Shay Snyder

April 30, 2020

CSCI 4957 Data Analytics

**Capstone Project**

***Introduction / Description***

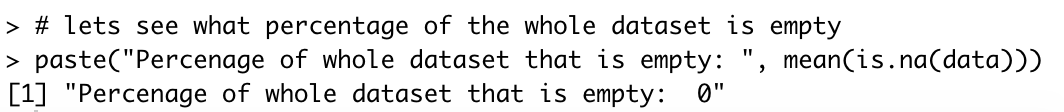
A health care organization believes they suffer from over $30 million in losses annually due to the readmission of patients who are discharged from the hospital too soon. Keeping all patients in the hospital longer, regardless of their condition, is a costly measure and inconvenient to patients.

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.

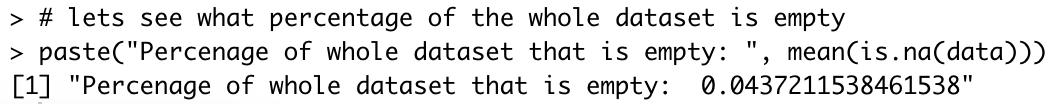
To help doctors understand and trust my approach, I will thoroughly explain my thought processes and various visualizations.

***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.



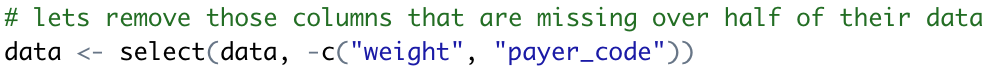
Since the function returns 0, this implies that 0% of the dataset is empty when that is not reality. The data set must contain empty Strings or other variables to represent empty values. Upon further investigation, we see that many empty values are accounted for by empty Strings (“”), question marks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

At this point of the project, it is time to see if there are interesting patterns or association rules in the 10kPatients.csv dataset. Note: At this point, I switched to the slightly pre-processed dataset from D2L. Note: I also had various warnings throughout this process that said the Apriori function could not discretize the data even though the data wasn’t linear. I would really appreciate some insight into this issue. The rest of this section will answer and expand upon the following questions:

* What are the rules that cause a patient not to be readmitted?
* What are the rules that cause a patient to be readmitted?
* What are the rules that cause a patient to pass away?

The following image depicts the algorithm used to determine what rules cause a patient not to be readmitted to the hospital:

A screenshot of a cell phone

Description automatically generated

As you can see, the Apriori algorithm was crucial in the process. Since we are sorting the rules based upon lift, we can be more lenient on the minimum support and confidence. This allows us to see a greater range of rules and the prevalence to the question at hand. The following images depicts the top ten rules that cause a patient not to be readmitted to the hospital.

A screenshot of a cell phone

Description automatically generated

Since the above rules have the greatest lift compared to their counterparts, they have the strongest positive correlation with patients not being readmitted to the hospital. Let’s examine the rule with the largest lift (Note: I will not repeat this for all rules as this step can be easily used to evaluate all of them), {race = African American, gender = female, discharge disposition id = discharged to home} with 0.0591 support, 0.6724 confidence, and 1.1141 lift. This rule implies than an African American woman that is discharged to their home, is unlikely to be readmitted to the hospital. The rule’s support implies that 5.91% of the data conlytains all of the items included in both the left and right sides of the association rule. The rule’s confidence implies that 67.24% of the time, the rule proves to be true. The rule’s lift implies that there is a positive correlation between the occurrence of the rule and the patient not being readmitted to the hospital. Looking at the whole list, I find it very interesting that the majority of the rules contain individuals of African American race but the previous section showed that African Americans only accounted for ~20% of the data compared to Caucasians which accounted for approximately ~75% of the data. Does this mean that African Americans are much less likely to be readmitted to the hospital when compared to Caucasians and slightly less than all other races?

The following image depicts the algorithm used to determine what rules cause a patient to be readmitted to the hospital:

A screenshot of a cell phone

Description automatically generated

Again, the Apriori algorithm was crucial in the process. Since we are sorting the rules based upon lift, we can be more lenient on the minimum support and confidence. This allows us to see a greater range of rules and the prevalence to the question at hand. The following images depicts the top ten rules that cause a patient to be readmitted to the hospital.

