1. Classification Performance Evaluation

```
evaluationfuction = function (data, model, method)
{
 if (method == 'binary-class')
 {
  roc(model)
  auc(model)
  acc(model)
  precvsrec(model)
  aucpr(model)
  lift(model)
  prbe(model)
  print(Concordance(model$y, model$fitted.values))
  dstatistic (data,model)
  chart(data,model)
  matt(model)
  print(F1_Score(data$pred,data$true))
  print(ConfusionDF(data$pred, data$true))
  print(LogLoss(y_pred = model$fitted.Values,y_true=data$pred))
  print(LiftAUC(data$pred, data$true))
  print(GainAUC(data$pred, data$true))
 }
 if (method == 'multi-class')
 {
  print(ConfusionDF(data$pred, data$true))
  print(MultiLogLoss(model$fitted.values, data$true))
  acc(model)
 }
}
```

Documentation

The name of the function is evalutionfunction, the function evaluates for binary class and multiclass classification. For the binary class the model predicts the value of the accuracy, roc curve, area under curve for the ROC curve, precision vs Recall Curves, Lift curves, the break-even point, concordance and discordance, d-stastistics, K-S chart, Matthew correlation, the F1 score, confusion matrix, log loss, lift AUC and Gain AUC. The individual function in this function are written and defined outside the function and available in the r code.

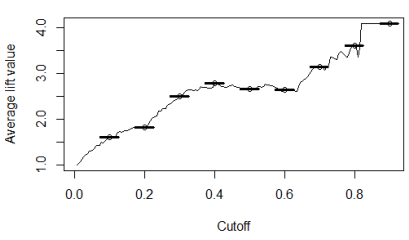
The multiclass model predicts the confusion matrix of the classification, the multi log loss, and the accuracy of the model. The two packages used for the function is the “ROCR” and “MLmetrics”

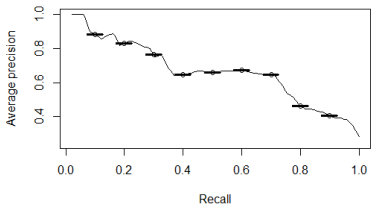
Input

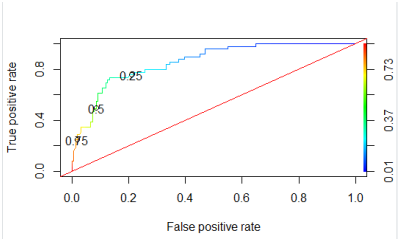
The input for the function is the data which contains the true and predicted labels, the model which is the out of the model fitting. The probabilities are gotten from this model and this is used for the analysis of the function. The method is either “binary-class” or “multi-class”.

Output

This is the output using the honors file to test the function. The result from the function is shown below.

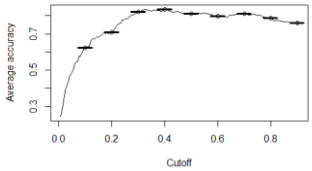


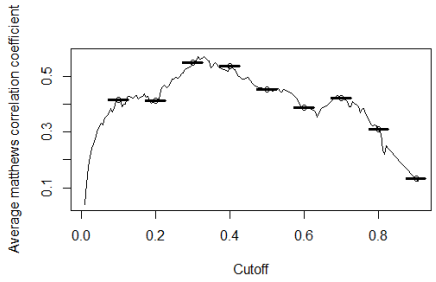





```
> evaluationfuction(honors, fit_honors, 'binary-class')
[1] "auc"
[[1]]
[1] 0.8569401

[1] "aucpr"
[[1]]
[1] 0.6810596

[1] "prbe"
[[1]]
[1] 0.6530612

$Concordance
[1] 0.8567374

$Discordance
[1] 0.1432626

$Tied
[1] 0

$Pairs
[1] 7399

[1] "dstatistics"
[1] 0.3250159
[1] "KS"
  Group   CumPct0    CumPct1        Dif
1     1 0.5165563 0.04081633 0.47573996
2     2 0.6821192 0.20408163 0.47803757
3     3 0.8543046 0.26530612 0.58899851
4     4 0.9072848 0.40816327 0.49912150
5     5 0.9139073 0.51020408 0.40370320
6     6 0.9337748 0.61224490 0.32152994
7     7 0.9668874 0.67346939 0.29341803
8     8 0.9933775 0.85714286 0.13623463
9     9 1.0000000 0.97959184 0.02040816
```

	y_true	y_pred	Freq
1	0	0	138
2	1	0	25
3	0	1	13
4	1	1	24





2a. Logistic Regression

```
fit2 <- glm(data=heartfailure, death ~ serum_creatinine+ejection_fraction+age+sex+
          creatinine_phosphokinase+serum_sodium,
        family="binomial")
summary(fit2)
```

The shows the estimate, standard errors, z-score and p values for each of the coefficients of the model. The model started with adding more predictors variables but the values of p for each of the predictors shows that they are not significant in determining the value of "death" the p vales for "serum creatine" and "ejection_fraction" shows that they are significant in the prediction as the p values are very low(<0.025). Additional age, and creatinine_phosphokinase are also useful in the analysis.

