CO3093 - Big Data and Predictive Analytics

Predicting Hospital Readmissions: Diabetes Patient Analysis

Part 1: Building up a basic predictive model

Data cleaning and transformation

For our initial diabetes readmission classification model we stuck to a simple data-cleaning process.
The dataset patient information from 130 American hospitals. In this dataset - the age column was originally an object data type with a range of numbers surrounded by brackets as its values. This made it unusable for our data analysis, so we stripped the brackets and transformed the data into middle point values - making the column easier to work with in our data analysis.

Data exploration

Regarding data exploration, we want to find the trend on aspects of the readmittance rate of patients. The columns that we will be looking at are age, race, gender and admission type. To start the analysis of the data set, we investigated which of the columns mentioned has the highest impact on the readmittance rate. For comparison, we have used a bar plot to have a simple overview of the trend in each comparison. We will first discuss each figure and the trends that can be inferred from the data. Second, we will come up with hypotheses and discuss their validity.

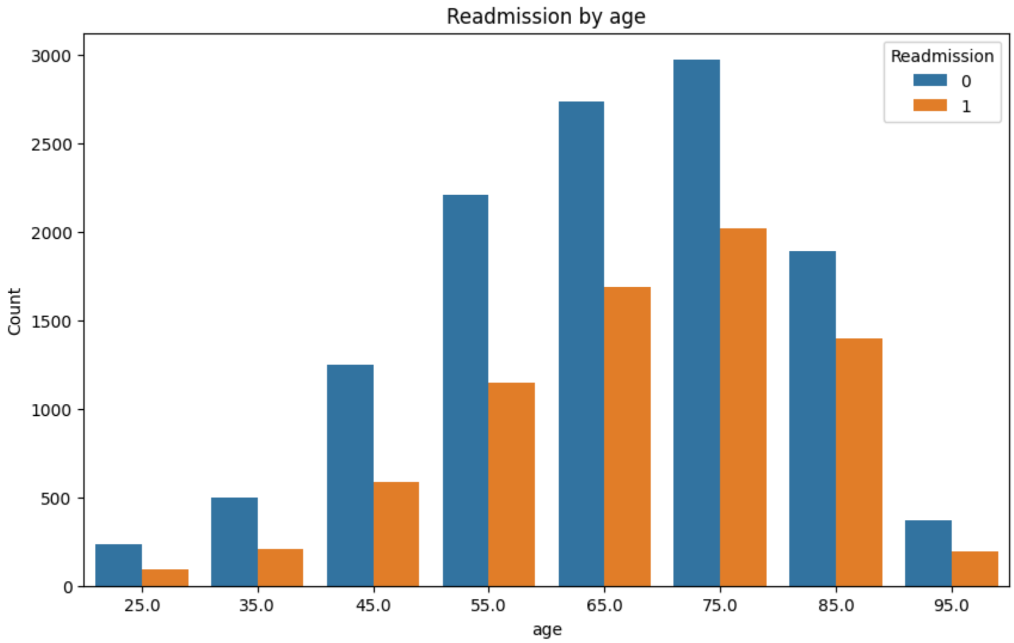


Figure 1.1 : Bar plot of readmission by age

Figure 1.1 shows the relationship between the age of patients with readmission. The immediate detail we can see is that most patients in the data set are between 55 and 85 years old. This highlights most diabetic patients in the data are elderly people. The highlighted trend implies that as people get older, they are more susceptible to diabetes. The inference that can be made is the trend of readmission by age. As the age increases, the rate of readmission increases. For example, in the age category of 35 and 45, the readmission rate is barely 50% of the non-readmission count but in the age category of 65 and 75, the readmission rate is higher than 50% of the non-readmission count. Overall, this trend shows the impact of age on the readmission rate where older patients are more likely to be readmitted than younger patients.

