Shay Snyder

April 30, 2020

CSCI 4957 Data Analytics

**Capstone Project**

***Introduction / Description***

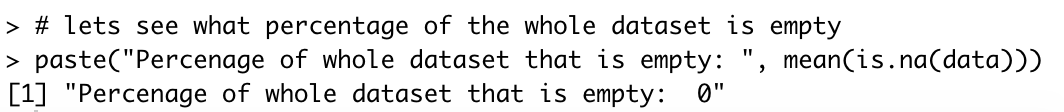
A health care organization believes they suffer from over $30 million in losses annually due to the readmission of patients who are discharged from the hospital too soon. Keeping all patients in the hospital longer, regardless of their condition, is a costly measure and inconvenient to patients.

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.

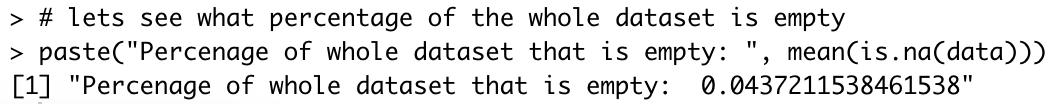
To help doctors understand and trust my approach, I will thoroughly explains my thought processes and various visualizations.

***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.



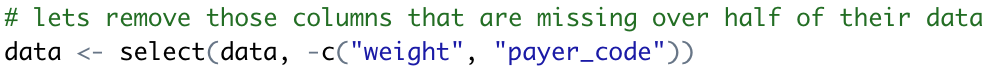
Since the function returns 0, this implies that 0% of the dataset is empty when that is not reality. The data set must contain empty Strings or other variables to represent empty values. Upon further investigation, we see that many empty values are accounted for by empty Strings (“”), question marks (“?”), “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! Note: If we continue to use the weight column throughout the rest of the process, we would need to make sure a patient’s weight is possible, if their weight isn’t possible, we would have to find some way to correct it.

***Exploratory Analytics – Is the dataset skewed?***

At this point of the data analytics process, we want to better understand our data through a variety of avenues. The first step in my process is to better visualize and understand the distribution of our data. The following plots show the distribution of various aspects in the dataset such as: patients readmitted against non-readmitted patients, patients on Diabetic medication vs patients who are not, patient race, patient age, patient gender, and patients passing away vs staying alive.

A screenshot of a cell phone

Description automatically generated

The histogram (created using ggplot) above visualizes the distribution of patient readmissions within the 10kPatients.csv dataset. We can see the data has a slight skew as the data is slightly more representative of non-readmitted patients (~60%) rather than readmitted patients (~40%).

A screenshot of a social media post

Description automatically generated

The histogram (created using ggplot) above visualizes the distribution of patient’s taking within the 10kPatients.csv dataset. We can see the data has a skew as the data is more representative of patients on diabetes medicine (~75%) rather than patients who aren’t (~25%).

A screenshot of a cell phone

Description automatically generated

The histogram (created using ggplot) above visualizes the distribution of patient race within the 10kPatients.csv dataset. We can see the data has an extreme skew towards Caucasians, and a slight skew for African Americans. Thus, the data is highly representative of Caucasians (~75%), slightly representative of African Americans (~21%), and hardly representative of Asians, Hispanics, and others (~4% combined between the three).

A picture containing drawing

Description automatically generated

The histogram (created using ggplot) above visualizes the distribution of patient age within the 10kPatients.csv dataset. We can see the data has a negative skew towards patients of older age, specifically ages 50 – 90 as they account for approximately 80% of the patients within the dataset. This means the data is highly representative of older individuals but not so representative of younger individuals.

A screenshot of a cell phone

Description automatically generated

The histogram (created using ggplot) above visualizes the distribution of patient gender within the 10kPatients.csv dataset. We can see the data almost has a normal distribution. This means that males (~47%) and females (~53%) are both represented extremely well.

A screenshot of a cell phone

Description automatically generated

The histogram (created using ggplot) above visualizes the distribution of patients who were able to leave the hospital in a living state. We can see the data is highly representative of individuals who were living at the end of their stay (~98%) rather than those who perished (~2%).

***Exploratory Analytics – Pattern Mining and Association Rule Learning***