BRIEF ANALYSIS ON THE PREDICTION OF OBSERVATION UNIT

Executive Summary

Even though most of the hospitals in the U.S. do not contain an observational unit, the Montanaro management believed that having a dedicated observational unit would decrease the overcrowding in emergency units and increase the hospital capacity. Following the evaluation, we tried to implement a data driven approach in order to improve the scenario in the observational units and thus to decrease the patients which signifies the added number of beds in the OU. Also, the approach was to predict the number of patients that would "flip". The historical data on observation-patients from Montanaro's Information Systems Department (ISD) were taken mostly for the analysis.

The appraisal showed that through various methods of data analysis we got the probability of each patient that would eventually flip to other wards. From the outcomes, the "flipped" were mainly modelled by the patients who are of gender male, also patients with diagnosis codes 580 and with lower blood pressure rates. Facilities that contributed in the exploration of the data included chief of emergency medicine, chief of medicine, the vice president and COO, clinical program administrator, director quality, chief nursing officer, director of social work and their respective sub divisions.

As the next step, we recommend to the management to initiate the implementation of an more accurate OU exclusion list which is more focused on including data that contain more of male gender, patients with diagnosis codes 580 and those with lower blood pressure rate. Moreover, with the implementation of this step we believe to increase the inpatient capacity and in an overall to increase the number of patients treated. But this would not be the end of the process, by repeated analysis of the effect that this suggestion brings we can achieve a more effective and functioning Observation unit.

Problem Statement

The recent observations in the units of Montanaro were taken to evaluate the average stay of patients in the observational units (OU) and the number of patients who flipped from the observational units to inpatient units. The average stay in the observational unit was 24 hours. Approximately 55 % of the patients were medicine service patients and the other 45% were patients that changed to inpatients.

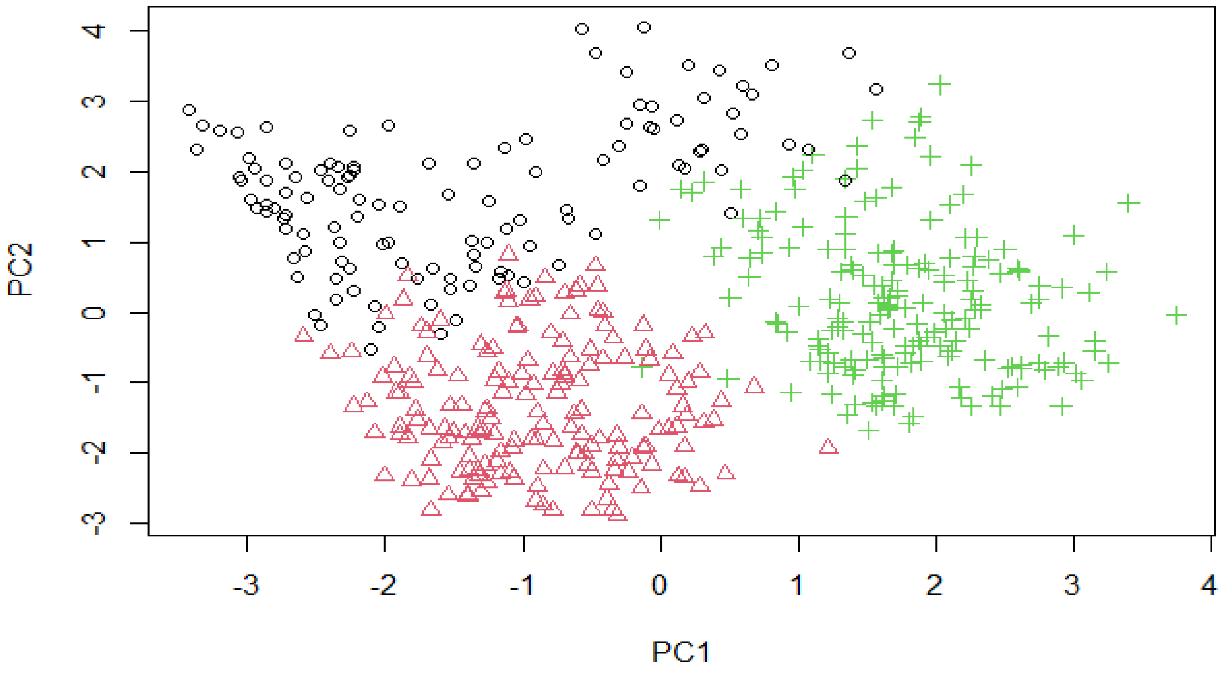
The objective of the analysis is to cluster the OU patients into three groups and to create models in order to predict whether patients will “flip” from the observation unit (OU) to an inpatient unit or whether they are discharged from the OU. Subsequently, applying the model to the prediction data. The analysis is also conducted in order to examine the OU exclusion list and to find the factors that have an effect on the prediction model.

Data Analysis

Our data set consists of 556 rows and 15 columns. In order to get better results we have done pre-processing. As part of this, first we took the summary of the data set to know about the NA values and the outliers. We plotted the box plots to detect the outliers and the procedure was done for six variables. They are OU_LOS_hrs, BloodPressureDiff, BloodPressureUpper, BloodPressureLower, Pulse and Respirations. Outliers were removed using the filter function from the dplyr package. Hence, the box plots were once again plotted to check whether the outliers were removed or not.

After the removal of outliers, the NA values were detected in the variable temperature. So, we imputed the NA values in the column of temperature using the median function. Afterwards, we removed the column ‘ObservationRecordKey’ which was not useful for the analysis. Additionally, we also removed the 'InitPatientClassAndFirstPostOUClass' because it is the same as the column 'Flipped' and decided to keep the variable 'Flipped'. These are the pre-processing that we did before the analysis. After the pre-processing, we were left with 480 rows and 13 columns.

The first objective is to cluster the patients into three groups using the data set. So, we used k means for clustering. Before clustering, we created the dummy variables for the categorical variables. We used the function 'dummy.data.frame' to create dummy variables. There are four categorical variables. They are Gender, Flipped, PrimaryInsuranceCategory and DRG01. Furthermore, we did the scaling using the function scale and clustering was done using the kmeans function. Then, the PCA was plotted and we obtained three clusters as shown below.