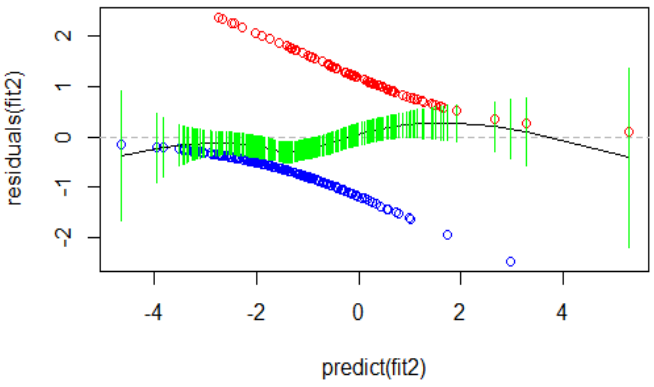
The null residual suggests the response by the model if we consider only the intercept. A lower value equates to a better model. This value is constant irrespective of the other predictors the residual

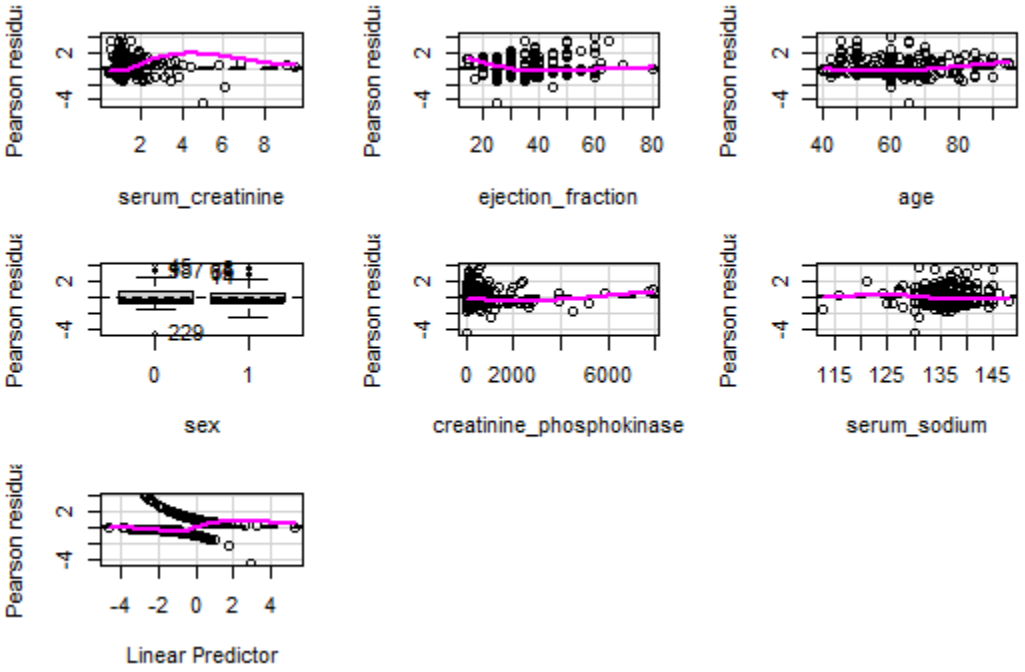
deviance indicates the response of the model when all the variables are included. These values was observed to reduce as some more predictor variables was added to the model.

Comparing the two models above with the AIC and deviance residuals show that the second model is better. The AIC is lower for the second model; from 283.11 to 272.55 and the residual deviance reduces from 277.11 to 258.55. Therefore the second model is picked as the working model.

The hosmer-lemeshow goodness of fit test. This is to show how well the model it's the data.
hoslem.test(heartfailure$death, fitted(fit2))
The p-value here is 0.5365 which means that there is no significant evidence to show that the model is a poor fit for the data.
hoslem.test(heartfailure$death, fitted(fit1))
The p-value here is 0.002587 which means that there is significant evidence to show that the model is a poor fit for the data.

