|  |
| --- |
| University of Leicester |
| CO3093 – Big Data & Predictive Analytics CW Assignment |
| Classification and Clustering |

|  |
| --- |
| Student Number: 199009725  3-24-2022 |

**Data Exploration**

**Chart, treemap chart

Description automatically generatedHeatmap of Correlation**

* To get an idea of what the relationship of each variable is to one another, I created a heat map to showcase the correlation.
* This gives an idea of how the outcome (readmitted) moves in relation to the other variables.

**Proving or Disproving Hypotheses**

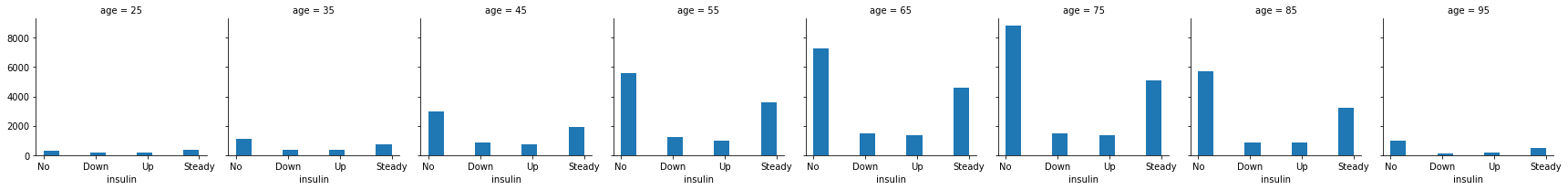
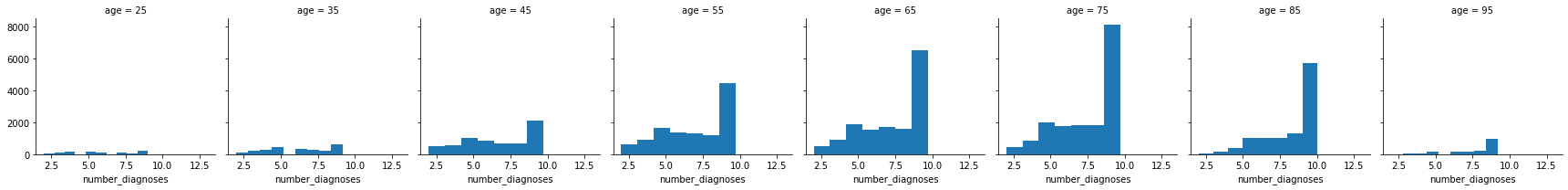
Chart, bar chart

Description automatically generated**Chart

Description automatically generatedHypothesis 1**: Age has a higher impact on readmission

* Grouped bar chart displaying the number of patients readmitted and non-readmitted for each age group.
* The result show patients in the 75-age group have been readmitted the most, closely followed by 65, 85 and then 55.
* This shows patients that are older, those in the 55+ group, would be more likely to be readmitted
* Bar chat showcasing the number of patients for each age group.
* Results obtained from this graph shows most of the patients age group are in the 75 range.
* With the next highest being 65 and 55.
* Bar chat showcasing the number of patients for each age group.
* Results obtained from this graph shows most of the patients age group are in the 75.
* With the next highest being 65 and 55.
* A green and white rectangle

  Description automatically generated with low confidenceA picture containing text, clipart, vector graphics

  Description automatically generatedHowever, 95 does not have a high readmitted number this could be due to not having many 95 aged patients in the dataset. But another reason why the number of patients drops after 75 is because the global life expectancy is 73.2 years [1], which would explain why the number of patients starts to drop at this point instead of the increase seen in the other ages.
* The bar charts above display the percentage of the age group readmitted and not readmitted. What these charts show is the percentage of age group being readmitted is higher in the 55+ ages and that even though there are more of these ages in the dataset, the percentage of not readmitted is lower. These result showcase that the older the patient is the more likely they will be readmitted.
* The insulin level in older ages for down and up are higher than younger ages. This suggests insulin levels are at more risk at changing in older ages which can lead to more serious health problems hence readmitted increasing.
* Using the correlation heatmap I can see that readmitted has a 0.1 correlation with number of diagnoses, showing it has a more positive relationship with it. When checking the correlation for number of diagnoses and age, it has a 0.21 correlation number, showcasing a high correlation.
* The graph above shows how as the age increases the number of diagnoses increase which also effects the number of patients being readmitted.

**Conclusion**

Findings show the older the patient is the chance they will be readmitted is higher, hence concluding **Hypothesis 1**: Age has a higher impact on readmission, to be True.

Chart, box and whisker chart

Description automatically generatedChart, pie chart

Description automatically generatedChart

Description automatically generatedA picture containing chart

Description automatically generated**Hypothesis 2**: African Americans are more likely to be re-admitted than other ethnic groups

* Since the previous hypothesis based on my findings was true, decided to check range of age for each race using a box plot.
* The box plot clearly shows that the ages of the patient that are African American is the same as Caucasian. The median is at 65 for both showing the age range for the races are similar.
* Bar chart for number of patients for each race.
* Chart shows the dataset consists of mainly Caucasian, not many African Americans in the dataset
* Group bar chart showcasing the count of readmission and non-readmitted for each race.
* As expected, since the count of Caucasian is much higher than any other race the number for readmission mainly consist of Caucasian people.
* The pie chart clearly shows this, with total readmission being 78.3% Caucasian.

Icon

Description automatically generated with low confidenceChart, bar chart

Description automatically generated**Conclusion**

* Checking the percentage of each race individually for readmission shows that more than 40 percent of Caucasian patients are being readmitted, which reflects in non-admission with it having the lowest percentage.
* African Americans are not too far behind Caucasian with it slightly beneath 40% in readmission. This shows that African Americans are not the most likely race to be readmitted.