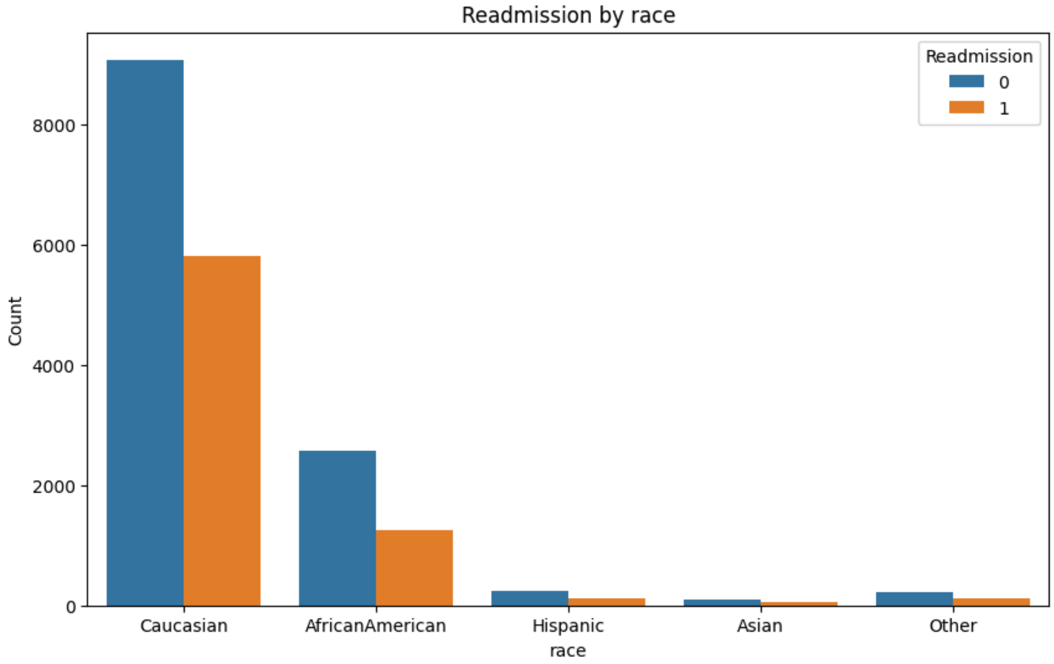


Figure 1.2 : Bar plot of readmission by race

In Figure 1.2, we can observe the trend between race and readmission. In the figure, there is one major observation that can be seen - the large gap between caucasian patients and patients of other races. This shows that the majority of the patients are caucasian but also African Americans have the second largest patient count in the data set. In this figure, we can see that Caucasians have a higher readmission rate than African Americans with the rate being above 50% and African Americans only having a readmission rate of about 50%.

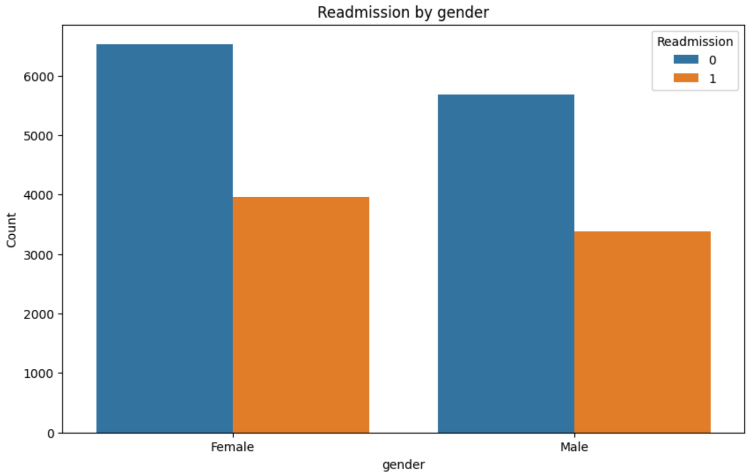
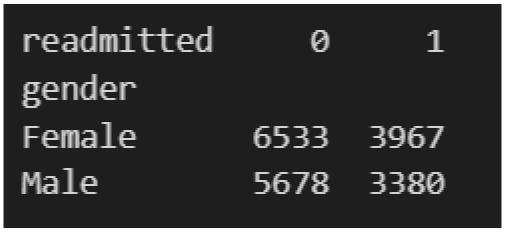


Figure 1.3 : Bar plot of readmission by gender



```
readmitted     0     1
gender
Female      6533  3967
Male        5678  3380
```

Figure 1.4 : Data for the bar plot in figure 1.3

Figure 1.3 allows us to observe the relationship between the genders and the readmission rate. One observation that can be made is that there are more female patients than male patients but there is no significant gap between the genders. Based on this observation, there is not enough evidence to conclude that gender has a significant impact on the patient's susceptibility to diabetes. Regarding readmission between the two genders, based on figure 1.3, the readmission rate is almost similar but based on figure 1.4 which shows the exact data of figure 1.3, female has a 61% readmission rate while male has a 60% readmission rate. This implies that gender does not affect the readmission rate based on the data set.