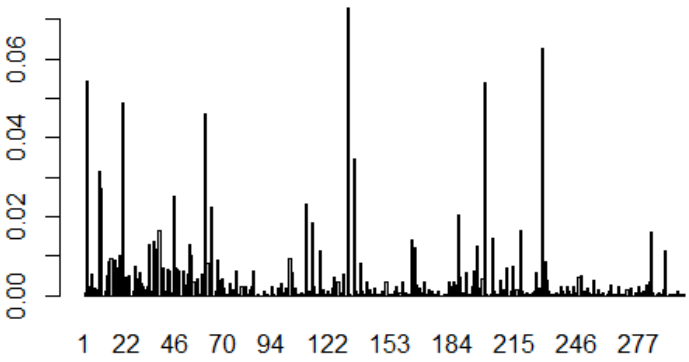
Residuals Diagnosis
The residual analysis show that when two lines, the blue line and the red line. If the true value is 0, then we always predict more and have a negative residuals (the blue line) and if it is 1, then we would under estimate and the residuals would be positive (the red points).
The green line is supposed to fit at the horizontal line for a good model and the line is along the zero line which means it is a good model with residuals almost zero.



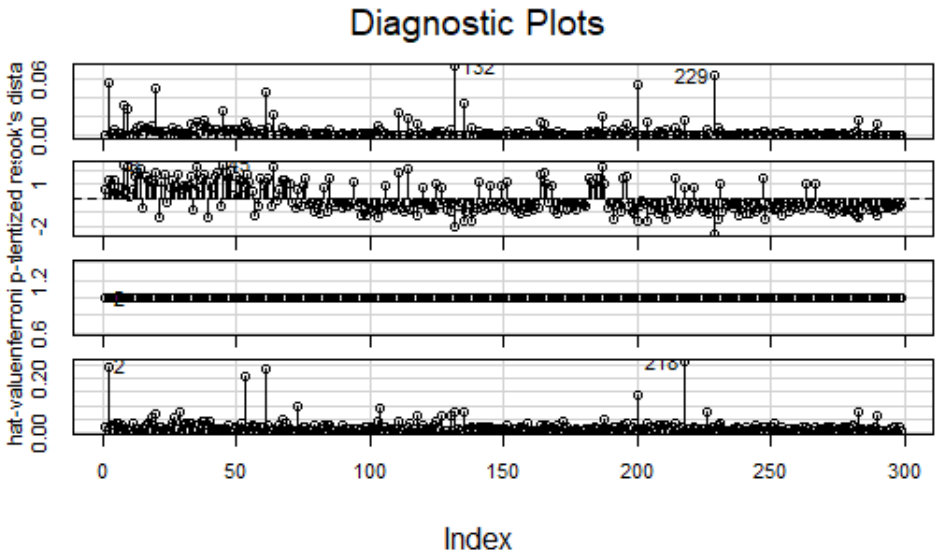


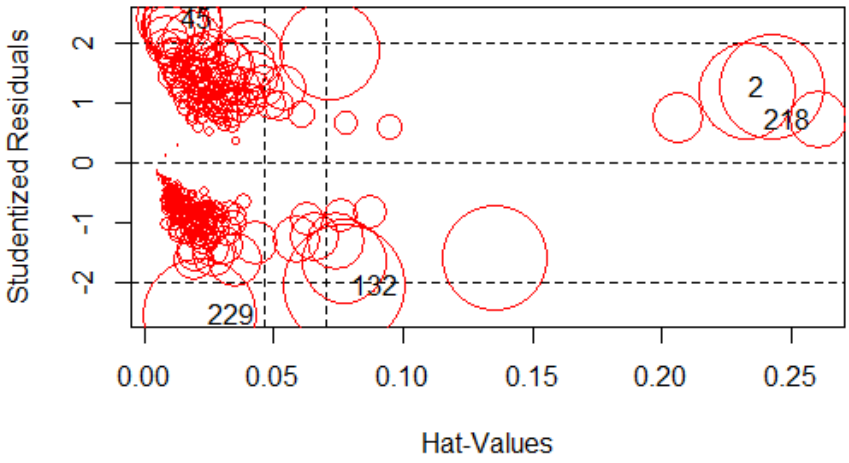
This plot shows how properly fitted the regression model is. It plots each of the predictors against the pearson residuals. if the model is properly fitted, there would be no correlation between the predictor and the residuals. The trend would be a horizontal line or very close to one. The lines for each of the predictors variables can approximate a horizontal line. The lines are horizontal, the model can be said to be well fitted. also, the stastiscal testing is done to check if any of the predictors is still statistically significant. The idea is that if the model is properly specified, no additional predictor variables that are statisifcally significant can be found.

Influence



Observations 2 and 218 have the highest hat values, 132 and 229 have the largest cook distance from the plot. this points could be outliers according to the cooks distance. If all the cook distance is high then the model is not well fitted to the data.In this case however, the cook distance are low and only few observations have a relatively high.





Observations that have a substatinal effect in the estimates of coefficient are called influential observations. Influence is a product of the leverage and outliers. The observations 229 has the largest standardized residuals and second largest cook distance. 132 has the largest cook distance.

