

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. $> 10^4\text{J}$) renders traditional analysis methods insufficient.

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 ($>10^4\text{J}$) 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.

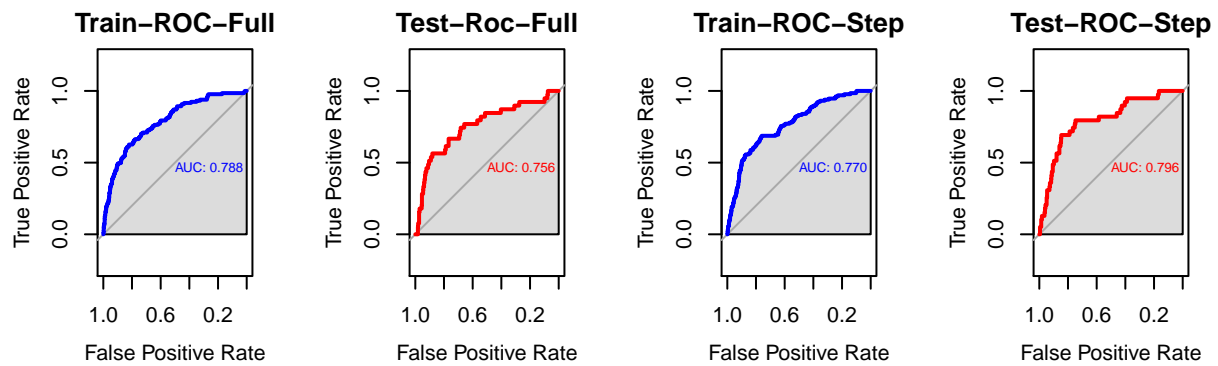
The purpose of this project is to find whether and how the other 18 variables can be used to determine the hazard status of the mine. In project 2, we utilized techniques such as the indicator matrix linear regression, logistic regression, linear discriminant analysis(LDA), quadratic discriminant analysis (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 ($\text{VIF} > 10$) 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.

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 Model Logistic Regression - Step Model



Logistic Regression - Lasso Model

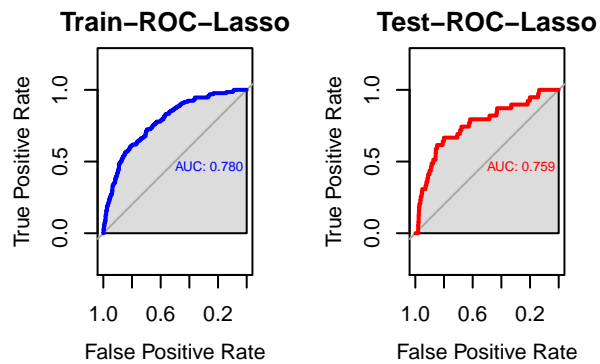
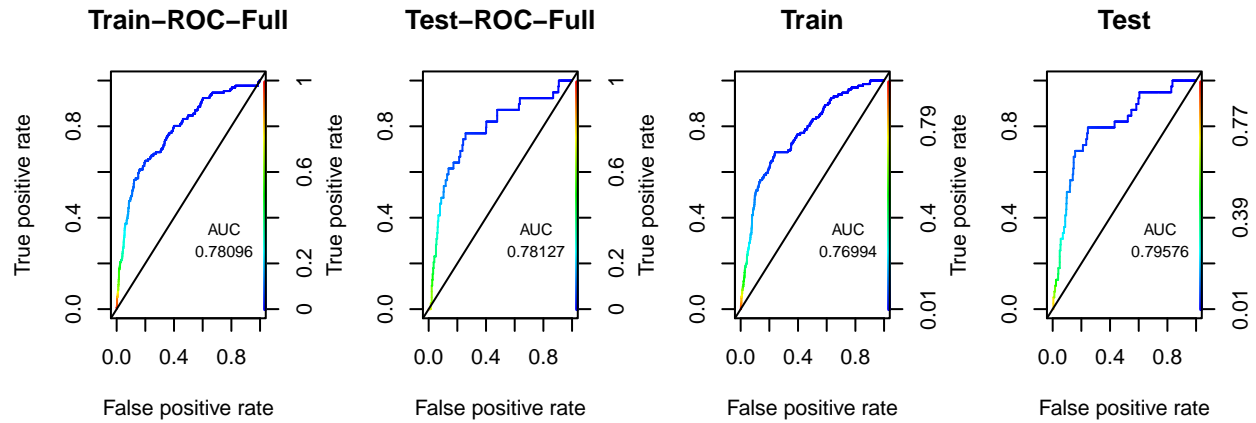


Table 1: Logistic Regression

	Full	Step	Lasso
Computing Time	0.116	0.076	0.075
Train Error Rates	0.067	0.070	0.068
Test Error Rates	0.065	0.062	0.062