Findings show that the dataset is too imbalanced in terms of race. The number of Caucasian is far higher than any other race in the dataset hence to conclude if African Americans are more likely to be re-admitted based on this data is unreliable. More data is needed to prove this hypothesis. Hence concluding, **Hypothesis 2**: African Americans are more likely to be re-admitted than other ethnic groups, to be False.

Chart, bar chart

Description automatically generatedA picture containing bar chart

Description automatically generated**Hypothesis 3**: Women patients are more likely to be re-admitted than men

* Females are being readmitted more than males as show in the grouped bar chart. This is also the case for non-readmitted.
* The number of patients per gender. Chart shows dataset consist mostly of females with the number being close to 35000, although males are less the number is not far behind standing at about 30000.

Bar chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, histogram

Description automatically generated

* The percentage of females’ readmission is higher than male’s readmission percentage, although it is only higher by a few percentages.
* Shows that females are being readmitted more than males.
* A box plot for gender against time in hospital shows how males has a few outliers different to females with just 1.
* The median and range is higher for females than male, meaning females are spending more time in the hospital.
* Since this has a positive correlation with number of diagnoses, means that as time increases diagnoses increase, which can lead to readmission increase for females.
* As stated earlier, about the correlation between readmitted and number of diagnose, decided to look at number of diagnoses. Females has a higher count but both genders are showcasing a similar shape, with both having a high number of diagnoses at 8 to 10.
* Looking at the correlation for number of diagnoses, it has a high positive correlation with time in hospital, which could be another reason females’ diagnoses is higher.

**Conclusion**

The data is fair, the difference in number of patients for each gender is small. The data shows that females are being readmitted more than males, and then looking at the correlation of variables and different graphs it shows the difference in time in hospital higher for female could be affecting other variables, hence increasing the readmission. Hence concluding, **Hypothesis 3**: Women patients are more likely to be re-admitted than men to be True.

Chart

Description automatically generatedChart

Description automatically generatedChart

Description automatically generated**Hypothesis 4**: Diagnose types have a higher impact on re-admission rates

* Grouped bar chart of readmission for each ICD code in each of the 3 diagnose columns.
* For this graph I have separated the 250 codes since they directly link to diabetes, to show if being diagnose with diabetes has a greater impact on readmitted or not.
* The main discovery from plotting it this way is that the 200 ICD codes have a greater number of patients than any other code, especially the 250 ICD codes.
* The readmission rate for other ICD codes is not as high suggesting if the diagnose lies in the 200s than the patient is more likely to be readmitted.

**Conclusion**

The data shows that the type of diagnose has an impact on readmission as clearly shown in the above graphs. Hence concluding, **Hypothesis 4**: Diagnose types have a higher impact on re-admission rates.

**Initial Model Building**

The initial model was built using logistic regression model. I used this model because it is good in predicting when the outcome variable is categorical. In this model the outcome is readmitted, with 1 representing will be readmitted and 0 not readmitted.

The initial model score was 60%. This means the current model is poor.

Then I built up the model using train and test data set, allocating 20% of the data to the test. The result of the model after this was also 60%, therefore no improvement in the score was found resulting in a poor model.

**Improved Model**

Improving the logistic regression model:

* Considering the whole dataset, to do that the categorical need to be converted to numerical.
* Normalise the dataset for better accuracy.
* Feature selection, to reduce the number of input variable to help improve the performance of the model.

Building the new improved model with training and testing datasets. 30% allocated to testing.

The model accuracy score saw an increase of 4%, meaning from 60% => 64%. This is increase means the model is more accurate, but the model is still poor.

The cross-validation score is 0.638 => 64%, meaning the effectiveness of the model has increased from the previous model although not a huge increase.

A picture containing treemap chart

Description automatically generatedThis is shown in the confusion matrix. The confusion matrix is a lot more viable unlike before where the values would be 0.

Logo

Description automatically generated with medium confidence

The AUC, (area under the curve) also shows the accuracy score, and from this the accuracy score is 0.66 which is higher than the accuracy score from before by 2%.

**Other attempts at improved model building**

Logistic Model building with over sampling and removed columns with p-values above 0.05.

This gave a model accuracy score of 62%, which can increase but not the biggest.

Neural Network model. This model was done using the neural network model but gave a model accuracy of 0.38.

The highest I achieved in accuracy score was 64%, which is still a poor model to be viable for a business the model needs to be at least 75% above. Thus this model is not business appropriate.

A picture containing background pattern

Description automatically generated**Cluster**

Using the elbow method, the optimal number of clusters was 6. Each cluster represents a separate group of patients that have similar traits. These traits are based on all the data in the dataset, comparing the diagnoses, medication, readmission, gender etc. to form these six clusters.

This information on clustering can be used to make decisions based on:

* The patients age with diabetes will most likely be in their late 50s and above, which all the clusters show.
* A high number of the patients will be needing diabetic medicine specifically cluster 4 and 5 patients will have the highest number, ensure enough medicine is in stock.
* Higher number of patients will be taking diabetic medicine, clusters 4 and 5 will be needing the most
* The readmission of patients in cluster 2 will be lower than the rest of the clusters, meaning cluster 2 will likely have less diabetics and serious health problems.
* The number of medications needed will be the same for each cluster expect for cluster 3.
* The number of lab procedures needed is less in cluster 1 and 5.
* Patients in cluster 5 will are less likely to spend time in the hospital compared to other clusters.

Bibliography

1. “Life expectancy of the World Population,” *Worldometer*. [Online]. Available: https://www.worldometers.info/demographics/life-expectancy/. [Accessed: 24-Mar 2022].