Variance Inflation

vif(fit2)

serum_creatinine	ejection_fraction	age
1.040885	1.124992	1.097697
sex	**creatinine_phosphokinase**	**serum_sodium**
1.062534	1.046058	1.044524

The variables do not show any correlation to each other as they have a low value of vif for each of the predictor variables. This means the model features are not highly multicorrelated.

Logistic Regression Coefficient Interpretation

exp(coefficients(fit2))

exp(confint(fit2))

(Intercept)	serum_creatinine	ejection_fraction
152.1219577	1.8766124	0.9321696
age	**sex1**	**creatinine_phosphokinase**
1.0569629	0.6396887	1.0002337
serum_sodium		
0.9471832		

	2.5 %	97.5 %
(Intercept)	0.02136931	1.351249e+06
serum_creatinine	1.40103346	2.679655e+00
ejection_fraction	0.90445657	9.584947e-01
age	1.03153221	1.084585e+00
sex1	0.35031338	1.162506e+00
creatinine_phos	0.99996294	1.000516e+00
serum_sodium	0.88658198	1.010016e+00

The intercept indicate the log odds of the whole population of interest to be on die from a heart attack with no predictor variables in the model.

The model is a multivariate logistic model.

1. serum_creatinine: after adjusting for all the cofounders (ejection_fraction, age, sex, creatinine_phosphokinase, serum_sodium and smoking), the odd ratio is 1.88, with 95% CI being 1.401 and 2.68. This means that with every 1 increase in the serum_creatinine level, the odd of dying from a heart attack increases by 88%.

2. ejection fraction: after adjusting for all the cofounders (serum_creatinine, age, sex, creatinine_phosphokinase, serum_sodium and smoking), the odd ratio is0.93219, with 95% CI being 0.90448 and 0.95852. This means that with every 1 increase in the ejection_fraction, the odd of dying from a heart attack decreases by 6.79%. this means the chnances of survival increases by 6.79% for a unit increase in ejection fracrtion.

3. age: after adjusting for all the cofounders (ejection_fraction, serum_creatnine, sex, creatinine_phosphokinase, serum_sodium and smoking), the odd ratio is

1.057, with 95% CI being 1.0316 and 1.08467. This means that with every 1 increase in the age, the odd of dying from a heart attack increases by 5.7%.

4. sex: after adjusting for all the cofounders (ejection_fraction, age, serum_creatnine, creatinine_phosphokinase, serum_sodium and smoking), the odd ratio is 0.6222, with 95% CI being 0.3175 and 1.204.this means that the odd of surviving for male (sex=0) is 37.78% less likely as compared to female (sex = 0)

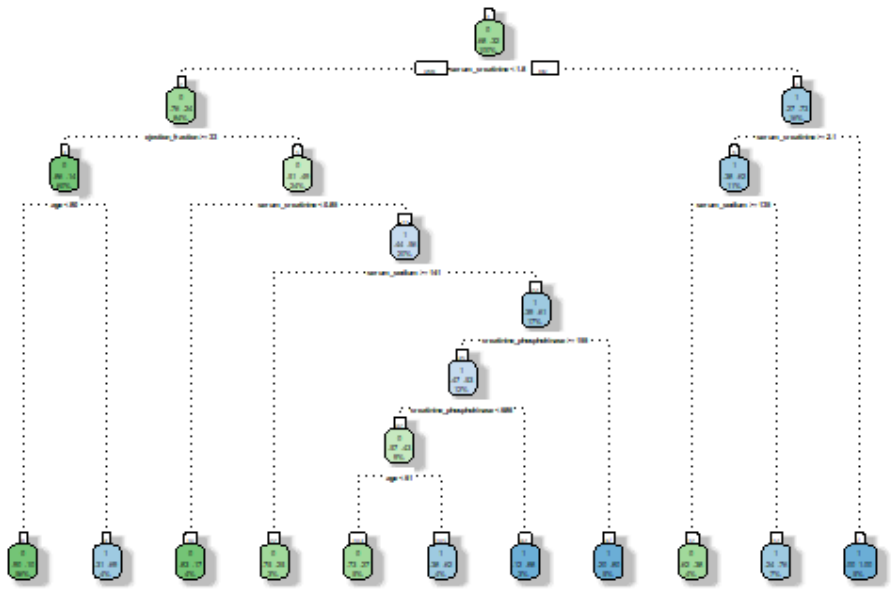
5. creatinine_phosphokinase: after adjusting for all the cofounders (ejection_fraction, serum_creatnine, sex, age, serum_sodium and smoking), the odd ratio is 1.00002337, with 95% CI being 0.9999 and 1.000519. This means that with odd ratio is very close to one which means the change in creatinine_phosphokinase does not have a significant impact on the survival rate.

