Data Mining Project 3

Ben Straub

April 18th, 2017

Introduction

Mining activity has long been associated with mining hazards, such as fires, floods, and toxic contaminants (Dozolme, P., 2016). Among these hazards, seismic hazards are the hardest to detect and predict (Sikora & Wróbel, 2010). Minimizing loss from seismic hazards requires advanced data collection and analysis. In recent years, more and more advanced seismic and seismoacoustic monitoring systems have come about. Still, the disproportionate number of low-energy versus high-energy seismic phenomena (e.g. $> 10^4$J) renders traditional analysis methods insufficient in making accurate predictions.

To investigate these seismic hazards and explore more advance analysis technique we used the seismic-bumps dataset provided by Sikora & Wróbel (2010), found in the UCI Machine Learning Repository. This seismic-bumps dataset comes from a coal mine located in Poland and contains 2584 observations of 19 attributes. Each observation summarizes seismic activity in the rock mass within one 8-hour shift. Note that the decision attribute, named "class", has values 1 and 0. This variable is the response variable we use in this project. A class value of "1" is categorized as "hazardous state", which essentially indicates a registered seismic bump with high energy ($>10^4$J) in the next shift. A class value "0" represents non-hazardous state in the next shift. Table 1 in the Appendix has a listing of all 18 variables and their descriptions.

The purpose of this project is to find whether and how the other 18 variables can be used to determine the hazard status of the mine. In project 2, we utlized techniques such as the indicator matrix linear regression, logistic regression, linear discriminant analysis(LDA), quadratic discriminant qnalysis (QDA), and regularized discriminant analysis (RDA) to try and find a model that would accurately predict the hazardous state. Unfortunately, all of the five project two methods performed poorly. We felt that there were two major issues at hand for this poor performance of the five methods. First, the low incidences of "1's" in the response variable class, which indicates a hazardous state in the mine. Only 170 "1's" for class out of 2584 were observed. A difficult problem for traditional method of analyses. The second issue was multicollinearity. Regression diagnostics indicate that the data, in general, meet most assumptions. However, we see that that data are somewhat skewed right, and there is severe multicollinearity (VIF > 10) between some of the covariates. Table 2 in the Appendix contains VIF's for the linear regression model.

Multicollinearity can be address by dimension reduction techniques such as PCA, step-wise regression, LASSO or ridge. In project 2, we utilized step-wise regression and LASSO to arrive at two candidate models. However, even with these dimension reduction techniques our models still performed poorly. Hopefully, to remedy this poor performance, we can utilize more advance techniques such as Boosting, Random Forest or Support Vector Machines. We only look at the model that was obtained through step-wise regression.

In section 2, we report ROC curves and missclassification rates for Logistic Regression, LDA, QDA and RDA. In section 3, we report **best technique** out of the three that we tried. In section 4, we provide concluding remarks as well as future work on seismic data.

2 Logistic Regression, LDA, QDA, RDA

2.1 Logistic Regression-Full and Step

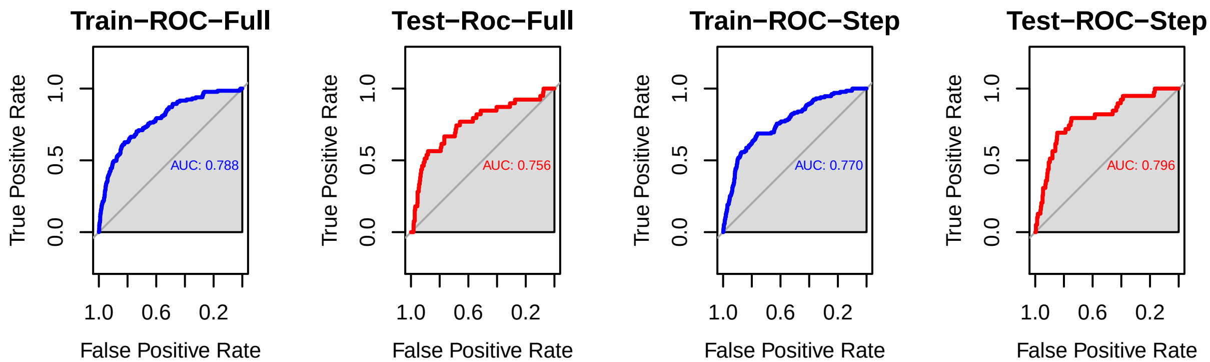


Table 1: Logistic Regression

	Full	Step
Computing Time	0.121	0.082
Train Error Rates	0.067	0.070
Test Error Rates	0.065	0.062

Multicollinearity can be address by dimension reduction techniques such as PCA, step-wise regression, LASSO or ridge. In project 2, we utilized step-wise regression and LASSO to arrive at two candidate models. However, even with these dimension reduction techniques our models still performed poorly. Hopefully, to remedy this poor performance, we can utilize more advance techniques such as Boosting, Random Forest or Support Vector Machines. We only look at the model that was obtained through step-wise regression.