We plotted the clusters on the first two principal component scores i.e., PC1 and PC2 to visually evaluate the quality of the clusters. In the plot there are three colors which represent three clusters. The colors are green, red and black. The red and the green clusters are mostly well separated. The black cluster is somewhat spread across. The red and the green clusters are clearly distinguished from each other. Many of the samples in the black cluster can be easily partitioned, but there is some overlap. Also, from the plot we get to know the variables Flipped0 and Flipped1 affects PC1 the most whereas, the variables PrimaryInsuranceCategoryMEDICARE OTHER and Age affects PC2 the most. In the positive part of the PC1 we have more flipped patients and less non flipped patients whereas in the negative part of PC1 we have more non flipped patients and less flipped patients. Moreover, in the negative part of PC2 we have more aged people and also more people with the insurance category of medicare other whereas in the positive part we have the people with less age and also less people with the insurance category of medicare other.

The red clusters are clustered towards the left bottom part. It has negative values in both PC1 and PC2. Here, based on PC1 the non-flipped patients are more and the flipped patients are less, whereas based on PC2 Primary Insurance Category MEDICARE OTHER and the age is more.

The green clusters are clustered towards the right bottom part. It is in the positive part of PC1 and negative part of PC2. So, according to PC1 it has more people flipped and less non flipped people on the other hand PC2 has more aged people and also more people with the insurance category of medicare other.

The black clusters are clustered towards the top left part. It has negative values in PC1 and positive values in PC2. Here, based on PC1 the non-flipped patients are more and the flipped patients are less, whereas based on PC2 PrimaryInsuranceCategoryMEDICARE OTHER and the age is less.

The second objective is to predict whether the patient is flipped or not. For that we took the data set that we had done the pre-processing and also, we removed the column OU_LOS_hrs, because the use of that column may create leakage in the overall prediction. The data was divided into train and test data sets, by taking 30 percent of the overall data in train. The train data had 144 rows and 12 columns whereas the test data had 336 rows and 12 columns.

We got to know there were no outliers and NA values in the partitioned data. For classification models we first used the logistic regression model. We used the glm function to do the logistic regression model. From the summary of the model, we were able to find that the variables which were significant in identifying whether our patients will flip or not were patients of male gender, also who are with diagnosis code 580 and with lower Blood Pressure rate. In order, to find the accuracy of the predicted model we then found the confusion matrix and the mis-classification rate, by predicting the result for the holdout data, the misclassification rate which we obtained was 40%.

The obtained confusion matrix is:

	0	1
0	113	69
1	66	88

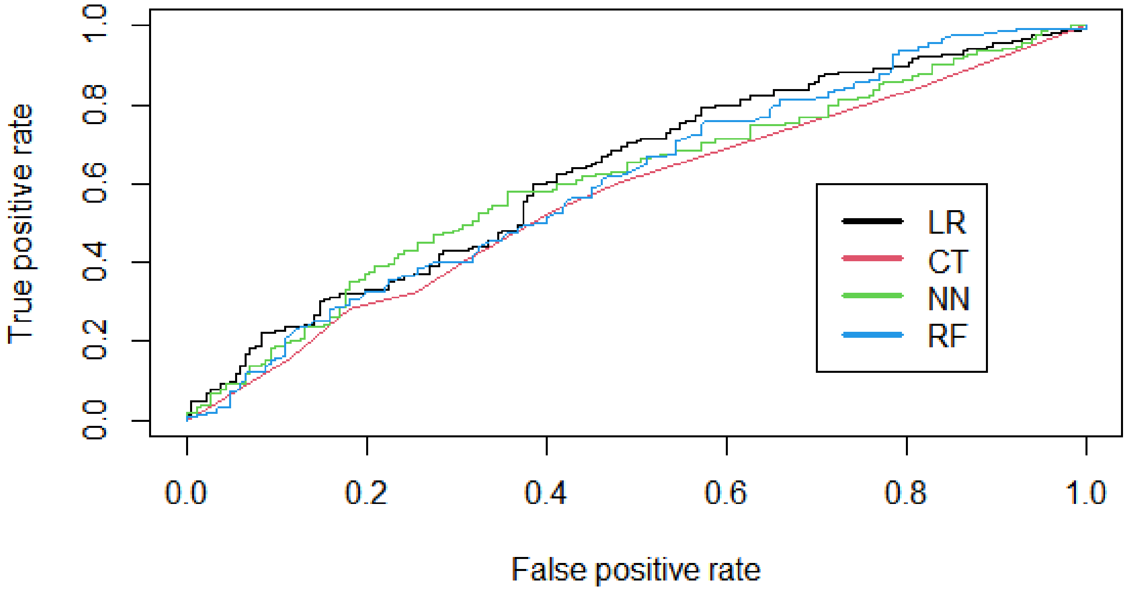
Next, we used the classification tree for the prediction. We used the rpart function for the classification tree. For creating the prediction model, we used the same train and test data as that of logistic regression. With the use of this resulting model's application on test data we obtained a misclassification rate of 44%. From the variable importance we were able to find that the variables which were significant in identifying whether our patients will flip or not were the diagnosis code, the patients by their gender and by their lower blood pressure.

Thirdly, we tried to classify our data with a neural network model, for this we created dummy variables for test and train data sets and then we scaled them to apply the train function. We also changed the level ordering for the application of the train function, with the use of the resulting model on the test data we obtained a misclassification rate of 41%. From the variable importance we were able to find that temperature, pulse rate and primary insurance medicare of the patients affect the flipping of the patients.

Finally, we did a Random Forest model, for this model we used the same train and test data set. However, we re-levelled the label ordering for the application of Random Forest function to run the model on the train dataset. Eventually, we used this model on test data and we obtained a mis-classification rate of 43%.

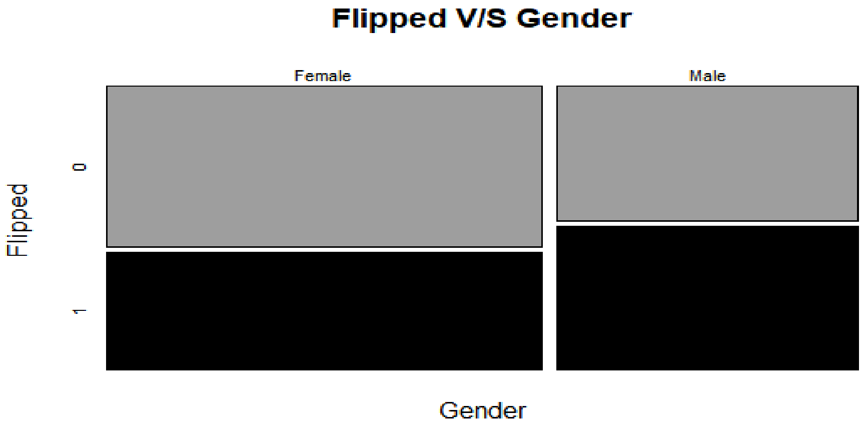
In order to find the most appropriate model for our prediction of the OU exclusion list, we plotted a ROC curve for all the four models we created earlier. Along with the help of the misclassification rate and this ROC curve we came to the conclusion of using Logistic regression for the further predictions.

