Project: Breast cancer analysis

Predicting benign vs malignant breast cancers using R

Submitted in partial fulfillment of the requirements of the course

BANA 4080 Data mining

By

Adrian Valles

Tarushi Ravindra

Kendra Mendoza

Alexandre Murray

Project Background:

The following data set was collected by Dr.William H Wolberg, a physician at the University of Wisconsin Hospitals. Donated by Olvi Mangasarian and received by David W. Aha in 1992,it is important to note that sample sets of the data arrived periodically therefore the dataset reflects this chronological grouping. Each variable is also normalized to a 1-10 scale, commonly implemented by doctors for categorizing potential indications of cancer. Variables include: Clump Thickness,Uniformity of Cell Size,Uniformity of Cell Shape, Marginal Adhesion,Single Epithelial Cell Size,Bare Nuclei,Bland Chromatin,and Normal Nucleoli. In the past this data set was sourced to prove multisurface methods of pattern separation for medical diagnosis applied to breast cytology. **Overall, the purpose for collecting the data was to detect benign vs malignant cancer cells and this is how we used the information.**

Cleaning the data:

1. This dataset has no column names, so we will need to add the column descriptions as names for each variable.
2. Variable 7 (BareNuclei) has some null ('?') values, so we will coerce it from it's original character form to an integer, where the ?'s will be parsed as proper N/A values.
3. We should then remove the observations associated with null Bare Nuclei values, since this variable is quite important and replacing the nulls with means in this case would likely not be a good practice given our focus on proper classification.
4. The ID variable can be coerced into a factor, since it should not be interpreted for its numeric value
5. The response variable, 'Class', is represented as a 4 or 2, representing a 1 and 0 respectively for the binary case. We will accordingly divide the variable by 2, then subtract one to get the appropriate binary representation.
6. We will convert 'Class' to a factor once necessary so that the random forests can have a classification type, as other variable types would be interpreted as a regression type.
7. We will split our set of 683 observations into training and testing sets on a 80/20% split.

EDA:

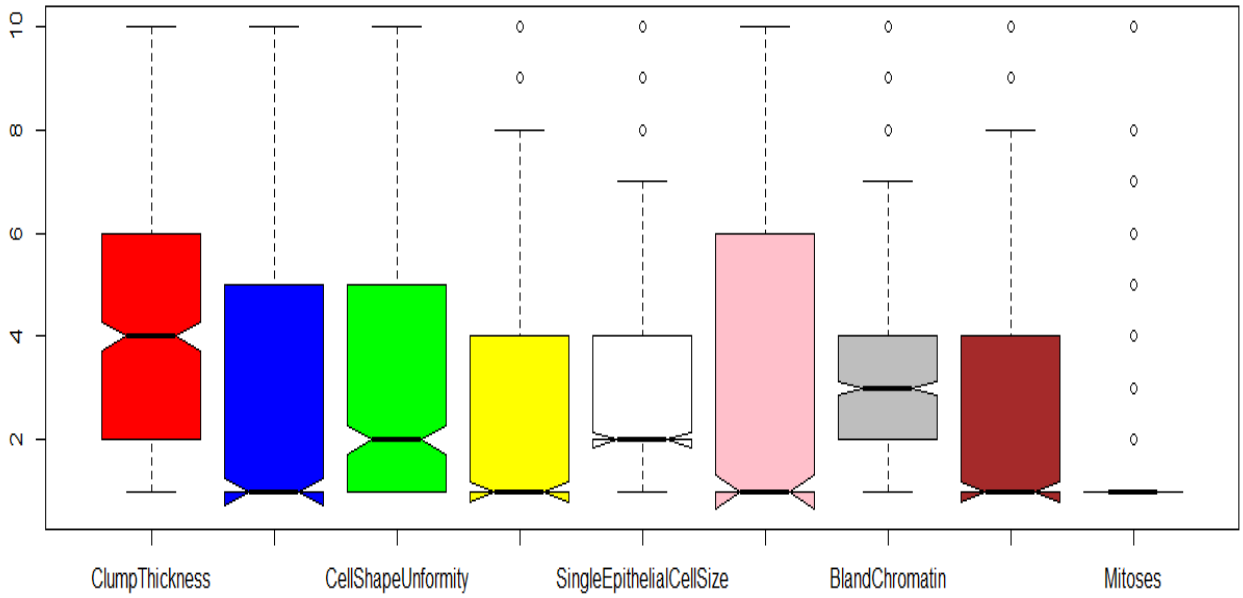
Our first step after cleaning our data in the project was to understand our data. We started analysing our dataset by simply exploring it through various EDA methods. To understand our variables better, we ran a statistics summary report and found the following statistics about our dataset:

Variables	Min	1st Quantile	Median	3rd Quantile	Max	Mean	SD
ClumpThickness	1	2	4	6	10	4.48	2.87
CellSizeUniformity	1	1	1	5	10	3.20	3.10
CellShapeUnformity	1	1	1	5	10	3.30	3.05
MarginalAdhesion	1	1	1	4	10	2.81	2.84
SingleEpithelialCellSize	1	2	2	4	10	3.28	2.27
BareNuclei	1	1	1	6	10	3.54	3.65
BlandChromatin	1	2	3	5	10	3.48	2.47
NormalNucleoli	1	1	1	4	10	2.86	3.06
Mitoses	1	1	1	1	10	1.61	1.77