In section 2, we report ROC curves and missclassification rates for Logistic Regression, LDA, QDA and RDA. In section 3, we report **best technique** out of the three that we tried. In section 4, we provide concluding remarks as well as future work on seismic data.

2.2 Linear Discriminant Analysis

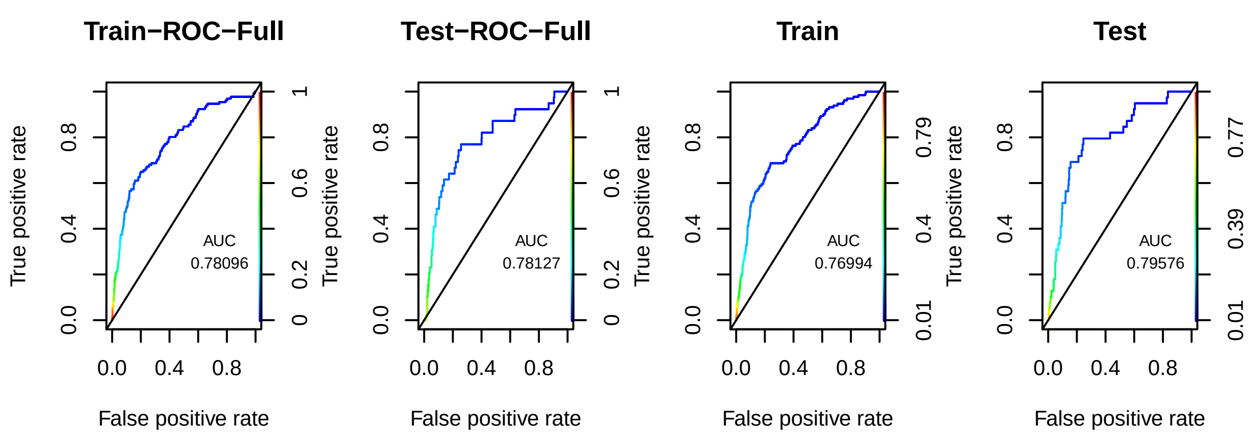


Table 2: Linear Discriminant Analysis

	Full	Step
Computing Time	0.461	1.340
Train Error Rates	0.074	0.081
Test Error Rates	0.077	0.076

Multicollinearity can be address by dimension reduction techniques such as PCA, step-wise regression, LASSO or ridge. In project 2, we utilized step-wise regression and LASSO to arrive at two candidate models. However, even with these dimension reduction techniques our models still performed poorly. Hopefully, to remedy this poor performance, we can utilize more advance techniques such as Boosting, Random Forest or Support Vector Machines. We only look at the model that was obtained through step-wise regression.

In section 2, we report ROC curves and missclassification rates for Logistic Regression, LDA, QDA and RDA. In section 3, we report **best technique** out of the three that we tried. In section 4, we provide concluding remarks as well as future work on seismic data.

2.3 Quadratic Discriminant Analysis

Full Model

Full Model not able to handle the multicollinearity of the data.

Quadratic Discriminant Analysis - Step

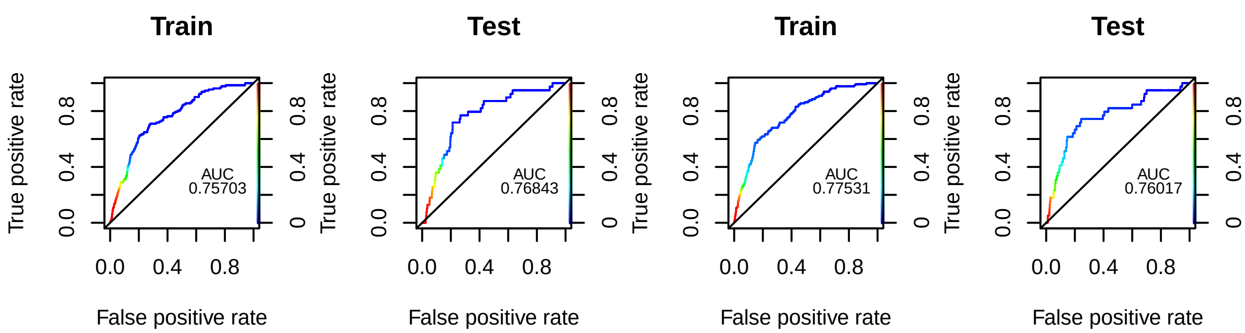


Table 3: Quadratic Discriminant Analysis

	Full	Step	Lasso
Computing Time	NA	0.431	1.28
Train Error Rates	0.149	0.109	NA
Test Error Rates	0.159	0.107	NA