6: serum_sodium: after adjusting for all the cofounders (ejection_fraction, serum_creatnine, sex, creatinine_phosphokinase, age and smoking), the odd ratio is 0.9471, with 95% CI being 0.88635 and 1.01 This means that with every 1 increase in the serum_sodium, the odd of dying from a heart attack decreases by 5.29%.this means the chances of survival increases by 5.29% for 1 increase in sodium_serum level.

bi. Decision Tree

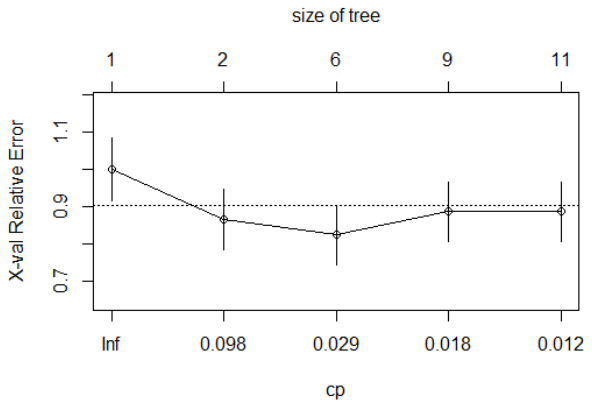
```
fitDT1<-rpart(death~serum_creatinine+ejection_fraction+age+sex+
              creatinine_phosphokinase+serum_sodium,data=heartfailure,
          parms=list(split="information"),
          control=rpart.control(xval=20))
summary(fitDT1)
```

bii.



Rattle 2020-Nov-02 21:10:44 oyino

The above tree is too complex and has to be pruned.



This plot shows the cp value that gives the minimum CV error. The minumimum point in this case is the 0.029 which is closely followed by 0.098.

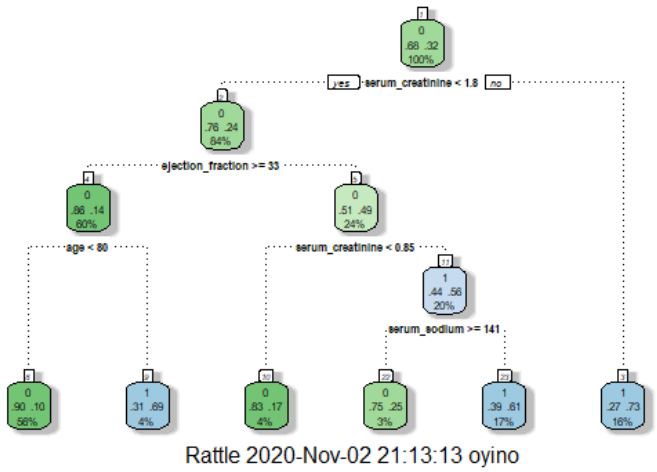
```
pfit1<-prune(fitDT1,cp=0.029)
fancyRpartPlot(pfit1) #pruning at the level of cp = 0.029 to remove the complexity of the tree.
```

Variable importance

serum_creatinine	ejection_fraction	age	creatinine_phosphokinase	serum_sodium	sex
40	22	20	9	8	1

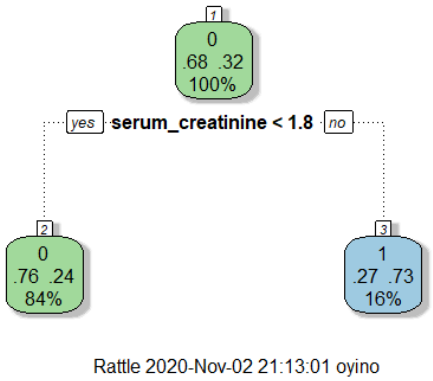
The most important varibale for the prediction of death from a stroke is the serum_creatinine. This follwed by ejection_fraction. This agrees with the paper done on this work. Age is the third most important feature. The sex is very low in important.

This tree uses the serum cretinine, ejection_fraction and also age to split the target varibale “death” as either 0 or 1. For example, if serum_cretainine is less than 1.8 and the ejection frcation is greater than 33 and the patient is below the age of 80, the patient survives the stroke. If above the age of 80, the patient does not survive the stroke.
And also, directly from the first root, if the serum_cretinine is greater than 1.8, then the patient dies from the stroke.



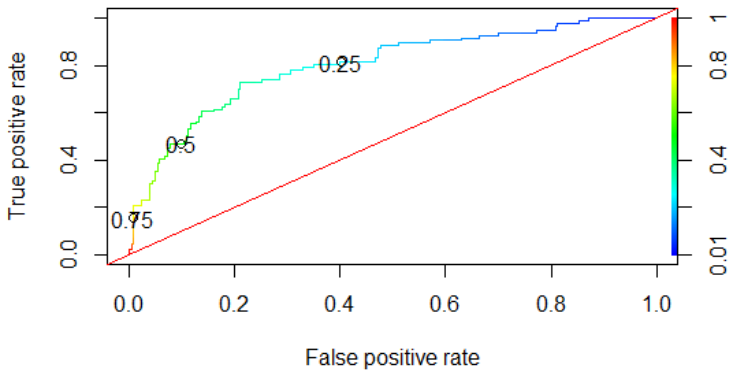
pfit<-prune(fitDT1,cp=0.098) #and we can prune to this level. This level gives
# a less complex tree with two leaves nodes.
fancyRpartPlot(pfit)

This is a simple model that uses only the serum_creatinie as the spilting criateria. This is the most important feature accoring to the model. If serum_creatinine is less than 1.8, then target is predicted as 0 (patients that survive) and if it is greater than 1.8 the target is 1 (patients that died).