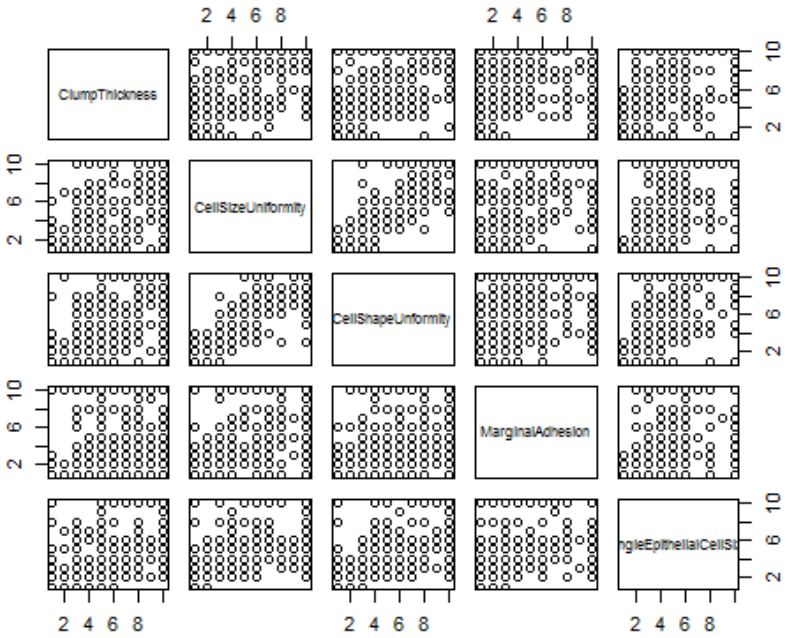
From the above summary, we can comprehend that all our independent variables lie between 1 and 10 (the values have standardized). Clump Thickness has highest median and mean suggesting its values are generally higher than the other variables.

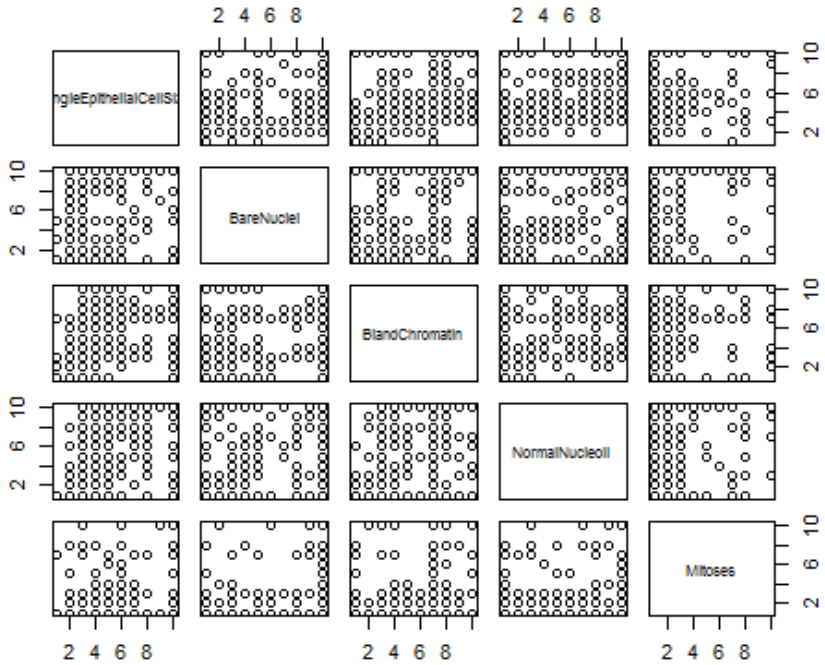
Box plot:

Next, we wanted to analyze our dataset visually, so we plotted a box plot. The box plot visually represents our summary statistics since it shows each variable's min, max, mean, standard deviation etc.



Next, to analyze how our independent variables behaved with each other, we show a scatter plot