Multicollinearity can be address by dimension reduction techniques such as PCA, step-wise regression, LASSO or ridge. In project 2, we utilized step-wise regression and LASSO to arrive at two candidate models. However, even with these dimension reduction techniques our models still performed poorly. Hopefully, to remedy this poor performance, we can utilize more advance techniques such as Boosting, Random Forest or Support Vector Machines. We only look at the model that was obtained through step-wise regression.

In section 2, we report ROC curves and missclassification rates for Logistic Regression, LDA, QDA and RDA.

In section 3, we report **best technique** out of the three that we tried. In section 4, we provide concluding remarks as well as future work on seismic data.

2.4 Regularized Discriminant Analysis

Regularized Discriminant Analysis -Full Regularized Discriminant Analysis -Step

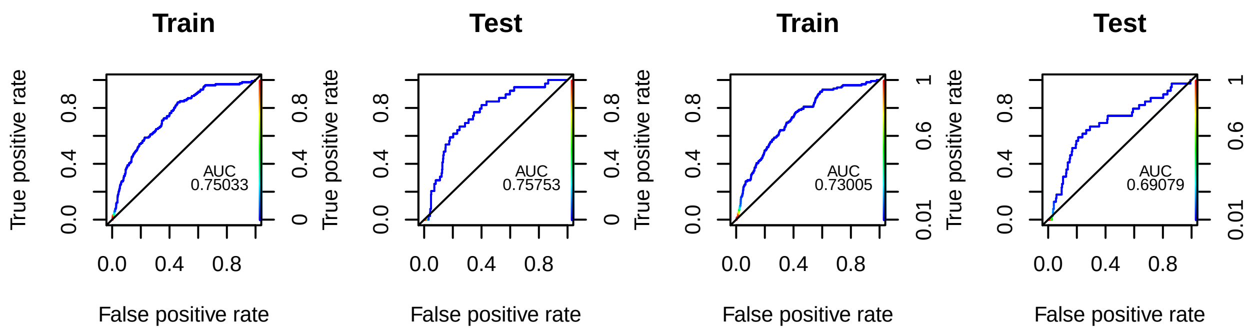


Table 4: Regularized Discriminant Analysis

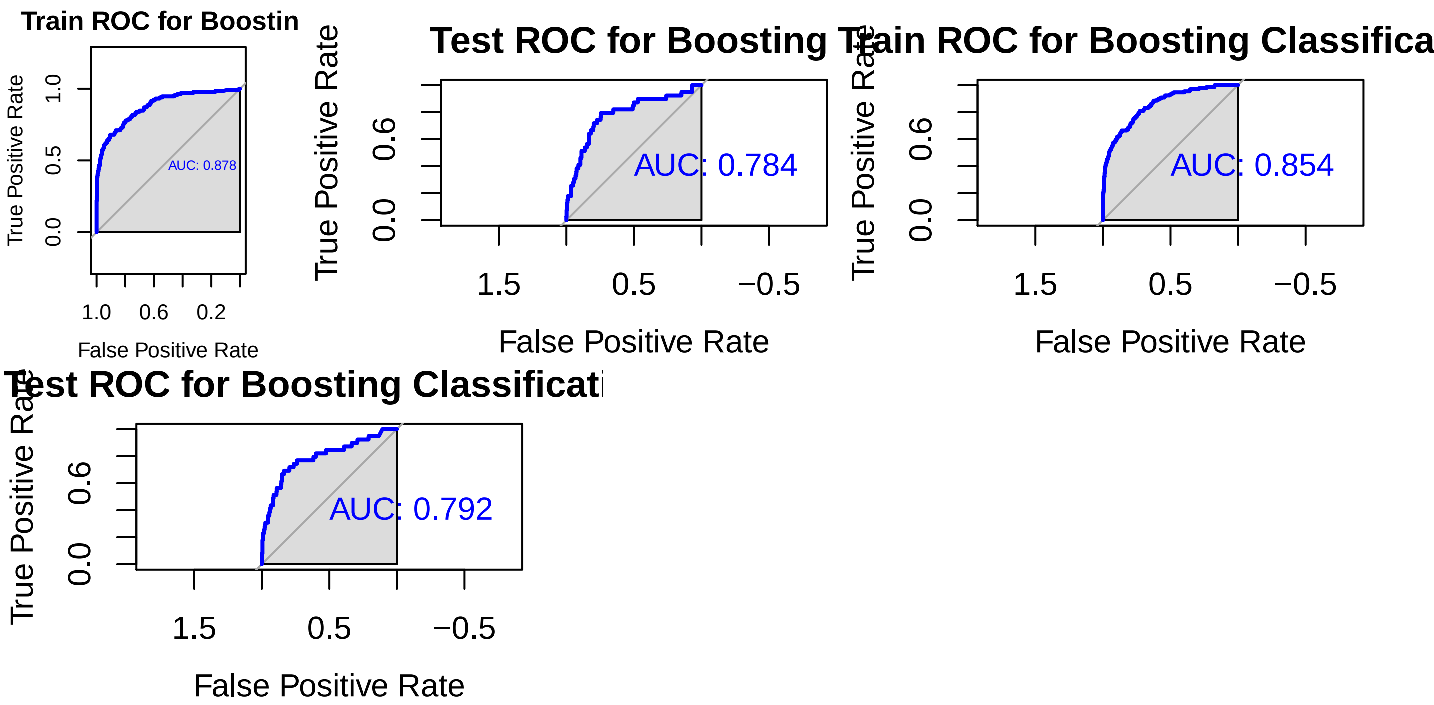
	Full	Step
Computing Time	3.423	2.041
Train Error Rates	0.076	0.082
Test Error Rates	0.082	0.085

Multicollinearity can be address by dimension reduction techniques such as PCA, step-wise regression, LASSO or ridge. In project 2, we utilized step-wise regression and LASSO to arrive at two candidate models. However, even with these dimension reduction techniques our models still performed poorly. Hopefully, to remedy this poor performance, we can utilize more advance techniques such as Boosting, Random Forest or Support Vector Machines. We only look at the model that was obtained through step-wise regression.

In section 2, we report ROC curves and missclassification rates for Logistic Regression, LDA, QDA and RDA. In section 3, we report **best technique** out of the three that we tried. In section 4, we provide concluding remarks as well as future work on seismic data.