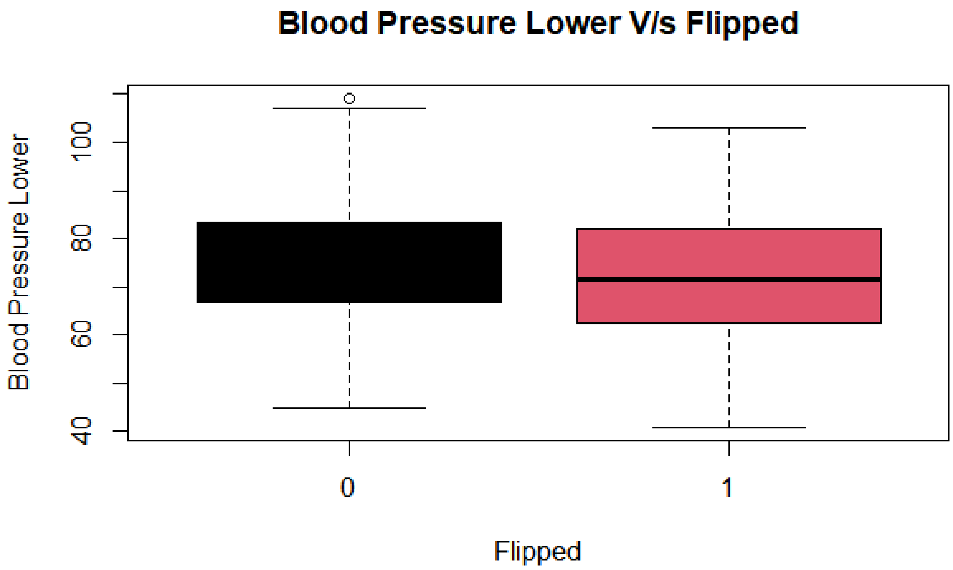
ROCcurve:

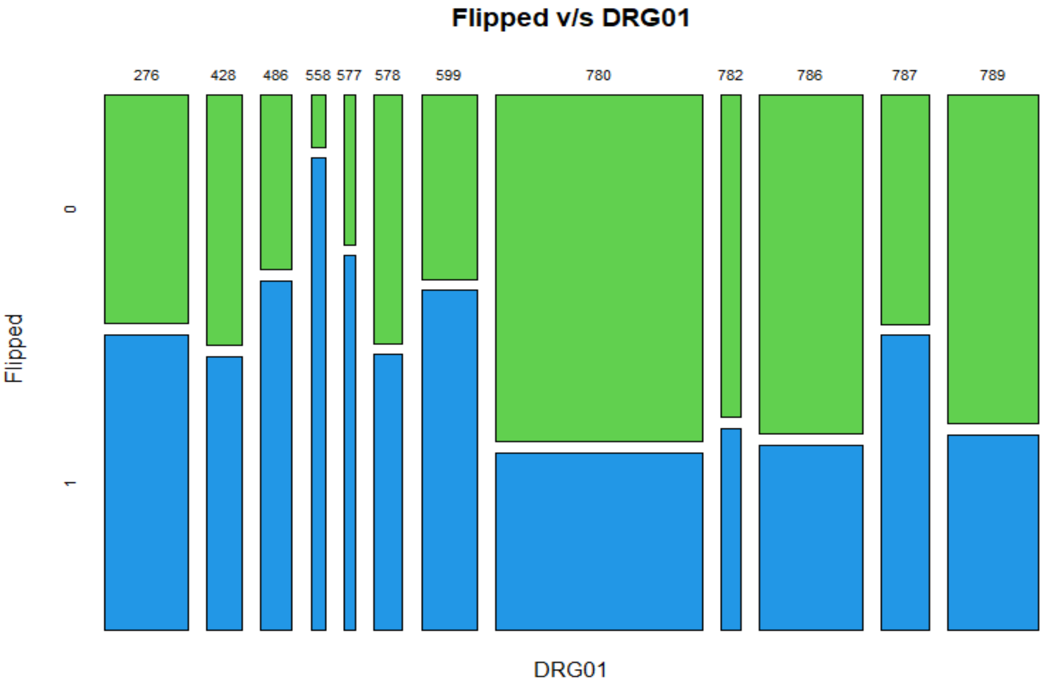


We also Visualized the important variable which we concluded from the logistic regression, to determine whether they resulted in the flipping of patients to inpatient status and validated the effect of them.

The plots that we obtained are given below:







At last, we then applied the model we generated from logistic regression on the prediction data, which was median imputed for the missing values, to obtain the probability for patients in the observation unit to find out whether they will flip or not, the predictions are attached along with this in a csv file.