For our diagnoses' graphs, we tried to create them from the cleaned dataset as we did with the rest of the data exploration, but from this, we could not draw any conclusions or see any clear trends. So, we created our diagnoses' graphs from the original dataset to better capture the nuances for these columns. Figures 1.4-1.6 and below are our findings.

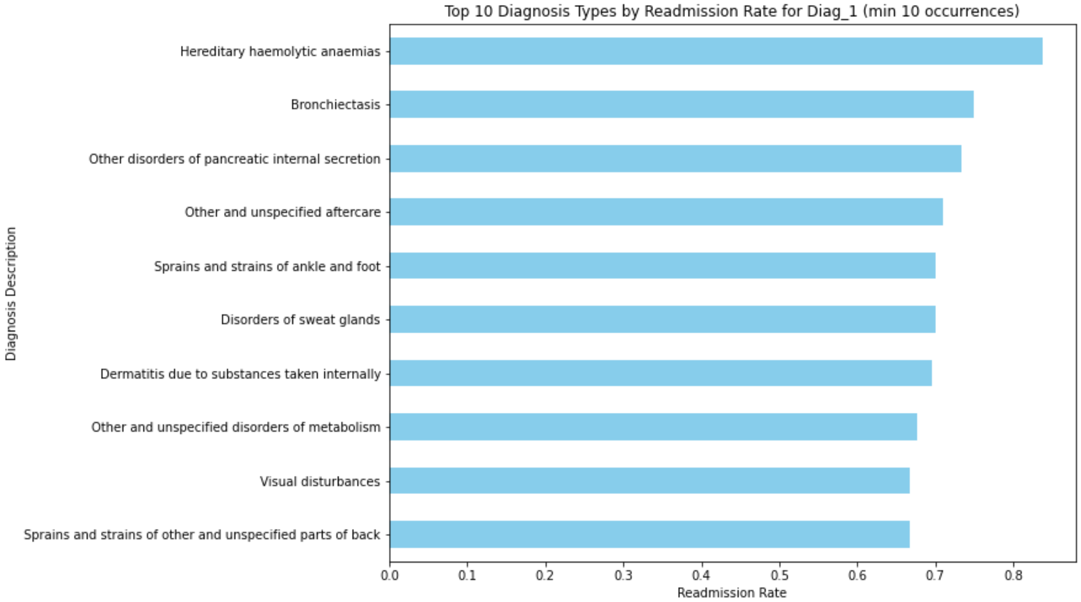


Fig 1.4

For the first diagnosis column (diag_1), the analysis revealed that hereditary 'haemolytic anaemia' had the highest readmission rate at 83.82%, with a notable occurrence of 68 cases, as depicted in Figure 1.4. The significant readmission rate for this diagnosis suggests a potentially critical nature of the condition that may require ongoing medical attention beyond initial treatment.

Following this, bronchiectasis and other disorders of pancreatic internal secretion also showed high readmission rates of 75% and 73.33%, respectively, though with lower occurrences. This pattern highlights that even less common conditions can pose substantial readmission risks.