Boosting before variable selection

Next we performed boosting to our dataset and see whether it brings improvement compared to previous methods. Boosting involves combining a large number of decision trees. In boosting, we slowly grow the tree according to residuals from the model. The construction of each tree depends strongly on the trees that have already been grown. (James et al., 2013)

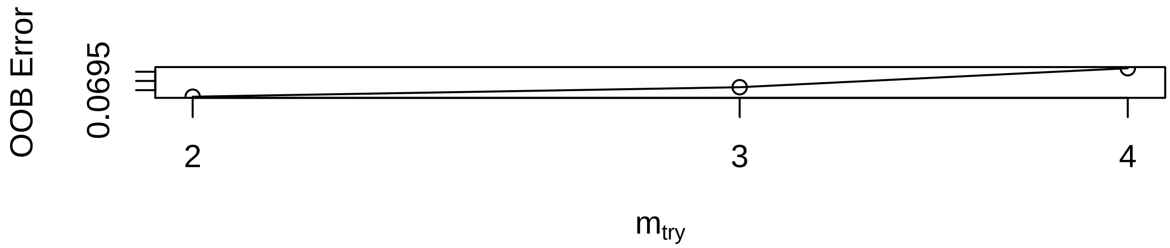


Boosting			
model	orignal	Model1	Model2
time	10.38	5.64	4.92
misclassification rate	.057	.060	.060

Random Forests Classification

Next, we use Random Forest classification method as it yields relatively better classifcation results among all tree-based methods. As opposed to growing silngle decision tree (as in CART), random forest grows multiple trees, with having each split to consider only a subset of all predictors.Then it takes average of all trees to make final tree. In this way, random forest can reduce amount of potetial correlation between trees and thereby help reduce the variance of the final tree. First, we used tuneRF function to find the optimal numbers of variables to try (mtry) splitting on at each node. We found mtry = 2 produces least out of the box (OBB) error, that means, 2 out of 15 predictors should be considered for each split.

```
mtry = 3  OOB error = 6.97%
Searching left ...
mtry = 2    OOB error = 6.91%
0.007407407 0.01
Searching right ...
mtry = 4    OOB error = 7.07%
-0.01481481 0.01
```



Then, we applied Random Forest formula on both train and test datasets for the models derived before and after variable selection. In each cases, the number of tress we used is 1000. We also calculated the 'variable importance' in order to see relative importance of each variable in the classifcation process.

