Big Data & Predictive Analytics

Coursework 2

CO3093

Author: Mohamed Mubaraak
Department of Informatics
Date of Submission: March 2018

***Objective**: Given a dataset outlining information related to diabetic patients, we would like to build up a predictive model which will help to predict whether a patient will be readmitted for their condition at a later date. This will be done with the assistance of a K-Means Approach.*

Exploring the data

The data-set shows information relating to diabetic patients that were admitted to US Hospitals between 1999 and 2008. The dataset shows personal information for each patient such as their race, gender and age and also medical information such as the types of drugs they are on, the type of admission they have and the length of time they spend at the hospital. It is therefore important to outline features that will have a direct impact on determining the objective.

Distribution of Race , Gender and Age in the Dataset

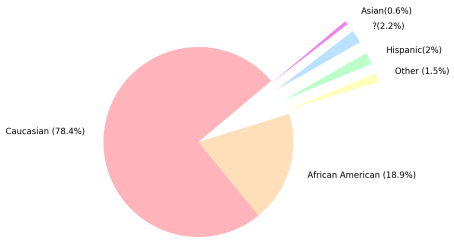


Figure 1: Race Pie Chart

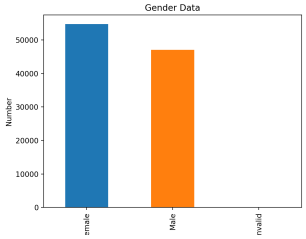


Figure 2: Gender Chart

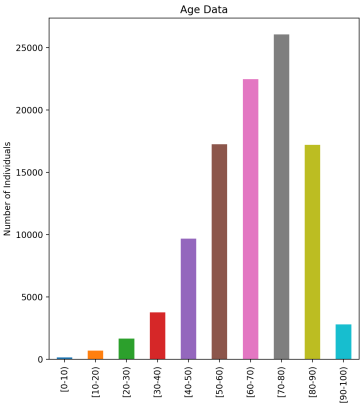


Figure 3: Age Bar Chart

The pie chart above visualises the distribution in the races of patients in the dataset. It is clearly evident that the majority of patients that are admitted are of Caucasian descent ***(78.4%)*** whilst Asians make up the fewest number of patients in the dataset ***(0.6%)***. The dataset also shows that the majority of patients in the diabetic dataset are Female ***(54708)*** compared to Males ***(47055)***. The dataset also consisted of a small fraction of ***(3)*** individuals that chose not to disclose their gender. Another point of interest in this dataset is the distribution of the age of patients since the data shows that the majority of patients at US hospitals during 1999 and 2008 fell between the age of 50 and 90 years of age. A possible cause for this somewhat bell curve distribution centred around the range [70-80) may be due to individuals at a younger age leading a more active lifestyle and are provided a healthier diet than the other individuals that are older.

What do the findings tell us?

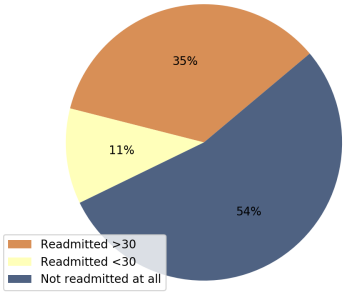


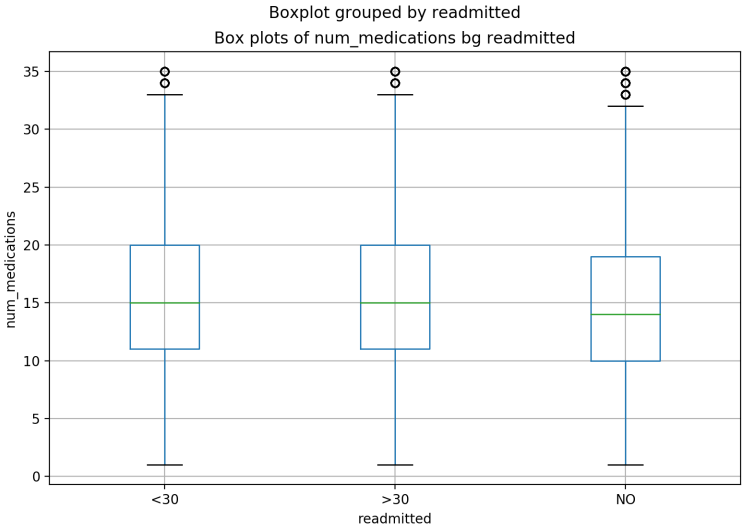
Figure 4: Readmission Information

Readmitted (if at all)	Number of individuals
Patient was readmitted	46902
Not readmitted	54864