Conclusion

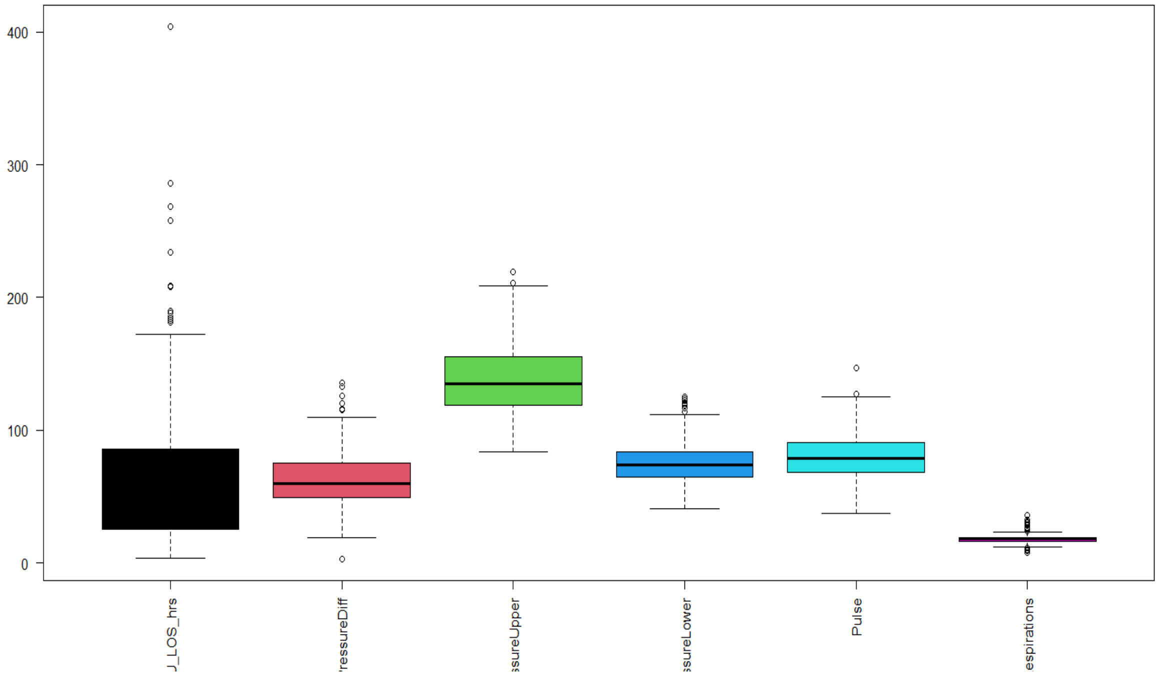
From the overall analysis, for predicting the OU exclusion list, as a hospital, we have the duty to create an environment which helps the patients to have a fast recovery. For this matter Montanaro hospital management has decided to have special units for those patients that are to be in observation after monitoring their present medical conditions. The observational unit that is specially constructed to have better inspection on the patients help to determine the health status of the patients and also to determine their health in future.

After the evaluation and analysis of the models we got a predicted probability of each patient that might flip from the observational unit to inpatient units. Different kinds of predicting methods were executed in order to have an enhanced model to foresee the number of flipped cases.

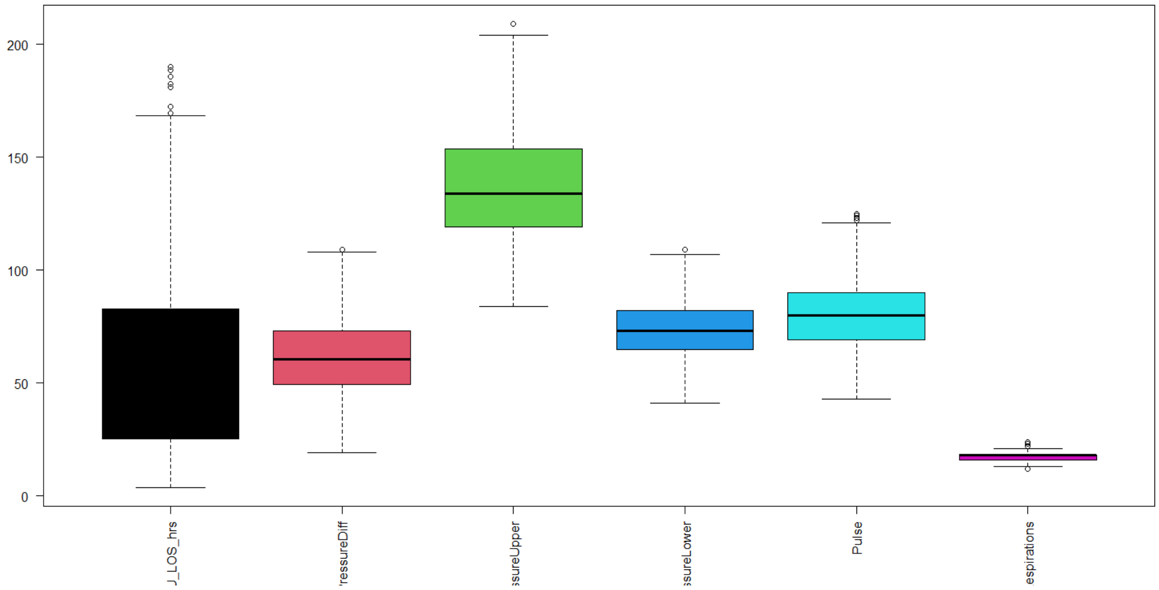
Furthermore, this analysis also exposes us to the variables that affect the model and they are patients of gender male, with diagnosis code of 580 and those with lower blood pressure rate. Hence, with the implementation of an OU exclusion list with this model consideration in mind we can obtain proper use of bed spaces in the hospital and hence increase the overall patient intake.

Appendix

Box plot before the outlier removal:

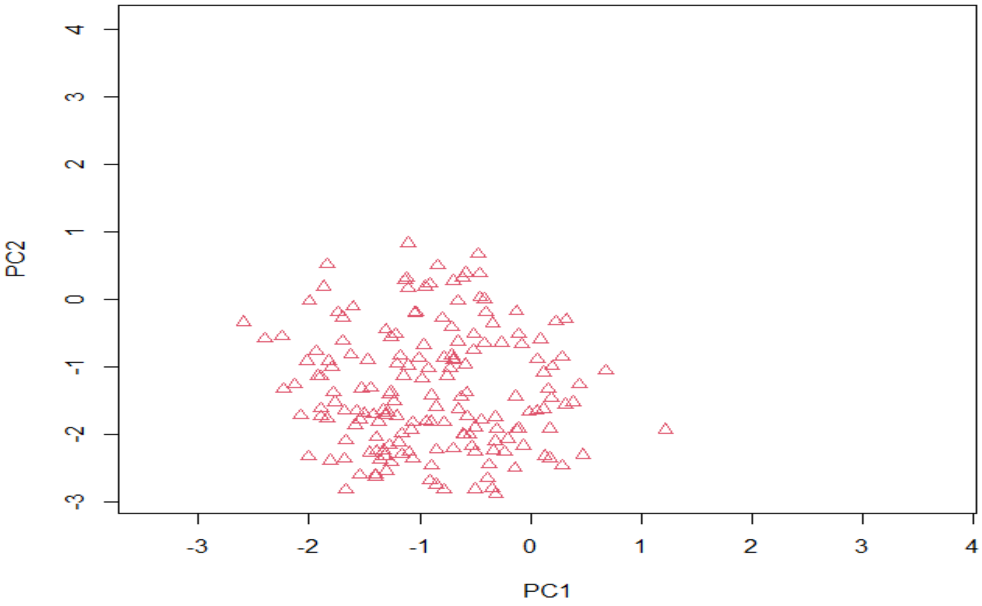


Box plot after the outlier removal:

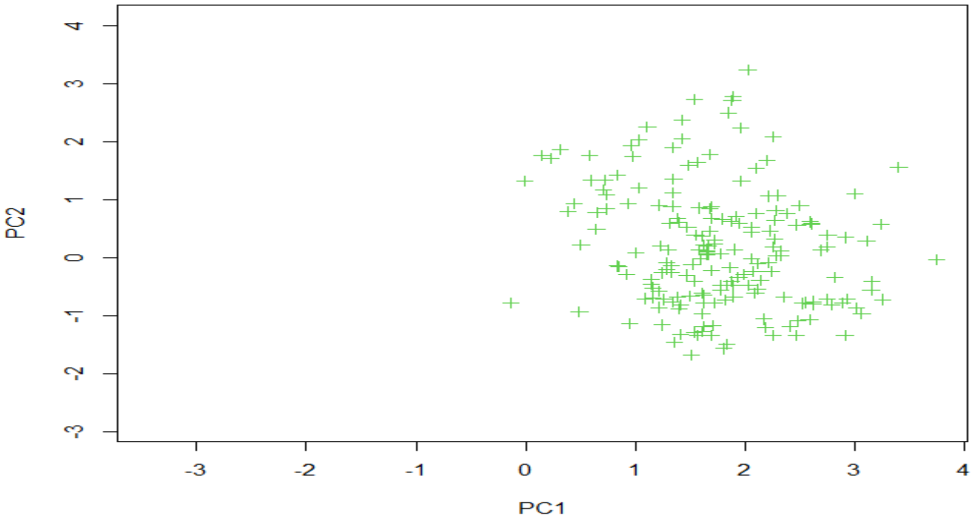


Clustering:

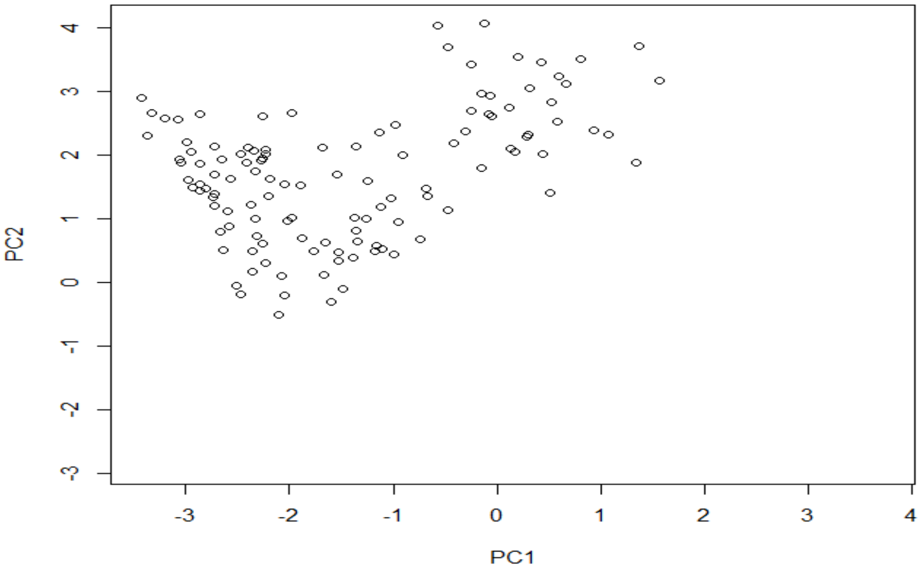
- Red cluster plot:



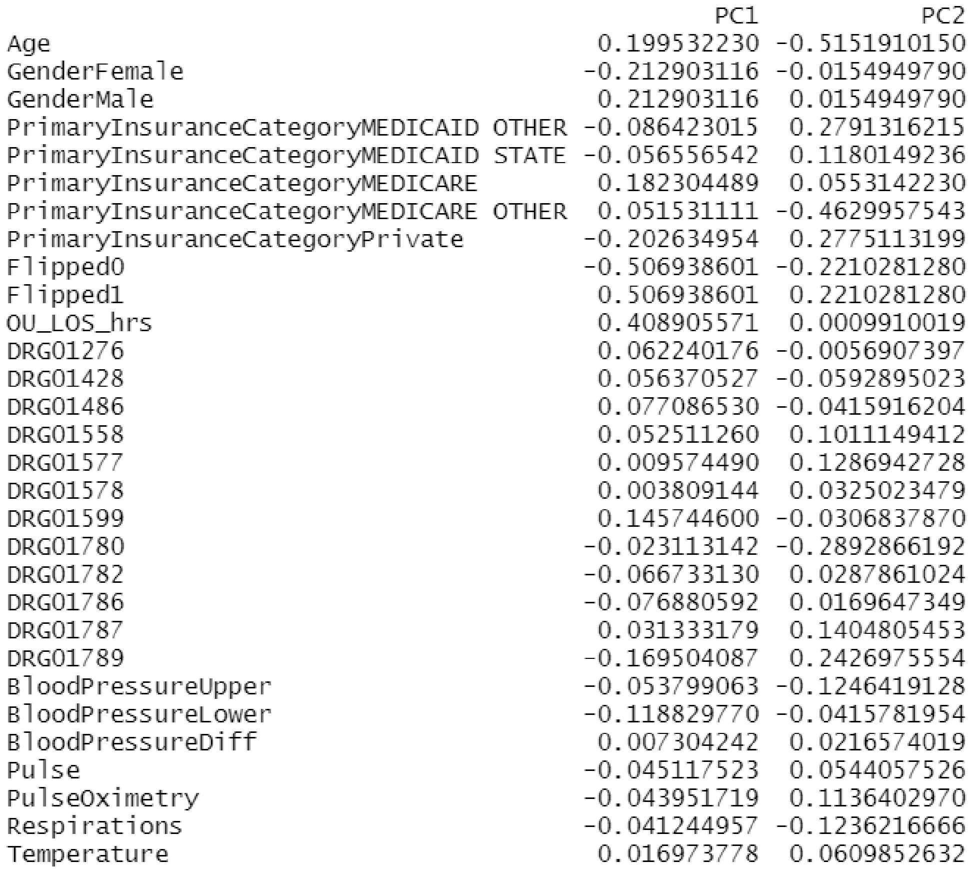
- Green cluster:



- Black Cluster:


PC2
4
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
4
PC1

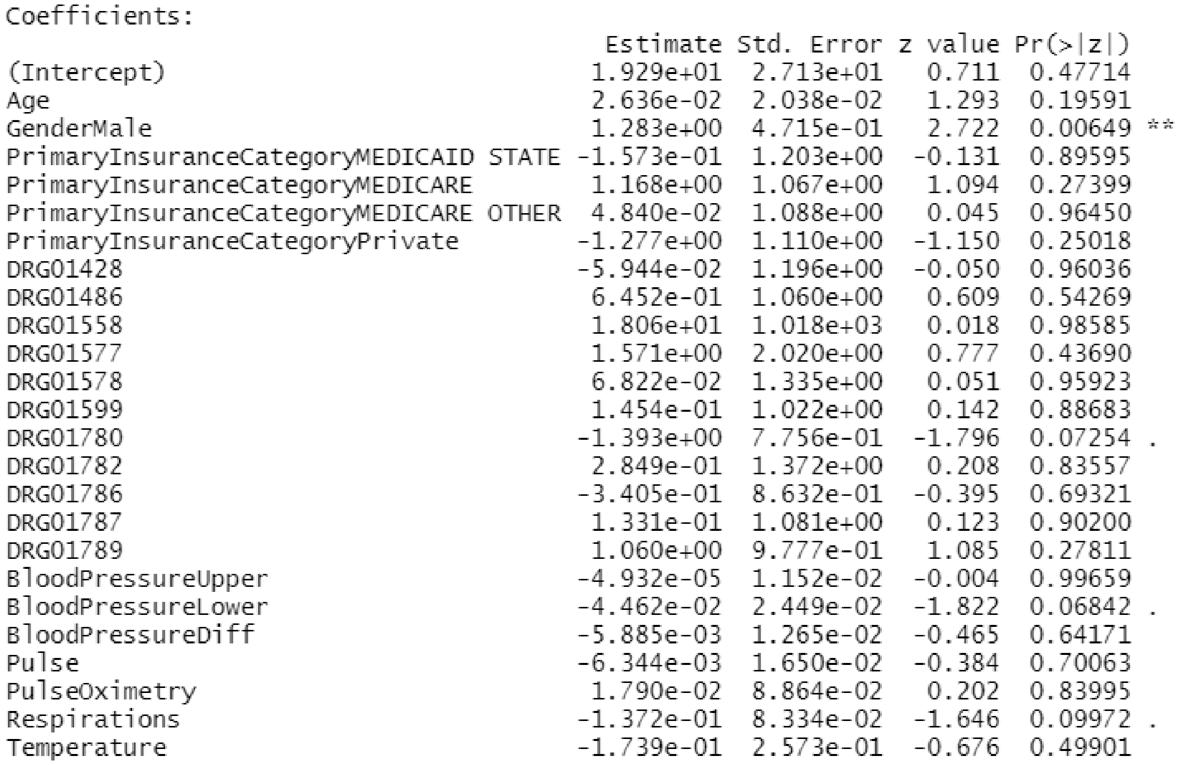
Rotation of PCA:



```
                                              PC1          PC2
Age                                   0.199532230 -0.5151910150
GenderFemale                         -0.212903116 -0.0154949790
GenderMale                            0.212903116  0.0154949790
PrimaryInsuranceCategoryMEDICAID OTHER -0.086423015  0.2791316215
PrimaryInsuranceCategoryMEDICAID STATE -0.056556542  0.1180149236
PrimaryInsuranceCategoryMEDICARE       0.182304489  0.0553142230
PrimaryInsuranceCategoryMEDICARE OTHER  0.051531111 -0.4629957543
PrimaryInsuranceCategoryPrivate       -0.202634954  0.2775113199
Flipped0                             -0.506938601 -0.2210281280
Flipped1                              0.506938601  0.2210281280
OU_LOS_hrs                            0.408905571  0.0009910019
DRG01276                              0.062240176 -0.0056907397
DRG01428                              0.056370527 -0.0592895023
DRG01486                              0.077086530 -0.0415916204
DRG01558                              0.052511260  0.1011149412
DRG01577                              0.009574490  0.1286942728
DRG01578                              0.003809144  0.0325023479
DRG01599                              0.145744600 -0.0306837870
DRG01780                             -0.023113142 -0.2892866192
DRG01782                             -0.066733130  0.0287861024
DRG01786                             -0.076880592  0.0169647349
DRG01787                              0.031333179  0.1404805453
DRG01789                             -0.169504087  0.2426975554
BloodPressureUpper                   -0.053799063 -0.1246419128
BloodPressureLower                   -0.118829770 -0.0415781954
BloodPressureDiff                     0.007304242  0.0216574019
Pulse                                -0.045117523  0.0544057526
PulseOximetry                        -0.043951719  0.1136402970
Respirations                         -0.041244957 -0.1236216666
Temperature                           0.016973778  0.0609852632
```

