CO3093-Big Data and Predictive Analytics Report

CW Assignment 2 by Zachary and Sarim

**Data exploration and findings**

A Data Exploration was implemented in data\_exploration.py to identify trends and patterns within Diabetic patients. Within this segment, we will be proving or disapproving the following hypothesis based on the data we have gathered.

Hypothesis 1: Age has a higher impact on readmission.

A graph of age vs readminister

Description automatically generated

Appendix Figure 1 - Age vs Readmission

As shown in Figure 1 above, we can clearly see a trend that the likeliness of readmission increases as the patient’s age goes up. The readmission rate is lowest for the youngest age group and progressively rises, peaking in the 70-80 age bracket before slightly declining for the oldest age group. This supports the hypothesis that age has a higher impact on readmission, proving our hypothesis that Age has a higher impact on readmission.

Hypothesis 2: African Americans are more likely to be re-admitted than other ethnic groups.

A graph of blue rectangular bars

Description automatically generated with medium confidence

Appendix Figure 2 - Race vs Readmission

The second graph shows that African Americans have a mean readmission rate that is slightly higher than other ethnic groups like Asians and Hispanics, but not markedly so. On top of that it is shown that people of Caucasian Race have a higher mean readmission rate. That along with the fact that all racial groups fall within a narrow range of readmission rates show that this hypothesis does not hold up and race might not be an influential factor in readmission rates.

Hypothesis 3: Women patients are more likely to be re-admitted than men.

A blue rectangular object with white text

Description automatically generated

Appendix Figure 3 - Gender vs Readmission

In the 3rd Graph, we see the comparison between the readmissions rates of males and females. As shown in the graph, it reveals females have a marginally higher rate. However, the difference displayed above is not substantial enough to draw a conclusion, with no other trends or patterns to look at we are not able to conclusively support the hypothesis that women are more likely to be readmitted than men.

Hypothesis 4: Diagnosis types have a higher impact on re-admission rates.

To further study the data related to this hypothesis, we will be looking at 3 different graphs.

A graph of blue lines with black text

Description automatically generatedA graph of blue and black text

Description automatically generatedA graph of blue vertical lines with text

Description automatically generated with medium confidence

As we observe the different graphs, we can see that the rate of readmission varies heavily based on the different diagnosis types. Out of all the diagnoses, diseases such as heart failure, heart disease, kidney disease and chronic airway obstruction stand out with the highest rates of readmissions. These 3 graphs constructed from the data display a wide variation in readmission rates across different diagnosis types, suggesting that certain conditions are indeed more associated with readmission.

**Description of model and evaluation**

As shown in our python file first\_model.py, we created a model that uses ML to predict diabetic patient’s readmission. We made use of the sub-dataset for the following columns: ['num\_medications', 'number\_outpatient', 'number\_emergency', 'time\_in\_hospital', 'number\_inpatient','encounter\_id', 'age', 'num\_lab\_procedures', 'number\_diagnoses','num\_procedures', 'readmitted']

This model was made using RandomForestClassifier, a popular ensemble learning method that constructs a multitude of decision trees at training time to output the mode of the classes for classification tasks.

Evaluation score

Confusion Matrix: [[2053 382] [ 992 519]]

F1 score: 0.64

Example findings and clusters from the first model

A graph of a number of bars

Description automatically generated A graph of a number of people

Description automatically generated with medium confidenceA graph of a number of objects

Description automatically generated with medium confidenceA graph of a number of diagnoses

Description automatically generated

A graph with different colored bars

Description automatically generatedA graph with a number of colored bars

Description automatically generated with medium confidence

A graph with numbers and a number of points

Description automatically generated with medium confidence

The analysis of hospital readmission data reveals significant patterns across various patient demographics and clinical characteristics. Through the model, patients are grouped into clusters that exhibit distinct traits, such as length of hospital stay, age, number of lab procedures, number of diagnoses, and medical procedures undertaken, all of which contribute to varying readmission risks. For example, some clusters indicate shorter hospital stays, while others suggest a higher complexity of care due to more procedures and diagnoses. The readmission distribution graph shows the potential for targeted post-discharge interventions, especially in clusters with higher readmission rates.

These statistics offer insights that healthcare staff would pay attention to mitigate risks of readmission of diabetic patients.

**Improved Model and K-clustering**

Description of the model and improvement from the first model

The improved model incorporates interaction terms and polynomial features, enabling the capture of complex, non-linear relationships between variables that the initial model, relying on raw features, might miss. By one-hot encoding categorical variables such as race and gender, the improved model can better utilize these features, which were previously not utilized.

In this improved model, new techniques were implemented such as SMOTE (Synthetic Minority Over-sampling Technique) and GridSearchCV.

SMOTE was employed to generate synthetic samples of the minority class, leading to a more balanced dataset that allows the model to learn more about the underrepresented class, this helps improve the model’s sensitivity predicting.

GridSearchCV was used to find optimal model parameters, such as the number of trees in the forest (n\_estimators), the depth of the trees (max\_depth), and the minimum number of samples required to split an internal node (min\_samples\_split). This exhaustive search ensures that the model is not just fitting well to the training data but also generalizing effectively to unseen data.

Evaluation score

Confusion Matrix: [[1446 366] [ 723 425]]

F1 Score: 0.694

K-Means Clustering

The K-Means clustering graphs below depict the segregation of patient data into clusters based on two features: the number of lab procedures and the number of medications. These graphs use a scatter plot format where each point represents a patient, with the position indicating their respective counts for lab procedures or medications, and the color representing the cluster to which they have been assigned to. The color intensity in both graphs represent cluster density, with more intense colors showing a higher concentration of patients within that specific cluster region.

A screenshot of a computer screen

Description automatically generated

Appendix Figure 4 - No. Medications vs Time in Hospital

This graph shows patient clustering based on the number of medications prescribed. These clusters appear to be defined by medication frequency. Some of these clusters seem to include patients on multiple medications. This might imply that certain patients could be dealing with serious health issues, hence requiring complex pharmacological management. other clusters might correspond to patients with fewer lab procedures, reflecting less severe health issues.

A screen shot of a graph

Description automatically generated

Appendix Figure 5 - No. Lab Procedures vs No. Diagnoses

The 2nd graph shows the distribution of patients across clusters based on the number of lab procedures they underwent. This identifies the clusters based on a range of frequencies. For example, there may be a cluster that specifically groups patients who have had a higher number of lab procedures, which could indicate more severe health conditions that require more treatment. On the other hand, other clusters might correspond to patients with fewer lab procedures, reflecting less severe medical situations.

**Take aways and Actions that can be taken**

Based on all the given data, The clusters may reveal insights into patient demographics or treatment profiles that correlate with readmission rates. For example, if older patients are predominantly in clusters with higher medication counts, this could indicate age-related polypharmacy risks.

Clinical Complexity: The observed trend reversal could be attributed to an increase in clinical complexity or severity of conditions among patients requiring a higher number of procedures. As patients' healthcare needs become more complex, they may require additional interventions and procedures, leading to an uptick in procedural frequency.

Healthcare Interventions: Alternatively, the increase in procedural frequency beyond a certain threshold may reflect proactive healthcare interventions aimed at managing or addressing underlying health issues. Healthcare providers may escalate the frequency of procedures to monitor patients' conditions closely or implement targeted treatments or interventions.