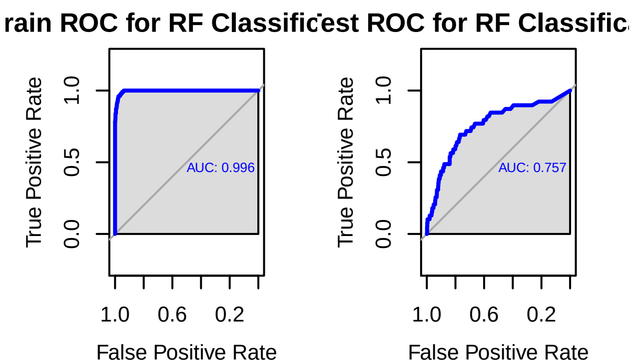
RF Classification BEFORE Variable Selection

Here we performed random forest classification on training and test datasets individually, using mtry = 2 and ntree = 1000. We found slightly lower test missclassification rate (5.7%) than train's (9.4%).

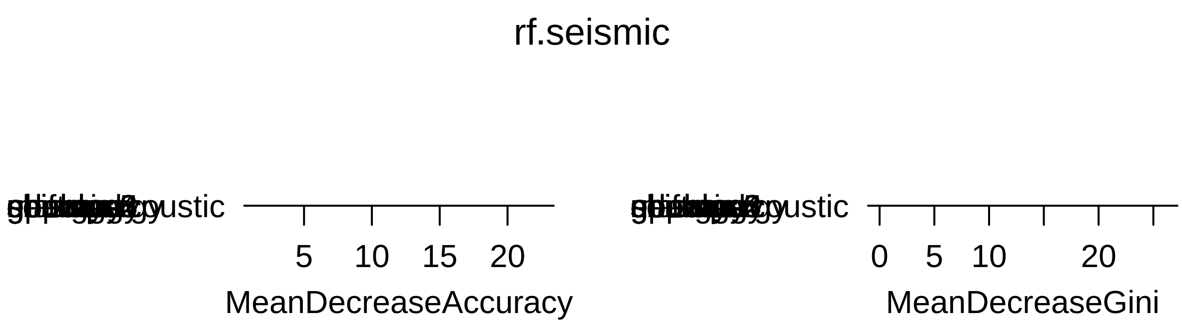
However, ROC curves show that predictions on test dataset are leass accurate than predictions on test dataset.

```
[1] 0.094
```

```
[1] 0.057
```



We see from the Variable importance plot, that the most important variables are nbumps2, nbumps3, genergy, nbumps4, nbumps, maxenergy, gdenergy, gpuls, and energy. Variables like shift, ghazard, nbumps5 and seismoacoustic are of less important for predicting seismic events.



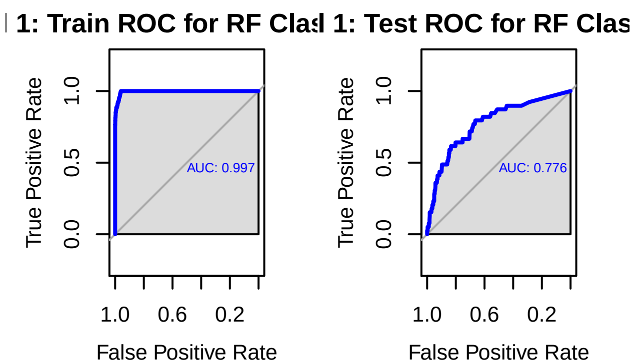
RF Classification AFTER Variable Selection

Here we performed random forest classifcations on resulting models (e.g., Model 1 and Model 2) from variable slection procedures. Model 1 = genergy + gpuls + nbumps + nbumps2 + nbumps4 Model 2 = seismic + shift + gpuls + nbumps Missclassificaion rates for Model 1 training and test datasets are 6.9% and 5.4%, respectively. This time the ROC curves revealed slightly improved test AUC.