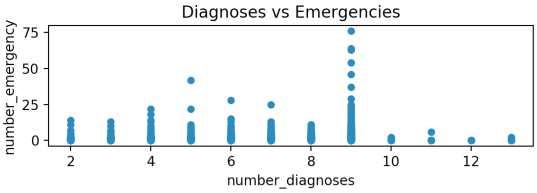
This pie chart yields some interesting information concerning the dataset. According to the readmission information, the majority ***(53.9%)*** of patients at the US hospitals in the decade spanning between 1999 and 2008 were not readmitted to the hospital after their initial admission, compared to ***(46.1%)*** of patients that were readmitted at some point. This is further broken down to '<30' and '>30' within the percentage of patients that were readmitted. This perhaps suggests that the doctors and medication prescribed for the patients in that decade were good however at this point in time it is safe to state that a possible correlation does not imply causation, .i.e. that just because fewer patients in this dataset were readmitted in that decade, it doesn't mean that those that were not readmitted will not be readmitted in the future.

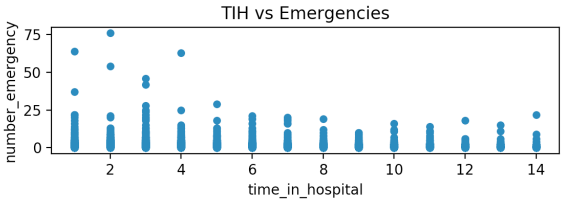
Relationships in the dataset

Given a dataset like this, it would be realistic to observe relationships between the columns for each individual. Relationships that would perhaps provide us with a reasoning for the readmittance of a patient either in the past or in the future. An example of this could be, that with an increase in the number of medications being prescribed for a patient, those patients being prescribed more medications are likely to be readmitted since they are higher priority or have a slightly more severe case of diabetes. This can be shown in the data below as, regardless of the duration of re-admittance ('>30'/'<30') patients with higher number of medications had been readmitted too.



Another interesting point to explore is the likelihood of there being a relationship between the number of diagnoses for a patient compared to the number of emergencies recorded for that patient. As is obvious, a patient who has been diagnosed for more illnesses, is more likely to come to the hospital as an emergency for such illnesses or as a by-product of those illnesses. The data shows a steadying growth in the pattern of increased diagnoses resulting in increased emergencies however it peaks at 9 diagnoses resulting in 75 emergencies for a patient. This is perhaps data considering outliers as this value is almost triple the number of emergencies for 8 diagnoses. Furthermore,