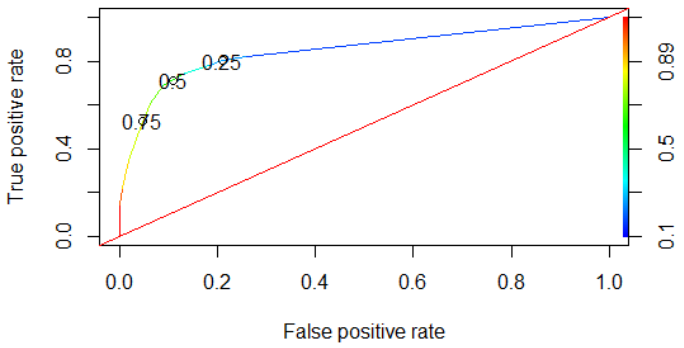
c. Model Metrics and Perfromance

1. **ROC and AUC**

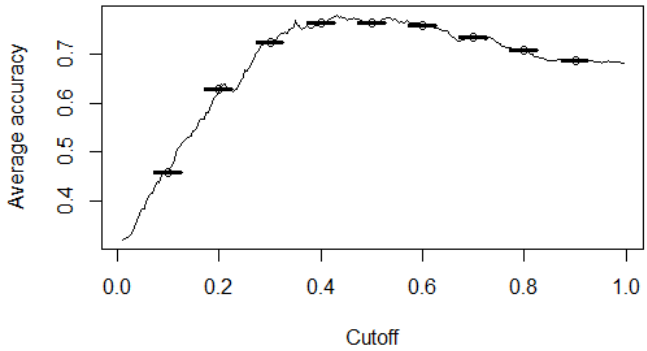


The roc shows the tpr vs the fnr. This the most effective evaluation metrics because it visualizes the accuracy of predictions for a whole range of cutoff values. if we had a perfect model, the ROC curve would pass through the upper left corner — indicating no error. The most important takeout from a roc is the auc. the auc is the measure of the ROC curve. The auc is **0.7992**. Usually, a auc above 0.7 is a good model. The closer the auc to 1, the better the model.

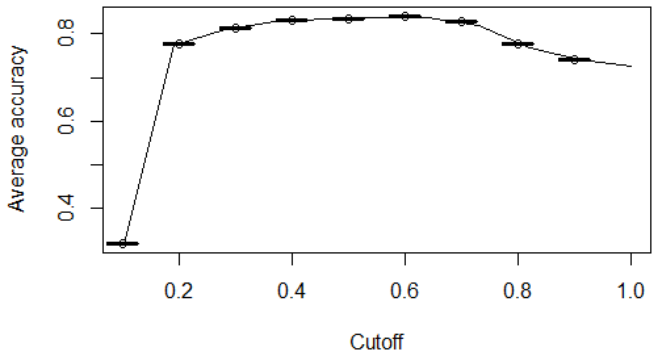
The second ROC shows the result for the decision tree, the curve is closer to the upper left side of the plot. The ROC curve is higher at the cut off. **AUC value is 0.8487018. The results for the decision tree is better.**



2. Accuracy

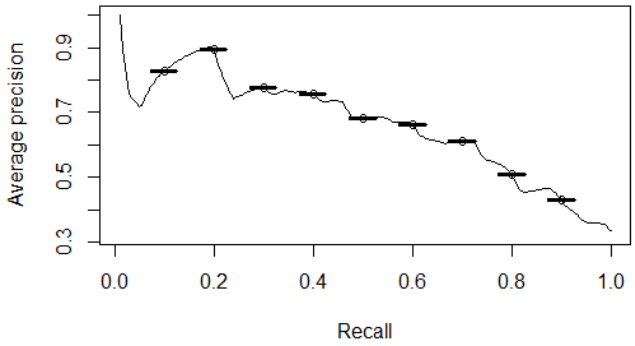


The highest accuracy is at a cut off of 0.5. Theses plot can be used to select the cut for the classification. The highest accuracy obtained is about 75%.