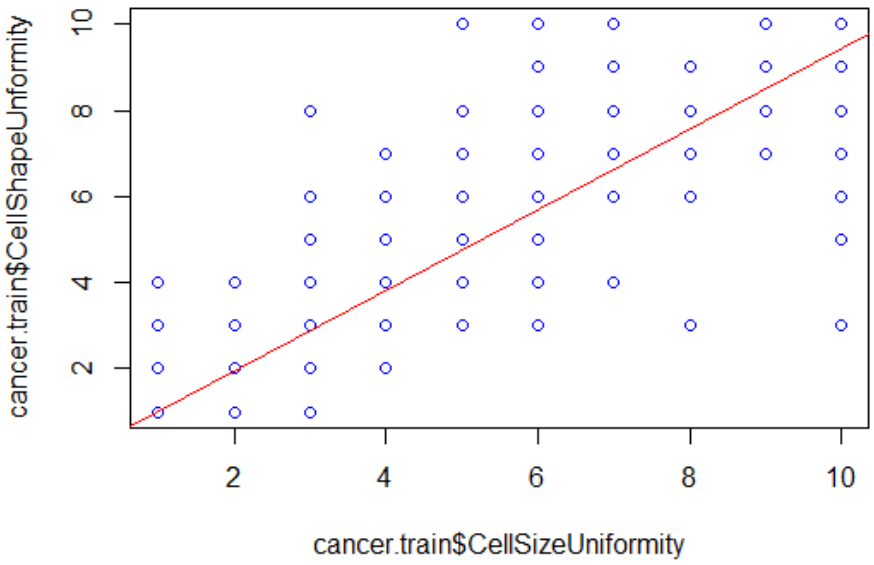




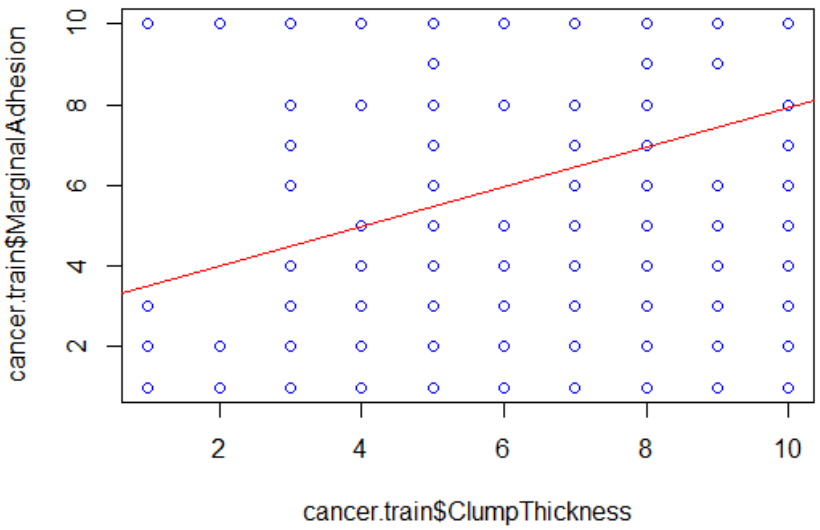
At a first glance, we can see that most of the variables show no correlation (+ve or -ve) with each other except two variables which are Cell Size Uniformity and Cell Shape Uniformity. These 2 variables show a somewhat positive correlation depicting similar behavior.

Closer Analysis:

Further analysis between some of the variables individually confirmed our initial hypotheses. We simply plotted some variables and tried fitting the best regression line.



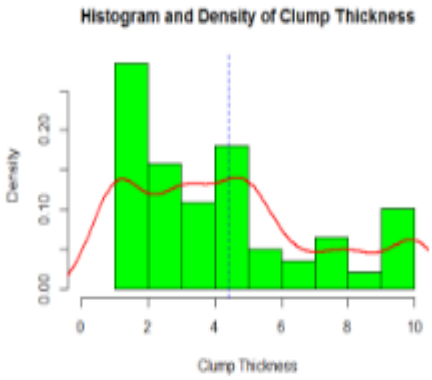
As can be seen, Cell Size Uniformity and Cell Shape Uniformity show a linear and positive relationship.

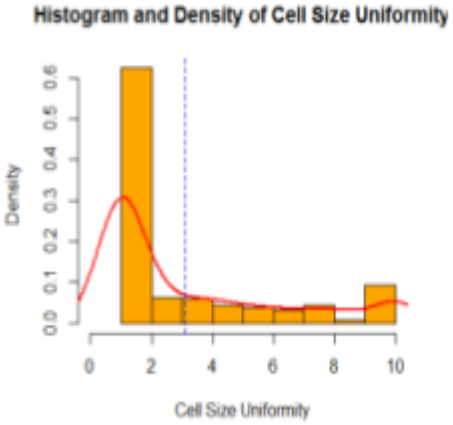


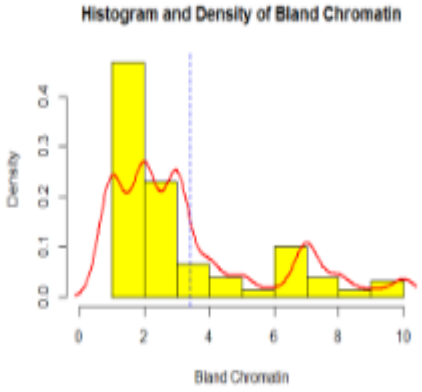
The other variables however don't behave typically with each other and hence show a scattered and noisy scatter plot.

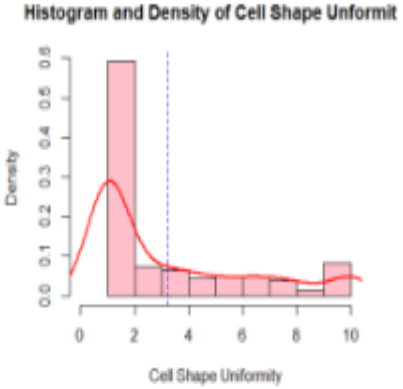
Histograms:

Our last graphical analysis method was to plot histograms. We wanted to see the kind of distribution each variable had.