A close up of text on a white background

Description automatically generated

Since the above rules have the greatest lift compared to their counterparts, they have the strongest positive correlation with patients not being readmitted to the hospital. Let’s examine the rule with the largest lift (Note: I will not repeat this for all rules as this step can be easily used to evaluate all of them), {race = Caucasian, gender = Female, admission type id = Urgent, discharge disposition id = Discharged/transferred to home health service} with 0.0070 support, 0.6364 confidence and 1.605 lift. The rule’s 0.0070 support implies that 0.7% percent of the data in the dataset contains all of the items included in both the left and right sides of the association rule. The rule’s 0.6364 confidence implies that 63.64% of the time this rule proves to be true. The rule’s 1.605 lift implies that there is a strong positive correlation between the occurrence of these attributes and the patient being readmitted to the hospital. I find it quite interesting that all of these rules have the same discharge disposition id: Discharges/transferred to home with home health service and most of these rules have the same admission type id: Urgent. This means that individuals who come to the hospital with urgent medical conditions and are discharged to a home health service have a high possibility of being readmitted to the hospital in the future.

The following the image depicts the algorithm used to determine what rules cause a patient to pass away during their stay at the hospital.

A screenshot of a cell phone

Description automatically generated

Again, the Apriori algorithm was crucial in the process. Since we are sorting the rules based upon lift, we can be more lenient on the minimum support and confidence. This allows us to see a greater range of rules and the prevalence to the question at hand. The following images depicts the top ten rules that cause a patient pass away during their hospital stay.

A screenshot of text

Description automatically generated

Since the above rules have the greatest lift compared to their counterparts, they have the strongest positive correlation with patients passing away during their stay at the hospital. Let’s examine the rule with the largest lift (Note: I will not repeat this for all rules as this step can be easily used to evaluate all of them), {admission type id = Emergency, diabetes medication = no, readmitted = No} with 0.0058 support, 0.0594 confidence, and 3.0166 lift. This rule implies than an individual without any known major health conditions, such as diabetes, that comes to the hospital with an emergent situation, is unlikely to survive their stay at the hospital. The rule’s support implies that 0.58% of the data contains all of the items included in both the left and right sides of the association rule. The rule’s confidence implies that 5.94% of the time, the rule proves to be true. The rule’s lift implies that there is a strong positive correlation between the occurrence of the rule and the patient not being readmitted to the hospital. Looking at a whole list, we can see that most of these individuals likely didn’t have any major known health problems but came to the hospital with an emergent situation. This conclusion coincides logical thought as these individuals most likely had extremely serious complications such as heart attack, stroke, or major injuries.

***Predictive Analytics***

As the final step in the process, it is time to take a patient’s attributes (for example, 20 attributes i.e., race, gender, …., diabetes medication) and predict whether or not they will be readmitted to the hospital. This study will implement two machine learning algorithms, Naïve Bayes, and Decision Trees in order to predict patient outcomes. These algorithms will be compared to determine which model is most effective. The data was split in a 4 to 1 ratio of training and test data (80% training and 20% testing). The following graphs will display the results of the various algorithms and examine their capability in the form of a confusion matrix.

A screenshot of text

Description automatically generatedA screenshot of text

Description automatically generated

The leftmost confusion matrix depicts the results of the decision tree upon the 10kPatient.csv dataset. We can see that 57.8% of the time, our model correctly determines whether a patient will be readmitted to the hospital. Sensitivity specifies that 70.42% of the time, the model predicts ‘yes’, when it’s actually ‘yes’. Specificity specifies that 37.89% of the time, the model predicts ‘no’, when it’s actually ‘no’.

The rightmost confusion matrix depicts the results of the Naïve Bayes model upon the 10kPatients.csv dataset. We can see that 63.35% of the time, our model correctly determines whether a patient will be readmitted to the hospital. Sensitivity specifies that 83.5% of the time, the model predicts ‘yes’, when it’s actually ‘yes. Specificity specifies that 32.57% of the time, the model predicts ‘no’, when it’s actually ‘no’. We can see that this model (Naïve Bayes) is much more effective than the decision tree.

A screenshot of text

Description automatically generated

Finally, the hospital can now implement the Naïve Bayes model in order to combat their extreme revenue loss due to patient readmissions Assuming that every patient contributes a roughly equal amount to the 30 million in excess spending, the hospital, with the help of this research, they should be able to reduce that figure by 63% (in a perfect world). That is of 20.1 million dollars per year! The graph above depicts the financial savings from this research over the next 30 years. These excess funds could be spent on other hospital necessities in order to better help their community.