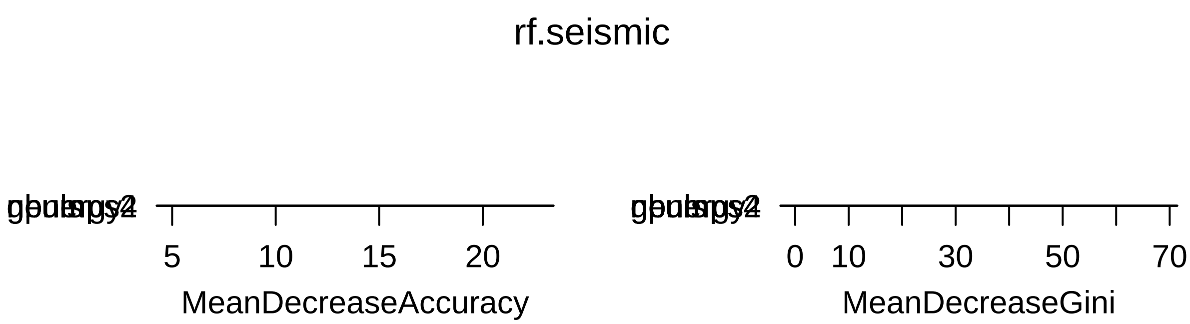
```
[1] 0.069
```

```
[1] 0.054
```

```
[1] 0.2724458
```



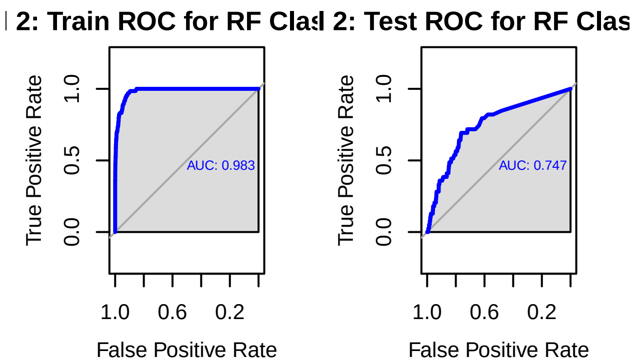
According to the variable importance plot, most important varibales for predicting next seismic events seem to be nbumps and nbumps4.



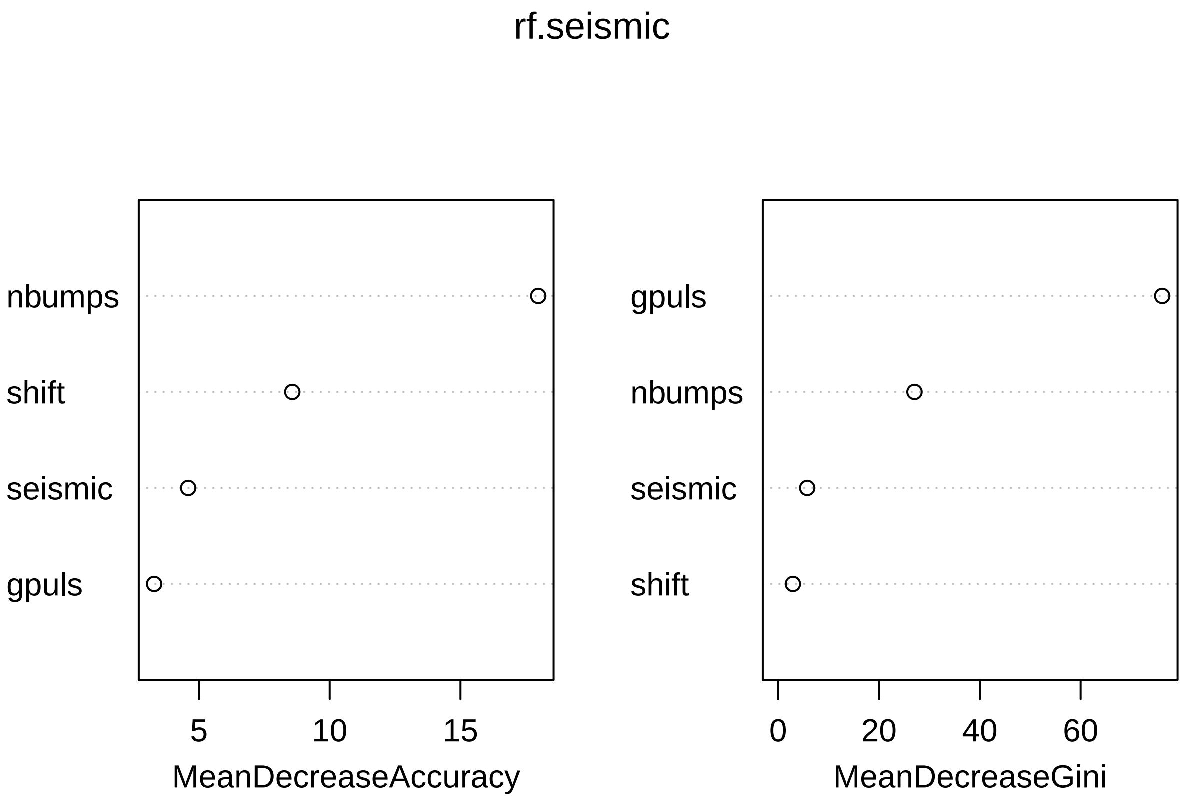
We can see from ROC curves, that train AUC is still slightly higher than test AUC. However, Model 1's test AUC (0.747) is slightly lower than the test AUC (0.776) from Model 1.

```
[1] 0.058
```

```
[1] 0.063
```



