**Introduction**

Hospital readmission is considered an effective measurement of care provided within healthcare. Being able to identify patients facing a high likelihood of unplanned hospital readmission in the next 30-days could allow for further investigation and possibly prevent the readmission. This can help healthcare providers improve inpatient diabetic care. The dataset we used represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes 50 features representing patient and hospital outcomes.

The data set provides the ability to predict the impact on re-admission rates for patients hospitalized with diabetes. We will first preprocess the data, select attributes and then used different algorithms to predict if a patient will be readmitted within 30 days based on attributes selected. Comparisons on how well the readmission is predicted will be made between algorithms using various metrics.

**Data Preprocessing**

The diabetic dataset consists of 101766 tuples and 50 attributes including the readmission result.

First of all, we have dealt with all the missing data and the redundant attributes. We have dropped attributes with a large percentage of unknown values because they can not provide us with enough information, including weight, payer id, and medical specialty. Then we dropped the rows with missing information to get a cleaner data set. Encounter id and patient number were dropped because they are not relevant to the readmission. Attributes that only contain one value ('No') were dropped, which were citoglipton and examide since they could not affect our prediction. We now have 43 attributes (including readmission) left.

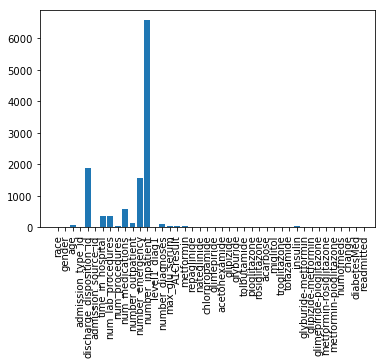
Then we converted categorical attributes to numeric in order to further process. Categorical attributes with only two categories were converted to binary data, for example, gender and some medicines. Categorical attributes with more than two categories such as race were coded with positive integers. Since age was given in intervals, we took their average to represent each bin. We have encoded the first diagnosis with integers by 9 larger categories using the ICD codes rather than using more than 900 detailer categories to avoid the complication. The second and third diagnosis have been dropped.

There are 21 attributes representing 21 types of medicines, our main idea was to see if a patient is using a medicine (including steady, up and down) or not (No). Therefore one more attribute was created to calculate the total number of listed medicines a patient uses, in order to capture more information. So we now 42 attributes (including readmission) in total.

The result we are trying to predict is if a patient is readmitted within 30 days after release, so we have coded ‘Not readmitted’ and ‘readmitted after 30 days’ with 0 and ‘readmitted within 30 days’ with 1.

**Feature Selection**

Since there are still over 40 attributes in our dataset, we want to boil it down so that we are only using the most important attributes for the prediction. A chi-squared test was performed for each attribute and the features with the top 15 scoring were selected. The following barplot shows the scoring of the attributes.



**Algorithm Implementation and Results**

After shuffling our data, we have used 80% data for training and 20% data for testing. We have implemented logistic regression, Decision tree, Naive Bayes, Neural network and Random forest to see which algorithm performs better with our dataset. And we have gotten the following results.

**Logistic regression**  
Cross Validation Score: 77.59%  
The confusion matrix is

|  |  |
| --- | --- |
| 7039 | 3508 |
| 918 | 8145 |

Accuracy is 0.77 Precision is 0.70 Recall is 0.90  
   
**Decision tree**  
Cross Validation score: 73.48%  
The confusion matrix is

|  |  |
| --- | --- |
| 7627 | 2920 |
| 2251 | 6812 |

Accuracy is 0.74

Precision is 0.70

Recall is 0.75  
   
**Naive Bayesian**

Cross Validation score: 72.05%  
The confusion matrix is

|  |  |
| --- | --- |
| 8367 | 2180 |
| 3355 | 5708 |

Accuracy is 0.72

Precision is 0.72

Recall is 0.63  
   
**Neural network**  
Cross Validation score: 77.69%  
The confusion matrix is

|  |  |
| --- | --- |
| 7171 | 3376 |
| 1408 | 8015 |

Accuracy is 0.78

Precision is 0.70

Recall is 0.89  
   
**Random forest**  
Cross Validation score: 75.07%

The confusion matrix is

|  |  |
| --- | --- |
| 8238 | 2309 |
| 2678 | 6385 |

Accuracy is 0.75

Precision is 0.74

Recall is 0.71

We can see that the overall accuracy is around 75% with Logistic regression and Neural network being the most accurate. Both precision and recall are around 70%, which are also acceptable, with Random forest having the highest precision at 0.74.

**Comments**

We can see that the performance with different algorithms across the board remains quite consistent. Even we have located that Logistic regression and Neural network are the most suitable models for our data, there is not a big differentiation compared to other models we used. We are basically satisfied with the performance on predictions we have made with our data. However, there are aspects we can maybe improve on. Due to the lack of medical background, we could not comprehend the meaning behind our attributes very well. We could not select attributes based on the medical relevance that they have with readmission, so we have relied on the chi-square solely for feature selection. Another thing to note is that the three diagnoses of each patient are considered fairly important by us. However, there are around 900 categories within each diagnosis, causing us not able to deal with those attributes in details. Although we have classified the first diagnosis into 9 larger categories, this process was very time-consuming. Therefore we could not process the second and third diagnosis in the same manner and had to drop them.