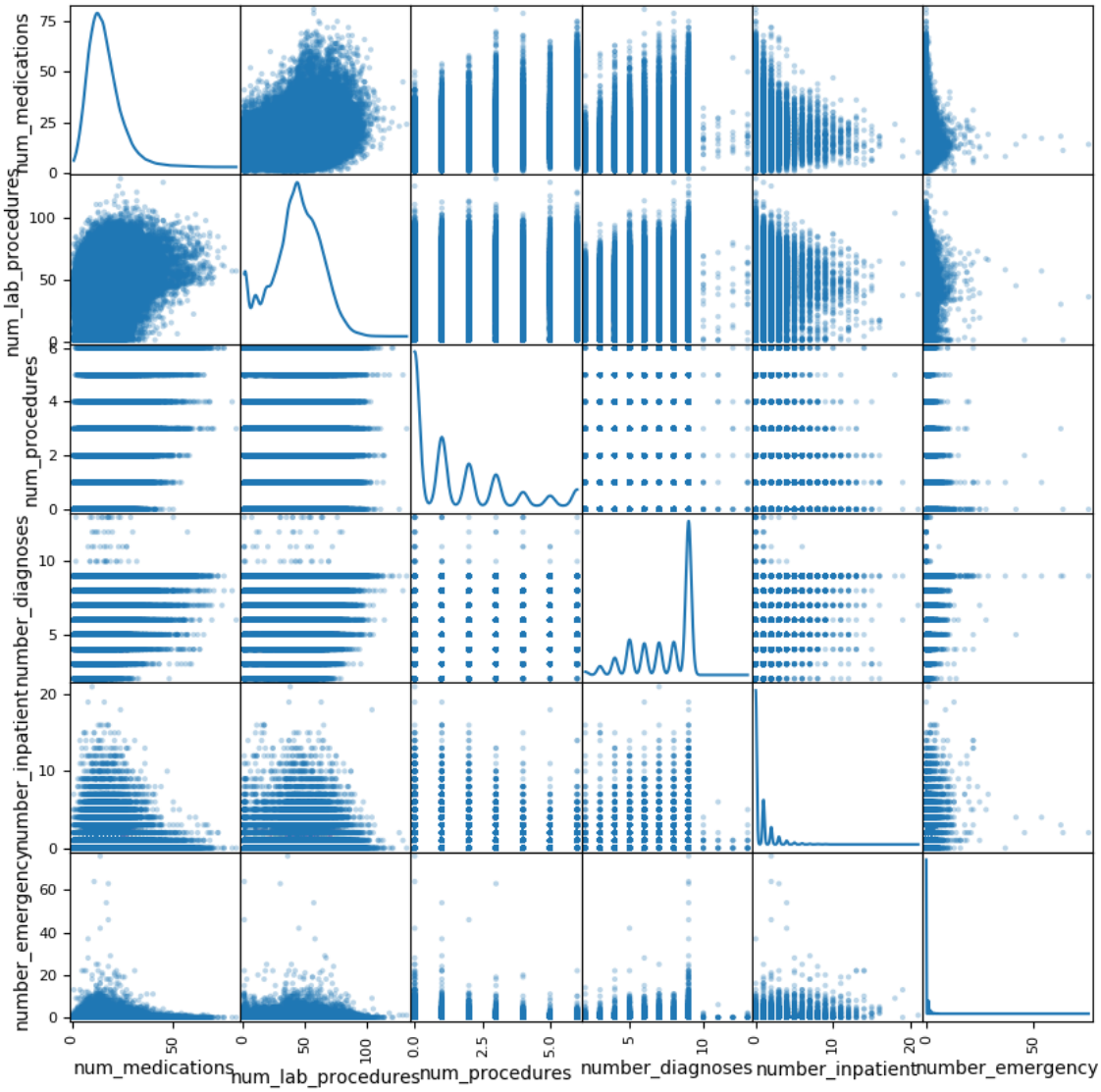




this information is contrasting with the information we gather for Time in Hospital for a patient compared to the number of emergencies recorded for that patient. Ideally, a patient who has a greater number of emergencies recorded in their records would be seen as a more high profile

patient and so therefore spend more time in hospital. However, the data suggests that as more time is spent in hospital, the number of emergencies remains the hovering around 25 emergencies. With some patients exhibiting outlier behaviour, e.g. a patient with nearly 75 recorded emergencies spending only 1,2 or 3 days in hospital.

Scatter Matrix for this Dataset



Outlier Handling

For this assignment, a value is considered an outlier if any data point falls outside of either 1.5 * interquartile range (IQR) below the first quartile OR 1.5 * IQR above the third quartile. This is easy to visualise in Figure 5 where there are numerous points above the third quartile in the boxplot that does not have the outliers handled.The function **identify_outlier** is only applied to numerical columns and its functionality can be see when it is applied to **Number of Medications** below. Once the function is applied to the relevant numerical columns in the dataset that are to be considered as input features, the outcome is to Figure 6 where the outliers exist no more.Upon filtering the data and removing outliers, we are left with 90383 rows and 15 columns.

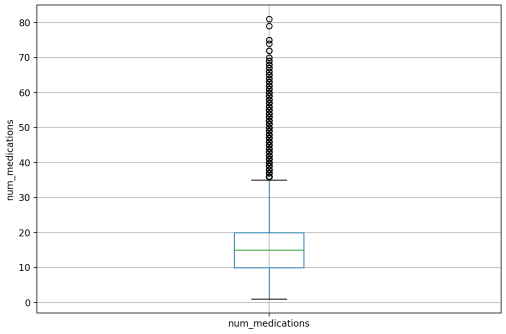


Figure 5: No. of Medications w/ Outliers

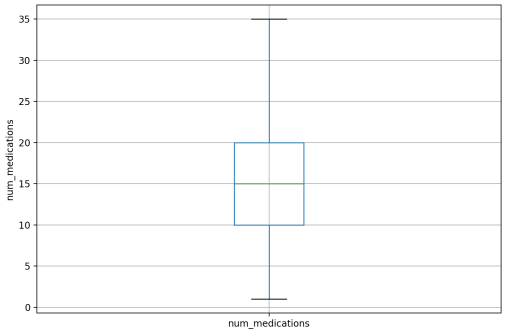


Figure 6: No. of Medications w/o Outliers

Logistic Regression Model

In order to make the most accurate predictions using a logistic regression model, it is vital to a) select the right features ***(X)*** that will yield a direct correlation to an output ***(Y)***. In the case of this model we are using 15 input features and trying to see how they impact which resulted in the following results:

```
Mean hits: 0.6133136640236032
Accuracy score: 0.6133136640236032
Cross validation mean scores: 0.6118853032679733
Intercept:
 [-0.69223983]
Coefficients:
      num_medications: -0.04924904752120697
   num_lab_procedures: 0.0862788437952141
       num_procedures: 0.407173535360584
     number_diagnoses: 0.0
     number_inpatient: -0.5015972768530171
     number_emergency: -0.19064255689950343
```

```
     Score against training data: 0.6135803249668078
Score against test data: 0.6116171860593768
```