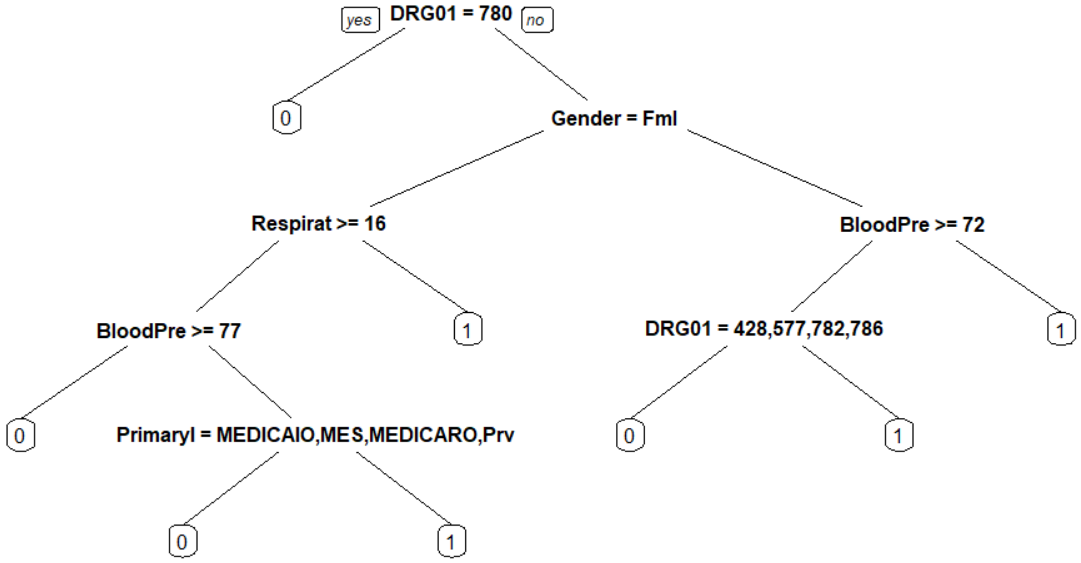
Logistic Regression

- The summary of logistic regression model is:



```
Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            1.929e+01  2.713e+01   0.711  0.47714
Age                                    2.636e-02  2.038e-02   1.293  0.19591
GenderMale                             1.283e+00  4.715e-01   2.722  0.00649 **
PrimaryInsuranceCategoryMEDICAID STATE -1.573e-01  1.203e+00  -0.131  0.89595
PrimaryInsuranceCategoryMEDICARE       1.168e+00  1.067e+00   1.094  0.27399
PrimaryInsuranceCategoryMEDICARE OTHER  4.840e-02  1.088e+00   0.045  0.96450
PrimaryInsuranceCategoryPrivate       -1.277e+00  1.110e+00  -1.150  0.25018
DRG01428                              -5.944e-02  1.196e+00  -0.050  0.96036
DRG01486                               6.452e-01  1.060e+00   0.609  0.54269
DRG01558                               1.806e+01  1.018e+03   0.018  0.98585
DRG01577                               1.571e+00  2.020e+00   0.777  0.43690
DRG01578                               6.822e-02  1.335e+00   0.051  0.95923
DRG01599                               1.454e-01  1.022e+00   0.142  0.88683
DRG01780                              -1.393e+00  7.756e-01  -1.796  0.07254 .
DRG01782                               2.849e-01  1.372e+00   0.208  0.83557
DRG01786                              -3.405e-01  8.632e-01  -0.395  0.69321
DRG01787                               1.331e-01  1.081e+00   0.123  0.90200
DRG01789                               1.060e+00  9.777e-01   1.085  0.27811
BloodPressureUpper                    -4.932e-05  1.152e-02  -0.004  0.99659
BloodPressureLower                    -4.462e-02  2.449e-02  -1.822  0.06842 .
BloodPressureDiff                     -5.885e-03  1.265e-02  -0.465  0.64171
Pulse                                 -6.344e-03  1.650e-02  -0.384  0.70063
PulseOximetry                          1.790e-02  8.864e-02   0.202  0.83995
Respirations                          -1.372e-01  8.334e-02  -1.646  0.09972 .
Temperature                           -1.739e-01  2.573e-01  -0.676  0.49901
```

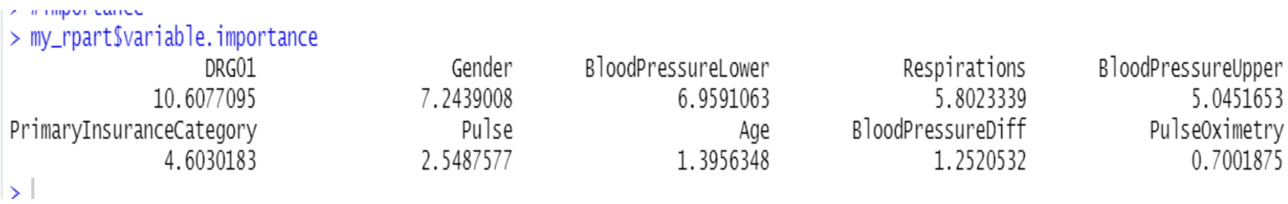
Classification tree



The confusion matrix we got from our data is:

	0	1
0	126	56
1	92	62

- To find the important variable:



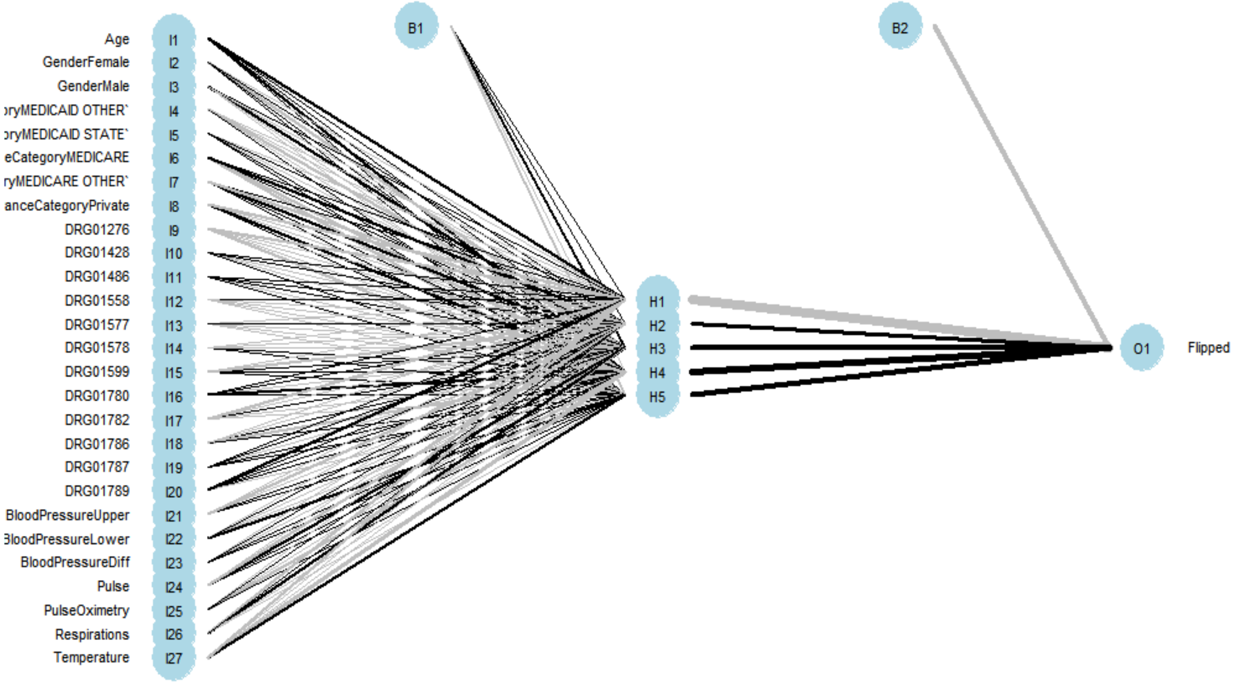
```
> # importance
> my_rpart$variable.importance
                   DRG01                   Gender      BloodPressureLower             Respirations      BloodPressureUpper
              10.6077095                7.2439008               6.9591063                5.8023339               5.0451653
PrimaryInsuranceCategory                    Pulse                     Age        BloodPressureDiff           PulseOximetry
               4.6030183                2.5487577               1.3956348                1.2520532               0.7001875
> |
```