Each of these graphs prove one thing: When dealing with real data, we always don't get the much-desired normal distribution. Clump Thickness data points seemed to be more evenly distributed compared to other variables. Cell Size Uniformity and Cell Shape Uniformity have their highest peak between 1 and 2 with 3 being the average depicted by the dashed line. Bland Chromatin also weighed heavily between 1 and 3 with an average of 3.5

Logistic regression:

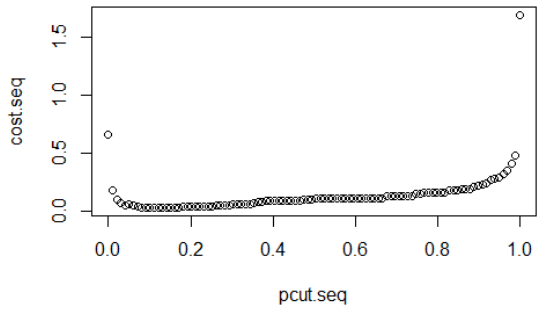
Next, we conducted logistics regression analysis on our dataset. We are trying to find the best model to predict whether the cancer would be benign or malign. Therefore we initiate the analysis with the full logistic regression model.

Class= -10.06706 + 0.47651ClumpThickness - 0.01403CellSizeUniformity +0.34678CellShapeUnformity+ 0.34073MarginalAdhesion+ 0.10985 SingleEpithalCellSize+0.35981BareNuclei+ 0.40305BlandChromatin +0.22539NormalNucleoli +0.62166Mitoses

In order to study this model, we are going to assume an asymmetric cost where the FN-FR ratio is 5:1. The idea behind this number is that obviously FN are more serious because predicting a cancer as benign when in fact is malign may result in horrible consequences. However, the ratio is not higher because FP errors may lead to unnecessary operations as well as life alterations. The table below shows a summary of this explanation.

	Pred=1(malignant)	**Pred=0 (benign)**
True=1 (malignant)		**FN-Weight=5**
True=0 (Benign)	**FP-Weight=1**	

Given the asymmetric cost, we studied the optimal p-cut for the model through the grid search method. As it can be seen in the table below, the p-cut that minimizes the total cost is p-cut=0.11111



Given this p-cut the confusion matrix obtained on the training and testing set are the following:

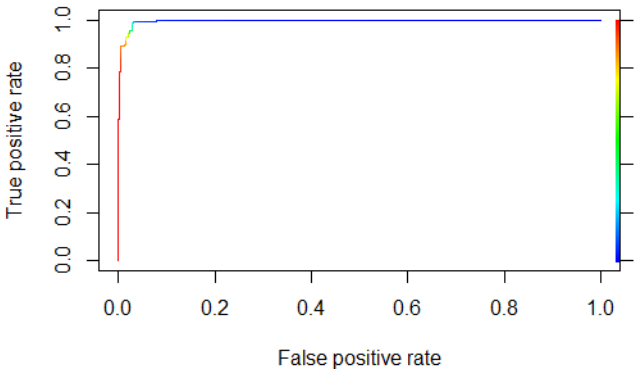
Confusion matrix on training set		
	0	1
0	350	11
1	1	183

Confusion matrix on testing set		
	0	1
0	78	4
1	1	54

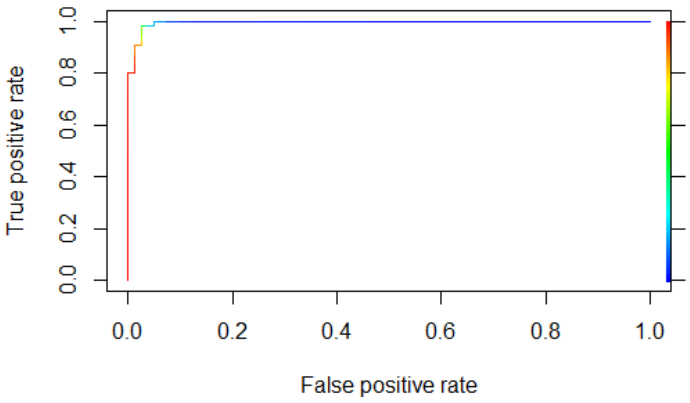
*A summary comparison with all the misclassification rates will be provided later.

Regarding the ROC curves, the pots obtained are the following.

In sample: (AUC=0.9963116)



Out of sample:(AUC=0.9960089)



In an effort to improve the model, we have tried to use a stepwise selection process to get a better mode. The model obtained is the following:

Class=-9.9197 + 0.4761ClumpThickness 0.3692CellShapeUnformity+ 0.3555 MarginalAdh +0.3601BareNuclei + 0.427 BlandChromatin1 +0.2326 NormalNucleoli +0.6142 Mitoses .