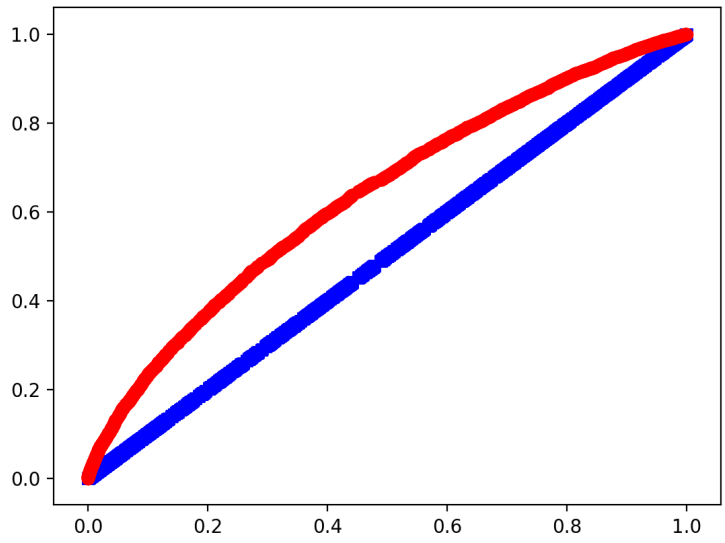
The mathematical equation for the fitted model is:

$$-0.69 + -0.05 * num_medications + -0.09 * num_lab_procedures + 0.41 * num_procedures$$
$$+0.0 * number_diagnoses + -0.50 * number_inpatient + -0.20 * number_emergency$$

These values are interesting because they give an indication on how effective the model is in its predictions and it also allows us to evaluate the performance of the model. The main conclusion to gather from this output is that this model has a predictive accuracy score of 0.61 or 61% which may seem perfectly fine if you wanted to predict something simple like the colour of a ball in a bag, however for something as serious as dealing with medical data relating to diabetic patients, this could prove dangerous with such a large margin of inaccuracy or error. The cross validation mean scores which is the score acquired from averaging measurements of predictive accuracy (usually used to provide a more realistic, true representation of the predictive model) also outputs a value of 0.61 which, in the case of this experiment with this sample test data, is simply just not good enough.

Tweaking the data & the ROC Curve

Initially the value for model accuracy was really low (0.52) but modifying the selected features considered in the predictive model by either reducing or adding to them resulted in increasing the value to 0.61 using 20 input features which is the value that resulted in the best results for this dataset. The ROC curve is shown below along with the value of the area under the curve:




```
AUC = 0.6365058371722709
```

Cluster Information

```
J-score =  6052.964678764797
 score =  -6052.964678764797
[0 5 4 ... 5 2 3]
centroids [[0.10790715 0.20535748 0.05700778 0.13023511 0.35235726 0.        ]
 [0.6454771  0.48489485 0.72931016 0.34849718 0.48296807 0.        ]
 [0.60659391 0.42226974 0.13692468 0.23136405 0.4642133  0.        ]
 [0.1919104  0.31503642 0.40911859 0.19205168 0.41890708 0.        ]
 [0.15426445 0.30841562 0.84680882 0.20429307 0.42097882 0.        ]
 [0.19691432 0.43742808 0.04256992 0.16724746 0.45572404 0.        ]]
```

Figure 7: Histogram of Clusters

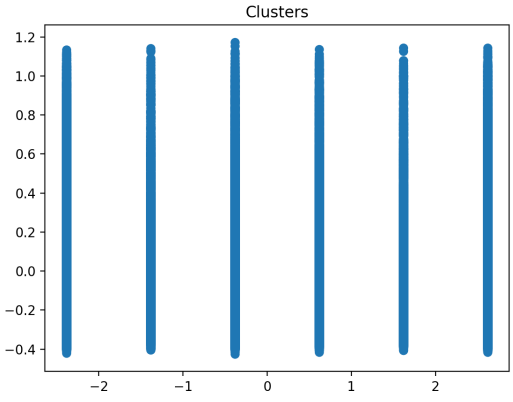


Figure 8: Clusters Scatterplot