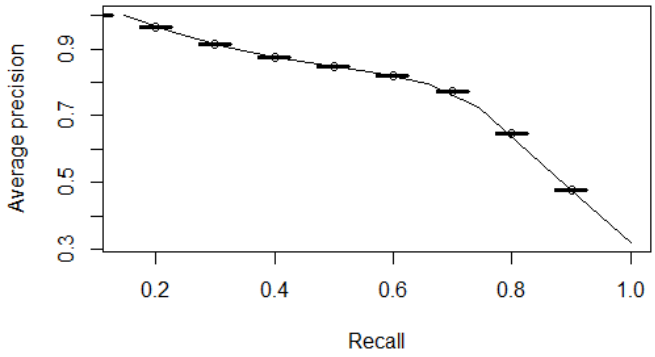
The accuracy curve is higher for the descion tree model. The highest accuracy is at a cut off of 0.5. Theses plot can be used to select the cut for the classification. The highest accuracy obtained is about 83%. The decision tree performed better.

3. Precison and Recall

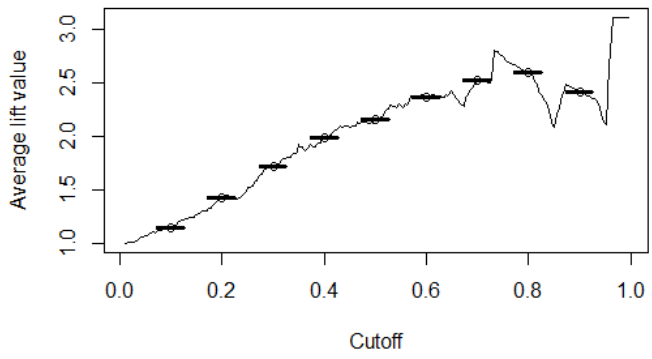


Recall is also known as the true positive rate: the amount of positives your model claims compared to the actual number of positives there are throughout the data. Precision is also known as the positive predictive value, and it is a measure of the amount of accurate positives your model claims compared to the number of positives it actually claims.

This shows the graph of precision vs recall and gives an aucpr of 0.6715687. These shows how good the precison and recall of the model is. The higher the area under the curve the better the model. This area is below 0.7 which is not very good.The decision tree model give a higer curve and a higher area under curve. The auc for the precison and recall curve is 0.791. This is higher than 0.7 which is good.

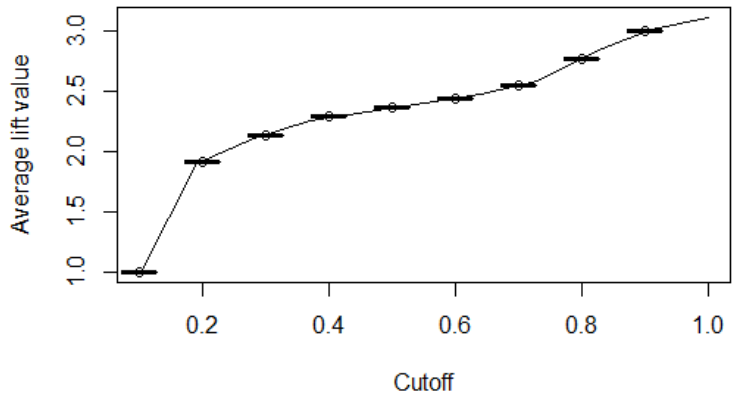


4. Lift Curves and AUC lift curve



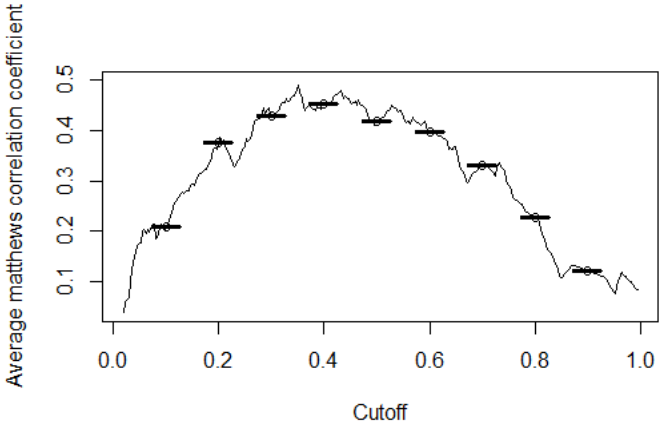
Using the predictive model using data from other patients, "0 / 1" is identified as the "target" field and other factors are used as predictors. As a result of the predictive model, we are able to sort entire patient list in decreasing order of expected death. Consequently, rather than a random selection, the patients can be sorted based on the most likely and so on.the lift line (response before predictive model) indicates the gained from using the predictive model. Gain > 1 means the results from the predictive model are better than random.