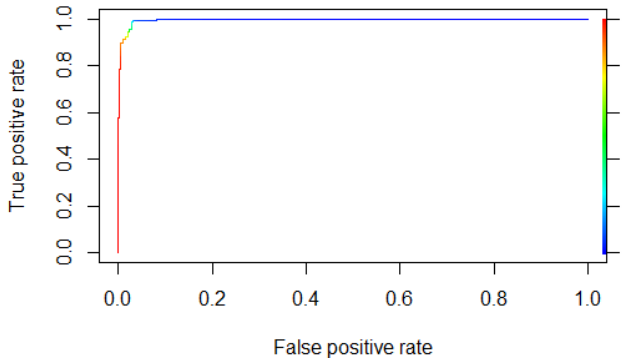
The confusion matrix for this model are the following

Confusion matrix on training set		
	0	1
0	350	11
1	1	183

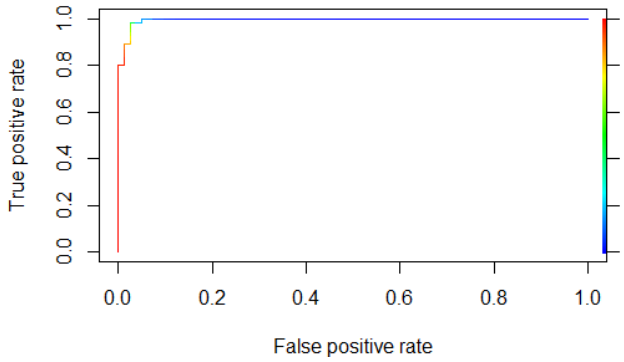
Confusion matrix on testing set		
	0	1
0	78	4
1	1	54

Regarding the ROC curves associated to these models, we got the following plots.

In sample(AUC=0.9963567)



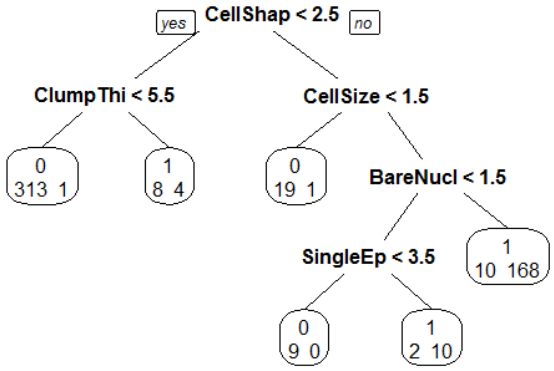
Out of sample(AUC=0.9957871)



As we will see in the summary table (located after the classification tree section) the full model and the improved one get very similar results (excellent ones). However, the best model slightly outperforms the full one in AIC and BIC.

Classification trees:

We start our analysis with an initial tree, given a default cp value=0.01 and keeping the asymmetric cost with the same values than in the logistic regression. The tree that we get is the following:



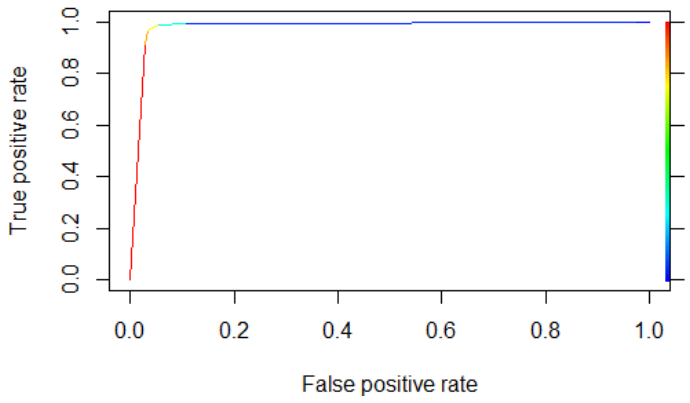
This tree leads to a confusion matrix with a higher misclassification rate and cost than both of the logistic regression models. The confusion matrices obtained are:

Confusion matrix on training set		
	0	1
0	341	20
1	2	182

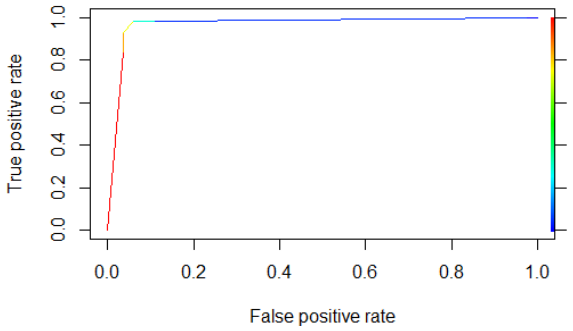
Confusion matrix on testing set		
	0	1
0	77	5
1	1	55

Following the same pattern, the ROC and AUC values are slightly worse than the logistic regression models

In sample (AUC=0.9813471)

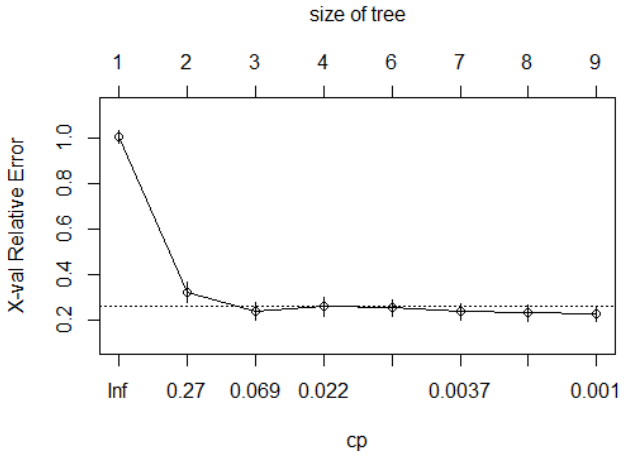


Out of sample (AUC=0.968847)