In this case, only nbumps seems to be imoportant varibale for predicting seismic events in the next shift.



At this point, it is clear that across all of the trees considered in the random forests so far, number of bumps in the previous shifts are by far the most important varibales for predicting potential seismic events in future mining shift(s).

Summary Table of Random Forest (RF) Classification:

RF on Models	Missclassification Rate	AUC	Important Variabl
Full Model (Train)	9.4%	0.996	nbumps2,3,4 , genergy, nbumps, maxenergy, gdenergy, gpuls, and ene
Full Model (Test)	5.7%	0.757	nbumps2,3,4 , genergy, nbumps, maxenergy, gdenergy, gpuls, and ene
Model 1 (Train)	6.9%	0.997	nbumps and nbum
Model 1 (Test)	5.4%	0.776	nbumps and nbum
Model 2 (Train)	5.8%	0.983	nbur
Model 2 (Test)	6.3%	0.747	nbur

Time elasped: 7.09

Support vector classifier and support vector machine

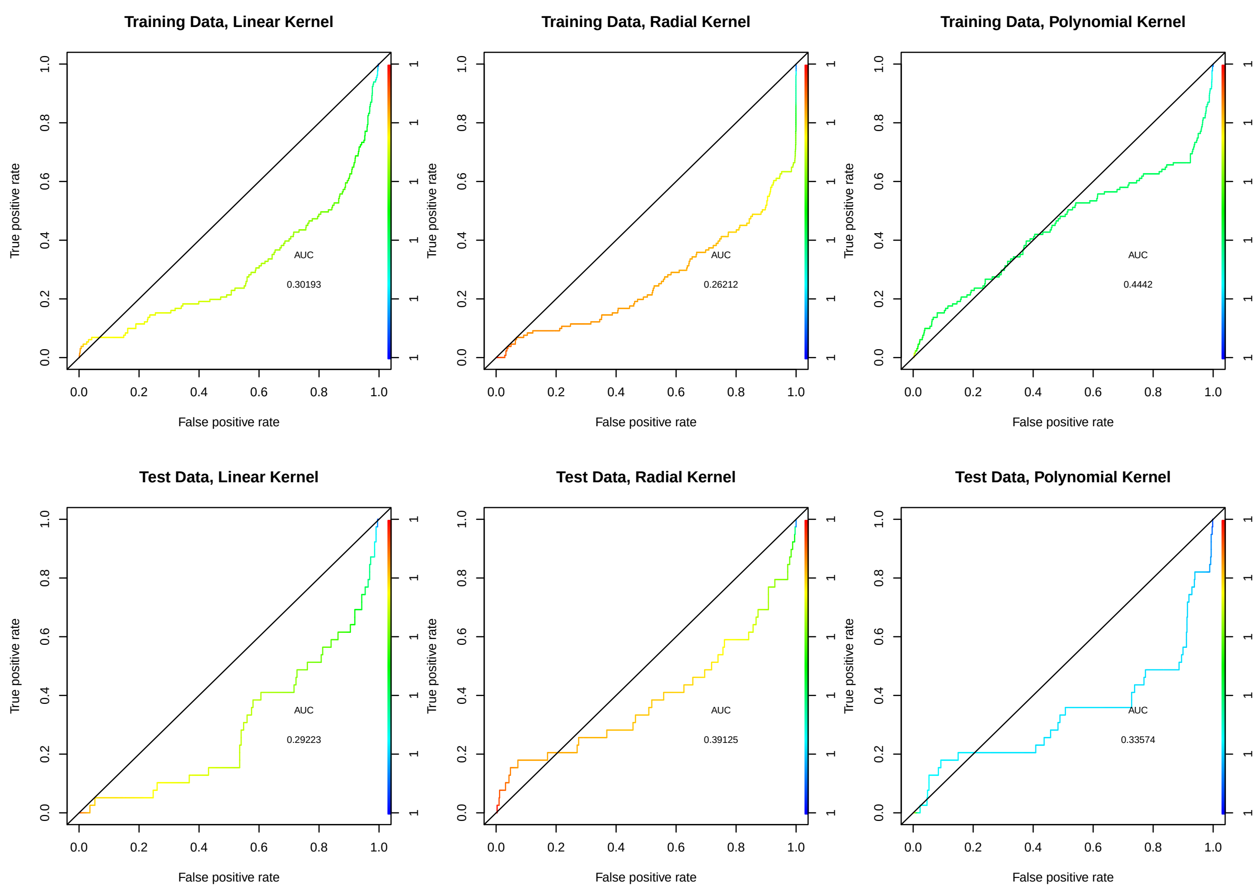
A support vector machine allows for the classification of data with, if desired, non-linear boundaries. We attempted to fit this approach to our data, using a linear, radial, and polynomial kernel. With an increased cost (a penalty term), the time to fit the prescribed SVM increased. Thankfully, our data favored lower costs which meant quicker computation in the end. For each kernel, we searched over a list of possible inputs for cost, gamma (where applicable), and degree (where applicable). Our final results are presented below.

