CO3093 Big Data and Predictive Analytics

Group 169

Develop a predictive model to predict diabetic patient readmission.

# Part 1 – building a basic predictive model based on a given specification.

## Data cleaning and transformation:

The initial dataframe shape was 101,766 rows and 50 columns. After dropping columns with 50% or more missing values the number of rows was reduced to 49, the weight column was removed from the dataset. Based upon our intuition weight potentially could have been a good predictor for readmission rates as it has a direct effect on a patient’s health. However, the absence of data heavily reduced its statistical power, increasing the probability of rejecting the null hypothesis. Therefore, no data cleaning technique would have meaningful effect in filling the missing data, such as using the mean or forward fill.

Furthermore, we also dropped columns that had over 95% of their values that were the same. These included many different medication changes that a patient may have had, this left the data frame to have 33 columns. The reason for this removal is because it would be hard to find justifications as to why they were readmitted or not as the columns don’t provide any extra information.

Once the columns that had lacking amounts of data or didn’t offer beneficial information were removed it was time to look at the other columns. Firstly, the age column, due to the ages being registered in the dataset as 10-year intervals they were transformed to use the middle value in that interval. This conversion between a categorical to quantitative variable would better allow our model to identify that the age 5 is closer to 15, whereas using categorical variables the model wouldn’t know the difference because it doesn’t recognise natural ordering.

After dealing with the age column, we then replaced all missing values within the diag 1-3 columns with 0’s. Afterwards, we dropped all rows that still had missing values, this left us with a total 26,755 rows of data which is a huge difference in the amount of data compared to the initial dataset.

Numerical features: ['age', 'time\_in\_hospital', 'num\_lab\_procedures', 'num\_procedures', 'num\_medications', 'number\_outpatient', 'number\_emergency', 'number\_inpatient', 'number\_diagnoses']

Categorical features: ['admission\_source\_id', 'change', 'encounter\_id', 'diag\_1', 'metformin', 'insulin', 'diabetesMed', 'diag\_2', 'patient\_nbr', 'glipizide', 'diag\_3', 'discharge\_disposition\_id', 'race', 'admission\_type\_id', 'payer\_code', 'pioglitazone', 'A1Cresult', 'glyburide', 'max\_glu\_serum', 'glimepiride', 'medical\_specialty', 'gender', 'rosiglitazone', 'readmitted']

After looking at the data within each of the features we identified that some columns were miss-matched in terms of whether they were categorical or numerical. Some of which would include the id columns, which were initially in the numeric category, for example the encounter\_id and patient\_nbr. These columns clearly are categorical as they are used as identifiers, the impact of this change will be further discussed in the model section of part 1.

Additionally, we also removed outliers in the numerical columns that were 3 standard deviations away from the mean. This also helped improve the statistical power as it reduced the variability in our numerical columns. Once the outliers were removed, we were left with a total of 24,858 rows of data. Finally in terms of the part 1 data cleaning and transformation we removed duplicates in the patient\_nbr column leaving us with a grand total of 18,575 rows of data to explore and build our model upon.

## Data exploration:

In order to fulfil meaningful data exploration a few more steps had to be taken to better plot and understand the data. A first step taken was to join both the diabetic dataset and an international classification of diseases dataset, this would help us further understand the differences in the three diag columns rather than seeing codes that mean nothing. Another step taken was to check for distinct values in various categorical columns, this would help reduce clutter as many of the values were very similar and came under the same overarching category. Therefore, this approach was also used on the diag columns, as a result reduced the number of unique values from 541, 541 and 546 to 9, 9 and 9 respectively. Finally, we also converted the readmitted column to binary values to suggest if they had been readmitted or not, this was instead of the previous three values (NO, <30 and >30). Before this conversion was made it was identified that a total of 11,649 patients were not readmitted, compared to the 5,383 patients greater than 30 days and 1,543 readmitted within a 30-day period.

In order to fulfil data exploration appropriate graph types must be chosen to better explore and help understand the data. Due to the readmitted column being a categorical variable this mean that we decided to use various bar and line plots to interpret the data and draw conclusions.

Hypothesis 1: Age has a higher impact on readmission.