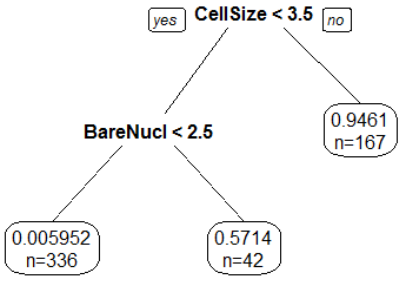


In an effort to improve the classification tree model, we prune it with the next data:

	CP	nsplit	rel error	xerror	xstd
1	0.731482558	0	1.00000000	1.0029639	0.02952059
2	0.097948298	1	0.26851744	0.3123877	0.04334902
3	0.048682024	2	0.17056914	0.2116042	0.03466583
4	0.009731879	3	0.12188712	0.1848823	0.03535058
5	0.004004756	5	0.10242336	0.1985473	0.03521463
6	0.003407049	6	0.09841861	0.1852183	0.03248205
7	0.001097641	7	0.09501156	0.1857067	0.03260420
8	0.001000000	8	0.09391392	0.1771720	0.03030360



Given both the table and the plot, we assign a cp=0.069 to the new tree, as it is a value that gives low error while keeping a low level of complexity. The new pruned tree is the following:



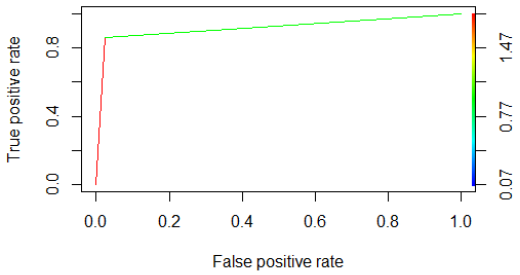
We quickly realize that this model is probably too simple and is not accurate enough for predicting such a serious topic. The model presents the following values:

Confusion matrix on training set		
	0	1
0	352	9
1	26	158

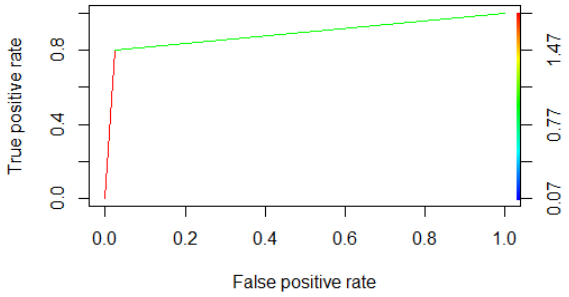
Confusion matrix on testing set		
	0	1
0	80	2
1	11	44

As well as the following roc curves.

In sample (AUC=0.911688)



Out of sample (AUC=0.8878049



Comparing all the logistic regression models as well as the classification trees, we obtain the following comparison table.

		In sample performance						Out of sample performance		
	Models	Residual deviance	AIC	BIC	Missclasification.rate	Cost	AUC	m.r	Cost	AUC
Logistic regression	Full model	81.23043	101.23	144.238	0.02201835	0.0293578	0.9963116	0.03649635	0.06569	0.9960089
	Best model(stepwise)	81.36279	97.3628	131.769	0.02201835	0.0293578	0.9963567	0.03649635	0.06569	0.9957871
Classification tree	Initial tree				0.04206501	0.0550459	0.9813471	0.04379562	0.07299	0.968847
	Pruned tree				0.06422018	0.2550459	0.9168825	0.09558824	0.41606	0.8878049

The most important conclusions that we get from this table is that the logistic regression models clearly outperform the classification trees ones. This can be seen in values such as the AUC or the cost.

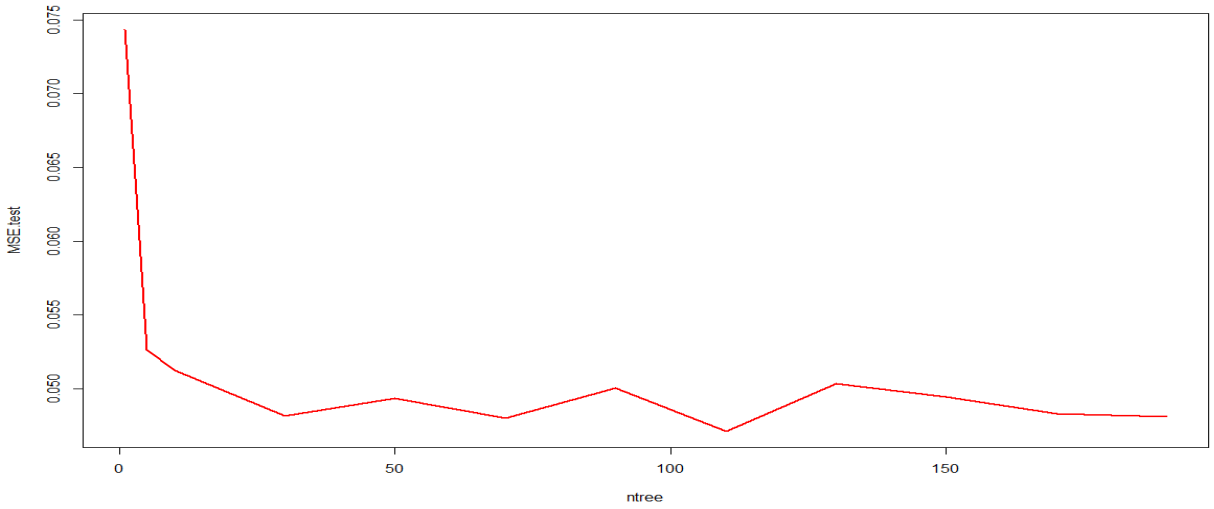
Also, the the best logistic regression model slightly outperforms the full one in AIC and BIC. However, the differences are so small that both models are excellent prediction models, and the selection of any of these two is perfectly justified.