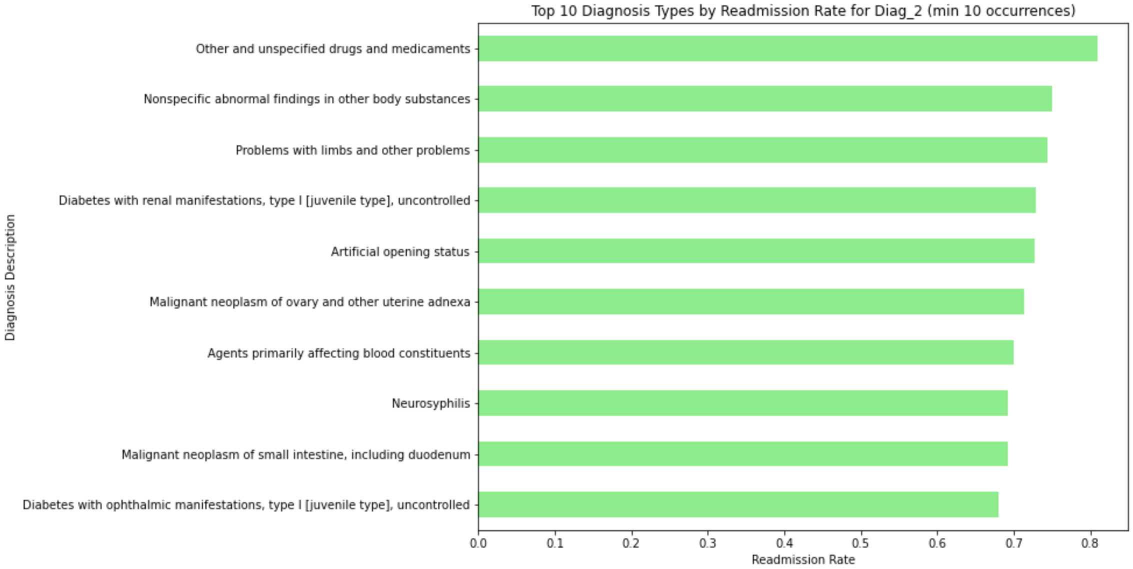


Figure 1.5 :

The secondary diagnosis column (diag_2) presented a different set of conditions leading the readmission rates, as shown in Figure 1.5. Here, the top spot was taken by 'Other and unspecified drugs and medicaments', with an 80.95% readmission rate over 21 cases. It emphasises the challenges in managing drug interactions and side effects, which can often lead to readmissions.

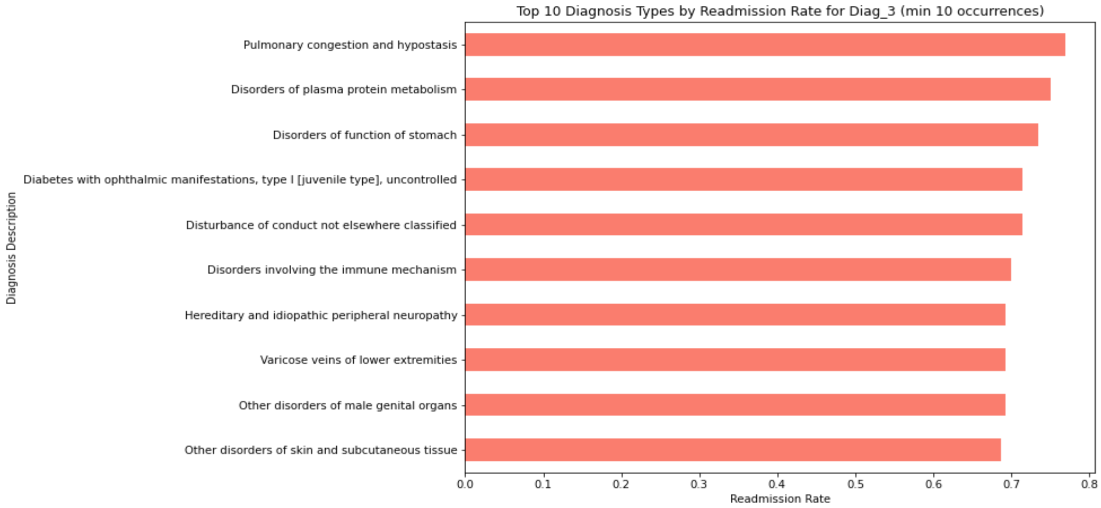


Figure 1.6

As for the third diagnosis (diag_3), 'Pulmonary congestion and hypostasis' led to the readmission rates of 76.92% over 13 occurrences, as illustrated in Figure 1.6. Disorders related to plasma protein metabolism and the function of the stomach were also among the top diagnoses with high readmission rates, suggesting these conditions may commonly lead to complications or co-morbid scenarios that result in readmission.