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
  cost
 0.001

- best performance: 0.06761391

- Detailed performance results:
   cost      error dispersion
1 0.001 0.06761391 0.02215124
2 0.010 0.06761391 0.02215124
3 0.100 0.06761391 0.02215124
4 1.000 0.06761391 0.02215124
5 5.000 0.06761391 0.02215124



Call:
best.tune(method = svm, train.x = factor(class) ~ genergy + gpuls +
    nbumps + nbumps2 + nbumps4, data = seismic[train, ], ranges = list(cost = c(0.001,
    0.01, 0.1, 1, 5)), kernel = "linear")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  0.001
      gamma:  0.2

Number of Support Vectors:  268

 ( 137 131 )


Number of Classes:  2

Levels:
 0 1
```

```
        truth
predict   0   1
      0 607  39
      1   0   0
```

```
        truth
predict   0   1
      0 607  39
      1   0   0
```

```
        truth
predict   0   1
      0 607  39
      1   0   0
```



Stepwise Model			
Kernel	linear	radial	polynomial
cost	.001	.001	.001
gamma	.2	1	.2
degree	N/A	N/A	2
time	5.4	26.47	25.31
misclassification rate	.06	.06	.06

Appendix

Table I. Attribute information of the seismic-bumps dataset

Table 8: Table II-VIFs of Linea

Data Attributes	Description
seismic	result of shift seismic hazard assessment: ‘a’ - lack of hazard, ‘b’ - low hazard, ‘c’ - high hazard, ‘d
seismoacoustic	result of shift seismic hazard assessment
shift	type of a shift: ‘W’ - coal-getting, ‘N’ - preparation shift
genergy	seismic energy recorded within previous shift by active geophones (GMax) monitoring the longwal
gpuls	number of pulses recorded within previous shift by GMax
gdenergy	deviation of recorded energy within previous shift from average energy recorded during eight prev
gdpuls	deviation of recorded pulses within previous shift from average number of pulses recorded during e
ghazard	result of shift seismic hazard assessment by the seismoacoustic method based on registration comi
nbumps	the number of seismic bumps recorded within previous shift
nbumps$i$, $i \in \{1, \ldots, 5\}$	the number of seismic bumps ($10^i - 10^{i+1}$ J) registered within previous shift
energy	total energy of seismic bumps registered within previous shift
maxenergy	maximum energy of the seismic bumps registered within previous shift
class	the decision attribute: ‘1’ - high energy seismic bump occurred in the next shift (‘hazardous state

Table 9: Table II-VIFs of Linear Model

seismic	seismoacoustic	shift	genergy	gpuls	gdenergy	gdpuls
1.21	1.29	1.41	2.89	4.06	3	3.43

ghazard	nbumps	nbumps2	nbumps3	nbumps4	nbumps5	energy	maxenergy
1.4	2414.69	798.96	769.13	104.4	11.56	110.28	93.76