Advanced Tree Models:

It would be useful to try out at least a couple different advanced classification techniques, so we'll start with bagging.

Bagging

First, we will get a plot to find the overall efficient number of bags that we would prefer in this case to minimize our MSE on the testing sample.



This graph shows that once we have about 30 trees we get diminishing and negligible returns to the minimization of MSE. However, with each time that I reran the bagging out-of-sample MSE, I got an MSE which was about 25% higher than the MSE from a single tree with most runs of my code, and the bagging MSE was never lower than the single tree MSE. For this reason, it seems that bagging is not very effective for this data.

Single Tree MSE	Bagging MSE
0.04017	0.04780526

OOB(Out of Bag)

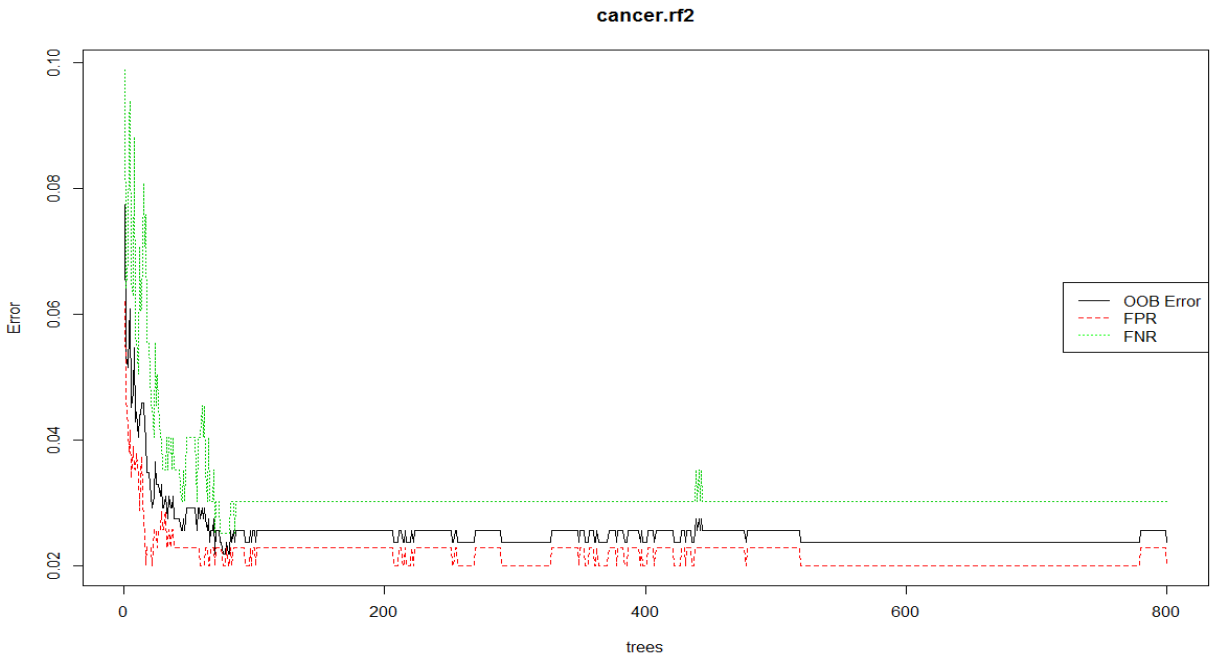
The out of bag method was also unsuccessful in the same fashion that the initial bagging methods were, regardless of the number of bags specified. The out of bag estimate for 150 bags was .2309, a rate much higher than the single tree MSE. By this, it is clear that the bagging methods are not adequate in this case, so we will move on to random forests.

Random Forests (Symmetric Cost)

By specifying a classification-based random forest with all variables other than ID included, we will immediately see much better performance in our model. This model does not take into account the possibility of an asymmetric cost for false positives and false negatives, but the performance is still impressive for classifying breast cancer. This symmetric random forest was run with the default parameters of 500 trees and 3 variables on each split, which provided the best results in the end, as seen below.