From the observations made in Figures 1.1, 1.2, 1.3, 1.4, 1.5 and 1.6 we have a good overview of how certain features affect readmission for patients. Some hypotheses were made before analysing the data set which are :

1) Age has a higher impact on readmission.
2) African Americans are more likely to be readmitted than other ethnic groups.
3) Women patients are more likely to be readmitted than men.
4) Diagnosis types have a higher impact on readmission rates

For the first hypothesis, we can confidently say that age has the highest impact on readmission than other aspects we mentioned. As the age increases, the rate of readmission increases. Based on the other aspects, the difference in race does not have a significant difference between other races and gender have the same admission rate for both male and female. As such, it can be concluded that age has a higher impact on readmission than other aspects.

Next, the second hypothesis is rejected as from Figure 1.2, we conclude that Caucasians have higher readmission rates compared to other ethnic groups.

The third hypothesis is also rejected as based on figures 1.3 and 1.4, it is observed that both female and male patients have a 61% and 60% admission rate. The difference of 1% is not conclusive evidence to conclude that female patients are more likely to be readmitted than men.

The fourth hypothesis has been accepted as the data suggests that certain diagnosis types are indeed associated with higher readmission rates, lending support to the hypothesis. In particular, Figure 1.4 demonstrates that certain conditions such as hereditary haemolytic anaemia, bronchiectasis, and disorders of pancreatic internal secretion are associated with notably high readmission rates, which could imply a stronger impact on readmissions relative to other factors. This could be due to the severity of these conditions, the complexity of their treatment, or the potential for these conditions to exacerbate other comorbidities. Furthermore, Figures 1.5 and 1.6 reinforce this point.

Model Building

For our predictive model, we utilised Logistic Regression because our problem is a classification problem, meaning that a relationship between numerical variables and a categorical data result needs to be investigated.

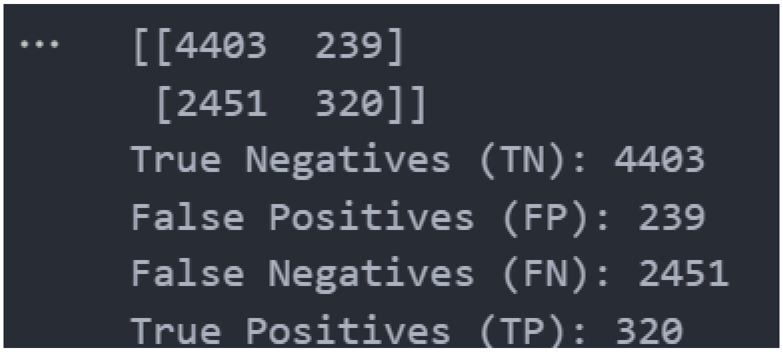
When we ran Recursive Feature Elimination, we generally found that the columns ['number_outpatient', 'number_emergency', 'number_inpatient'] were our most impactful variables in determining the outcome of a patient being readmitted. However, the columns do change occasionally. Nonetheless, the three columns suggest that the number of times a patient visits the hospital, whether short or long-term, has a strong correlation with whether or not they will be readmitted to the hospital.

From this we can theorise a relationship: the more a diabetic patient visits the hospital over a given year, the higher the likelihood of revisiting. I.e if a patient has been visiting regularly, for whatever reason, they are likely to continue that behaviour.

The R-squared values for our training and test data were 0.645 and 0.637 respectively. For the training data, the score indicates our model explains 64.5% of the variance of readmission concerning the model’s independent variables to a satisfactory degree. In short, it can explain 64.5% of all variability in the model.

When looking at the performance against our test data, a score of 63.7 indicates our model generalises fairly well. In both cases, there is substantial room for improvement.

For cross-validation, we performed a 10-fold cross-validation and had a mean score of 0.642. Again in line with the performance from



```
[[4403  239]
 [2451  320]]
True Negatives (TN): 4403
False Positives (FP): 239
False Negatives (FN): 2451
True Positives (TP): 320
```

Figure 1.7

In the confusion matrix above - we used a threshold value of 0.5. In the matrix, you can see that we had a high rate of false negatives. This implies our model is biassed to not re-admitting patients. Conversely, it rarely made false positives.