Neural Network

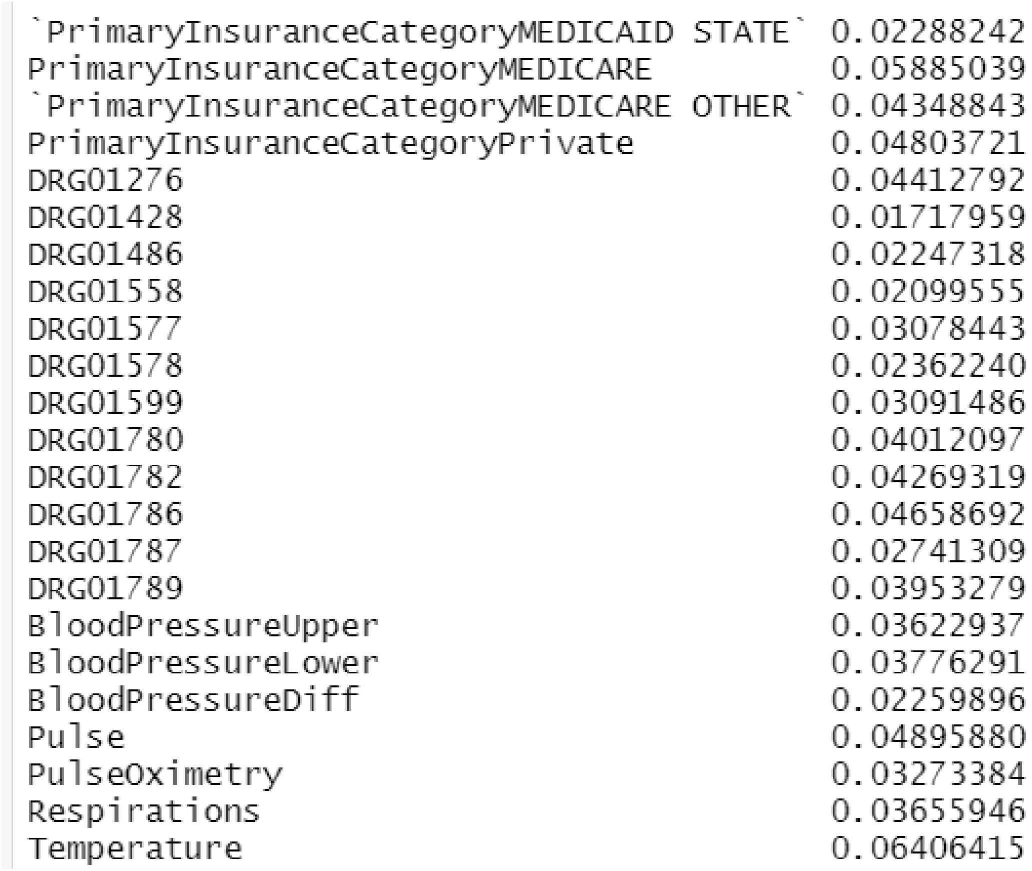
- The confusion matrix:

	Yes	No
Yes	90	64
No	74	108

- Graph:



- Important variables:



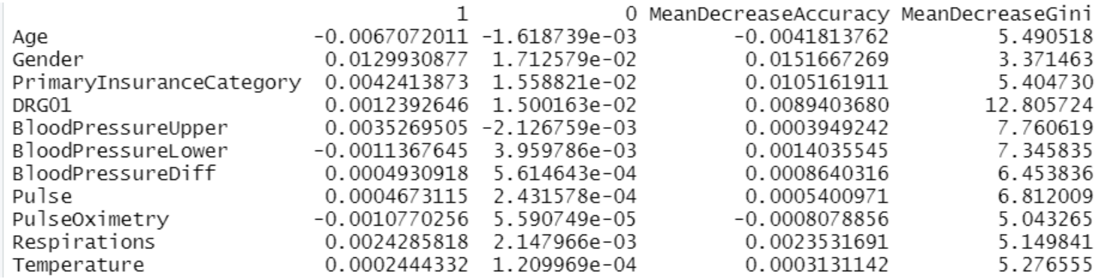
```
`PrimaryInsuranceCategoryMEDICAID STATE` 0.02288242
PrimaryInsuranceCategoryMEDICARE         0.05885039
`PrimaryInsuranceCategoryMEDICARE OTHER` 0.04348843
PrimaryInsuranceCategoryPrivate          0.04803721
DRG01276                                 0.04412792
DRG01428                                 0.01717959
DRG01486                                 0.02247318
DRG01558                                 0.02099555
DRG01577                                 0.03078443
DRG01578                                 0.02362240
DRG01599                                 0.03091486
DRG01780                                 0.04012097
DRG01782                                 0.04269319
DRG01786                                 0.04658692
DRG01787                                 0.02741309
DRG01789                                 0.03953279
BloodPressureUpper                       0.03622937
BloodPressureLower                       0.03776291
BloodPressureDiff                        0.02259896
Pulse                                    0.04895880
PulseOximetry                            0.03273384
Respirations                             0.03655946
Temperature                              0.06406415
```

Random Forest

- Confusion matrix

	1	0
1	76	78
0	69	113

- Important Variables:



```
                                     1             0 MeanDecreaseAccuracy MeanDecreaseGini
Age                      -0.0067072011 -1.618739e-03        -0.0041813762         5.490518
Gender                    0.0129930877  1.712579e-02         0.0151667269         3.371463
PrimaryInsuranceCategory  0.0042413873  1.558821e-02         0.0105161911         5.404730
DRG01                     0.0012392646  1.500163e-02         0.0089403680        12.805724
BloodPressureUpper        0.0035269505 -2.126759e-03         0.0003949242         7.760619
BloodPressureLower       -0.0011367645  3.959786e-03         0.0014035545         7.345835
BloodPressureDiff         0.0004930918  5.614643e-04         0.0008640316         6.453836
Pulse                     0.0004673115  2.431578e-04         0.0005400971         6.812009
PulseOximetry            -0.0010770256  5.590749e-05        -0.0008078856         5.043265
Respirations              0.0024285818  2.147966e-03         0.0023531691         5.149841
Temperature               0.0002444332  1.209969e-04         0.0003131142         5.276555
```