Symmetric Cost - Confusion Matrix on Training Set			
	0	1	FNR & FPR
0	341	7	0.0201
1	4	194	0.0202

In this case, the false negative and false positive rates are almost identical, leading to an overall misclassification rate of 2.01%. However, the false negatives are still typically bound to end up happening more often than false positives, as shown by this plot displaying the errors we attain from different numbers of trees



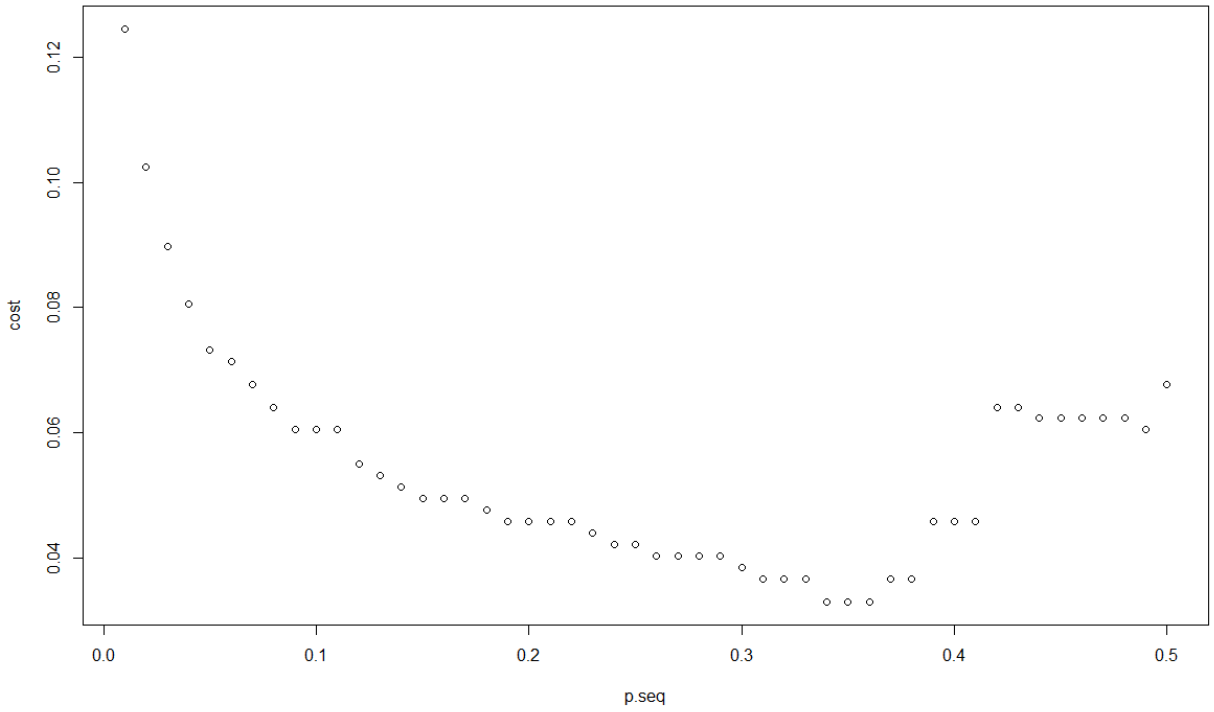
Discussion on Cost Allocation:

We would like to additionally make it so that false negatives are more costly than false positives. We have chosen to have a cost distribution of 5:1 for false negatives to false positives. We see this as adequate because certainly false negatives with cancer detection can imply the cost of life. Nonetheless, we do not slate false negatives to be the only costs we care about, given that

false positives can also lead to unnecessary interventions and worsen the quality of life/finances of the person diagnosed. False positives can still be tested further after mammographies, so we can be sure that we still want a false positive to be considered as being much more costly. Nonetheless, further research would help find to determine a true optimal distribution for false positives and false negatives in the case of breast cancer testing.

Random Forests with Pcut Values Optimized for Lowering Asymmetric Costs:

Once we specify our 5:1 cost distribution to find the minimal pcut value for such a cost measure, we can see expect optimized performance for limiting the number of false negatives.



This plot reveals that our optimal pcut is at .34, so we will use this as our cutoff for classification.

Asymmetric Cost - Confusion Matrix on Training Set			
	0	1	FNR & FPR
0	335	13	0.0018
1	1	197	0.0238

Asymmetric Cost - Confusion Matrix on Testing Set			
	0	1	FNR & FPR
0	91	5	0
1	0	41	0.0365

With this specification, the asymmetric cost confusion matrix shows an overall misclassification rate of 2.56% on the training set and a 3.649% MR on the testing set. This reveals that there is a cost for heavily preferring false positives over false negatives, but we find this cost to be appropriate. The number of false positives essentially doubles in order to reach a very low false negative rate. In this case, and indeed in all runs I've done on the testing set, there were zero false negatives on the testing set for the asymmetric cost model. These results are wonderful to hear; with 137 observations in the testing sample, there was not a single case where the cancer was not detected.

From this, it is rather clear that our asymmetric random forest model provides the most optimal output for unseen observations. The best logistic model is just behind it with one false negative and an asymmetric cost of 6.57%.The pruned and initial classification trees are also quite helpful, given that they still have misclassification rates below 10% and are very simple for doctors and patients to quickly interpret.It would be helpful to test these models on larger amounts of data to see if these classification rates remain consistent and to see where improvements could be made.