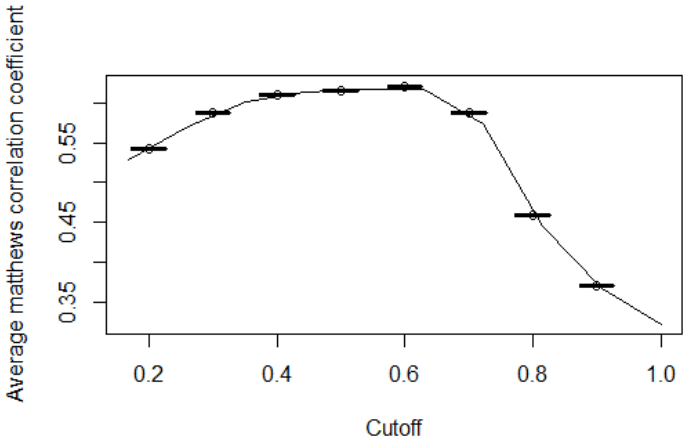
A lift chart shows the actual lift, which is the ratio between results with an without the model. In data mining, you can think of it as measuring "…the change in concentration of a particular class when the model is used to select a group from the general population." The higher the lift
(i.e. the further up it is from the baseline), the better the model. The lift curves here shows the lift values against the cutoff. This can help the lift vluae for the model for the cutoff vlaue of the model. The decision tree model has higher lift values at the same cutoffs.



5. Mattew Correlation

For the first curve; logistic regression, the mattew correlation is 0.48 at a cut off 0f 0.5. The second curve for decision trees has a mattew correlation of 0.6 at a cut off od 0.5. This shows that the decision tree has a better model for the dataset.





5. Confusion Matrix

Logistic Regression

y_true	y_pred	Freq
1	0	0 184
2	1	0 51
3	0	1 19
4	1	1 45

Decision Trees

y_true	y_pred	Freq
1	0	0 185
2	1	0 30
3	0	1 18
4	1	1 66

Comparing the Logisitic regression and Decision tree, the decision tree has a higher true positive of 185 and also high true negative of 66 as agaisnt the 184 and 45 of the logistic regression. The FP and FN was also lower for decision tree as 18 and 30 respectively as agaisnt 19 and 51 for the logistic regression.

6. **Log Loss**

The best model would have the lowest log loss possible. The log loss is the penalization for the bad predictions. Higher losses are given to prediction probalilities closer to 0 and lower losses to predictions closer to 1.

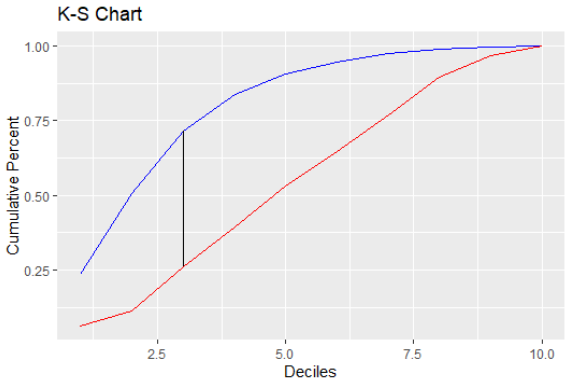
Log Loss

0.2938122

7. **F1 Score**

This can be used to determine the model performance. It however does not include true negatives in its analysis and should only be used in cases where true negatives are not very important. For the logistic model; F1 = 0.84 and for the decision tree, the F1 = 0.89. The decsion tree does a better job than the logitic regression.

8. K-S Chart



K-S measures the degree of separation between the distributions of the positive and negative death value. K-S = |cumulative % of total death — cumulative % of survivors|. The higher the value, the better the model is at separating the positive from negative cases. If a model cannot separate positive from negative cases, the K-S for all deciles will be 0. The best decile is the decile 3 has it the highest separation.

Group CumPct0 CumPct1 Dif

3 3 0.7142857 0.2604167 0.453869

The maximum separation is 0.4539.

9. **D-statistics**

The dstatistics shows the difference between the predictions for the 1 and 0. This shows the separation between the two classifications. The larger the value of separation the better the model. The d statistics value for this model is 0.2483.

"dstatistics"

0.2482999

10. Concordance and Disconcordance

Concordance: In how many pairs does the probability of ones is higher than the probability of zeros divided by the total number of possible pairs. The higher the values better is the model.
The value of concordance lies between 0 and 1. Similar to concordance, we have disconcordance which states in how many pairs the probability of ones was less than zeros. If the probability of ones is equal to 1 we say it is a tied pair.The number of tied pairs in this case is 0.

Logistic Regression

$Concordance	**$Discordance**
[1] 0.2002258	**[1] 0.7997742**

Decision Tree

$Concordance	**$Discordance**
0.775195	**0.224805**

Conclusion:

The decision tree has a model that does a better classification of the labels than the logitic regression based on this analysis of the model and it is a model good enough for these classification problem. I would trust the decision tree in the classification.