5.97% |
| Column 8 | 9.73% |
| Column 10 | 53.41% |
| Column 11 | 41% |
| Column 18 | 0.02% |
| Column 19 | 0.59% |
| Column 20 | 2.08% |
| Column 50 | 0.02% |
| Column 51 | 0.59% |
| Column 52 | 2.08% |

Let’s remove column 6 (“weight”) and column 10 (“payer\_code”) using the dplyr package.



Now that we have remove the columns that are mostly empty, lets attempt to fix some of the missing values in the data using kNN imputation (as taught in previous weeks). Before kNN imputation, the dataset was missing 7802 entries. After kNN, the dataset contained zero missing entries. That is simply amazing!

***Exploratory Analytics (Data Visualization and Plotting)***

At this point of the data analytics process, we want to better understand our data through a variety of avenues.