Part 2: Improved Model

For our improved model we took a variety of approaches and adopted Extreme Programming (XP). One group member went about increasing the rows of our dataset. By dropping two columns with large holes of missing data (‘payer_code’ and ‘medical_specialty’) and then filling some of the null values within numerical columns with mean-fill, mode-fill and zero-fill strategies. This brought the column count from ~18,000 up to ~69,000.

Afterwards, the same member manually adjusted the selected columns used in the recursive feature elimination section of the model building. This was handled manually by testing multiple 6-column combinations of the numerical columns due to computational constraints (we could not directly run 11 feature columns with any reasonable amount of data, especially not 68,000).

Finally, they tried to hyper-tune the parameters of the logistic regression model using the sklearn package GridSearchCV. Grid search is a hypertuning technique where you make a grid of hyperparameter values and compare the results of multiple combinations of them via cross-validation. There was unfortunately no notable improvement from this.

In tandem, the second group member utilised the data exploration findings and decided to emphasise the race feature in the model. This was achieved by using hot-one encoding to transform race into values of 0 or 1. 1 for Caucasian as in our data exploration we noted that Caucasians have higher readmission rates compared to other ethnic groups and 0 for all the other races in the data set.

After this, recursive feature elimination was carried out again and the selected features were:
['number_emergency', 'number_inpatient', 'race']

However, the columns do change occasionally.

Next, the second group member used several models to improve performance from the original model - Logistic Regression, Random Forest, eXtreme gradient boosting (XGBoost) and finally a Voting Classifier which is a combination of the above three models in an ensemble leveraging the wisdom of the crowd.

After comparing the results of both approaches, we chose to follow the second group members' approach.

Which lead to two improved models:

The XGB Classifier model with an accuracy score of 0.6457 and the Voting Classifier model with an accuracy score of 0.648.

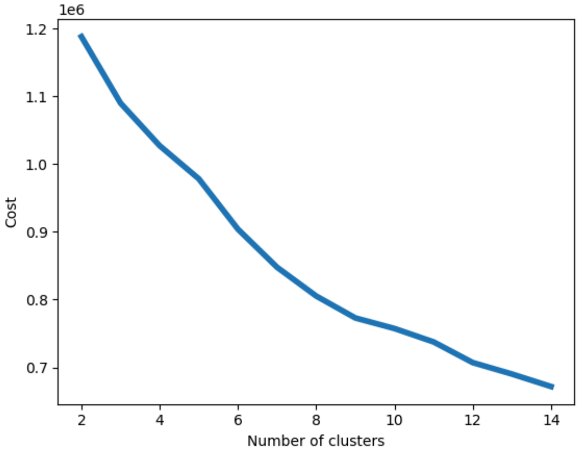


Figure 2.1 : Elbow method

Using the Elbow method as seen in figure 2.1, we have decided that the k = 3 cluster is the most appropriate number to use due to the sharp decline in the figure. Following this, we proceeded to use the Principal Component Analysis (PCA) to find a common trend between each points in each cluster.

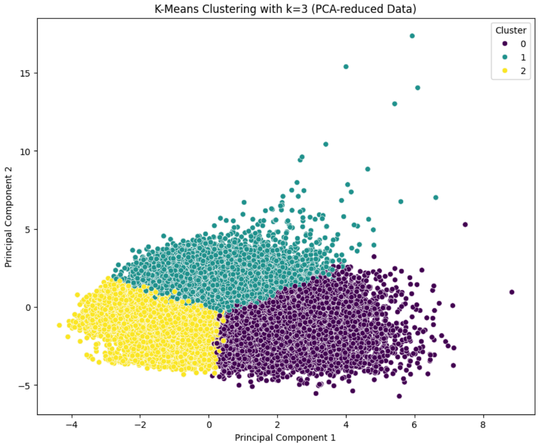


Figure 2.2 : PCA graph

In figure 2.2, we can see the correlations between all three clusters where each cluster is closer to each other with the exception of certain data in the cluster.