Firstly, looking at Figure 1 it highlights that there are greater number of diabetic patients between the ages of 45 and 85. Using external research [1] it shows that the average age in the US to be diagnosed with diabetes is 50 years old, this therefore explains the result we see in the graph. The dramatic drop in patients beyond 85 can also be acquitted to expiration, which is also an expected observation. However, this is not the case for determining whether age has a higher impact on readmission rates. Once again looking at Figure 1 it shows that the number of readmitted patients proportionately increases at the same rate as the number of patients not readmitted. Furthermore, a line plot, Figure 2, was also created to better show the change in readmission for each age. This graph also backed up our initial findings from the bar chart. However, it does show that the ratio between readmission rates at the age range of 75 – 95 is much closer. Which may influence our model building.

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

By evaluating Figure 3 it clearly outlines that most of the data collected is from Caucasian individuals, with the second highest being African Americans. However, based on this data it is hard to compare and confirm the hypothesis due to a redundancy in variation for each other ethnic groups. Therefore, this means it is much harder to draw conclusions as the data is extremely biased. Furthermore, based on the data that is presented in Figure 3 and using some statistical analysis was determined that 33% of the African American patients were readmitted, compared to the 38% readmission of the Caucasian patients, disproving the hypothesis.

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

Once again by looking at the bar and line plots, Figures 4 and 5 respectively, it shows that there does not seem to be much variation in whether gender has an impact on readmission. The data this time, however, can be seen as much more reliable as the total amount of records for both genders is much closer therefore resulting in less of a bias. Moreover, the percentage of Female readmissions is 37% of all Female records, compared to 36% for Males, highlighting that we should reject the hypothesis.

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

Looking at the different diagnosis types in relation to readmission has shown once again like the previous hypothesis that there may not be enough data for the rarer types of diagnoses. Looking at Figures 6 it showcases that there are much more patients diagnosed with circulatory diseased compared to various other types of diagnoses.

However, when comparing between Figures 6, 7 and 8 the values change only slightly for circulatory and massively for the diabetes and the “other” category, although there is not visible change in ratio of the not readmitted and readmitted patients. Therefore, making it hard to draw conclusions on diagnoses that have the higher number of records. This can be compared the injury/poisoning category which has a much better ratio between if patients are not reemitted and are. Therefore, this may mean that this would be a good category to investigate when building our model. However, the total number of patients diagnosed with injury/poisoning is far too low and will be incomparable when using other diagnoses that would bias the model, this can be said for the various other low data amount categories.

Model building:

Based on the following predictor features: ['num\_medications', 'number\_outpatient', 'number\_emergency', 'time\_in\_hospital', 'number\_inpatient', 'encounter\_id', 'age', 'num\_lab\_procedures', 'number\_diagnoses', 'num\_procedures', 'readmitted']

Part of the model building process we split our data into training and test datasets. Whilst splitting the data we decided to use a 70/30% split, without splitting in a stratified fashion or shuffling the data beforehand. We then decided we would use logistic regression to build our first model. From our model we got an initial accuracy score of 0.647 for our training data, compared to an accuracy of 0.652 for the test data. This shows that we are not overfitting as the test score is higher than the one using our training data. In terms of the accuracy, 65% seems high as we did not select the most impactful predictors, along with not fully transforming/cleaning the data to remove other biases we have identified. The cross-validation score of the model was 0.646, using a 10-fold cross-validation, which is very close to the previously observed accuracy. Following on we plotted a confusion matrix of the predicted and actual outputs of our model, Figure 9. As you can see our model is lack lustre when it comes to predicting the outcome of readmitted (A high rate of false positives).

We then decided to try another model instead, this time using random forest to predict if a patient would be readmitted. When using random forest, we got an increased training and test accuracy score compared to logistic regression, being 0.656 and 0.653 respectively. However, when looking at the cross validation mean score it gave a value of 0.373 which is extremely low compared to our accuracy score, this means that the random forest model is overfitting as our test and training error has a large gap between them. Figure 10 highlights the confusion matrix for random forest, this time being much worse at predicting values where patients are not readmitted.

Appendix

Figure 1:

Chart, bar chart

Description automatically generated

Figure 2:

Chart, line chart

Description automatically generated

Figure 3:

Chart, bar chart

Description automatically generated

Figure 4:

Chart, bar chart

Description automatically generated

Figure 5:

Graphical user interface

Description automatically generated with low confidence

Figure 6:

Chart, bar chart

Description automatically generated

Figure 7:

Chart, bar chart

Description automatically generated

Figure 8:

Chart, bar chart, waterfall chart

Description automatically generated

Figure 9:

Chart, treemap chart

Description automatically generated

Figure 10:

Chart, treemap chart

Description automatically generated

Figure 11:

References

[1] <https://news.northwestern.edu/stories/2021/september/black-and-mexican-american-adults-develop-diabetes-younger/>

[2]