2.2 Linear Discriminant Analysis



Linear Discriminant Analysis - Lasso

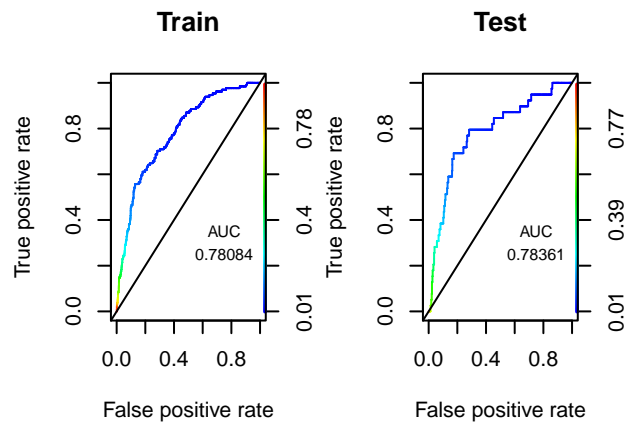


Table 2: Linear Discriminant Analysis

	Full	Step	Lasso
Computing Time	0.476	1.378	0.418
Train Error Rates	0.074	0.081	0.077
Test Error Rates	0.077	0.076	0.076

2.3 Quadratic Discriminant Analysis

Full Model

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

Quadratic Discriminant Analysis - Step

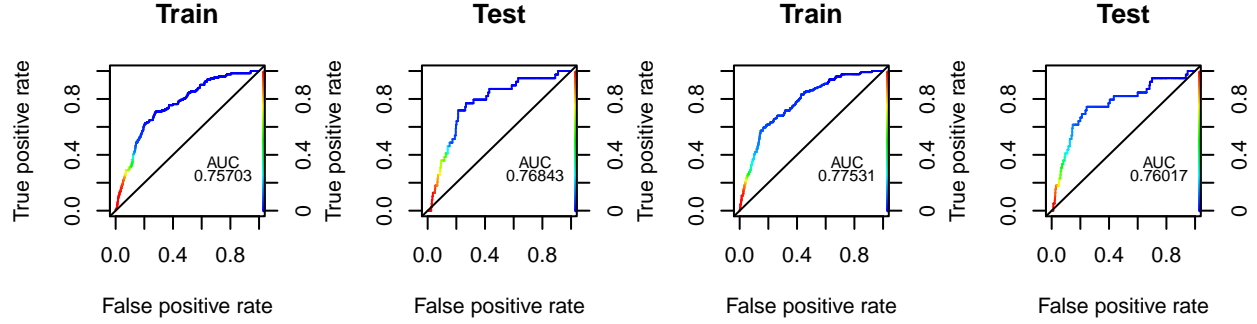
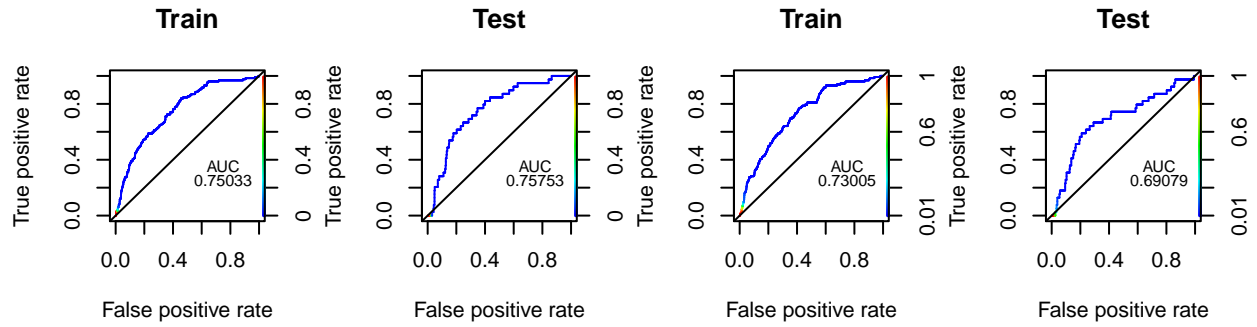


Table 3: Quadratic Discriminant Analysis

	Full	Step	Lasso
Computing Time	NA	1.165	0.4179999999999999
Train Error Rates	NA	0.109	0.077
Test Error Rates	NA	0.107	0.076

2.4 Regularized Discriminant Analysis

Regularized Discriminant Analysis -Full Regularized Discriminant Analysis -Step



Regularized Discriminant Analysis - Lasso

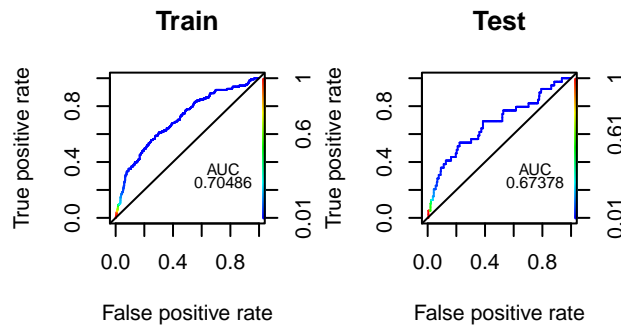
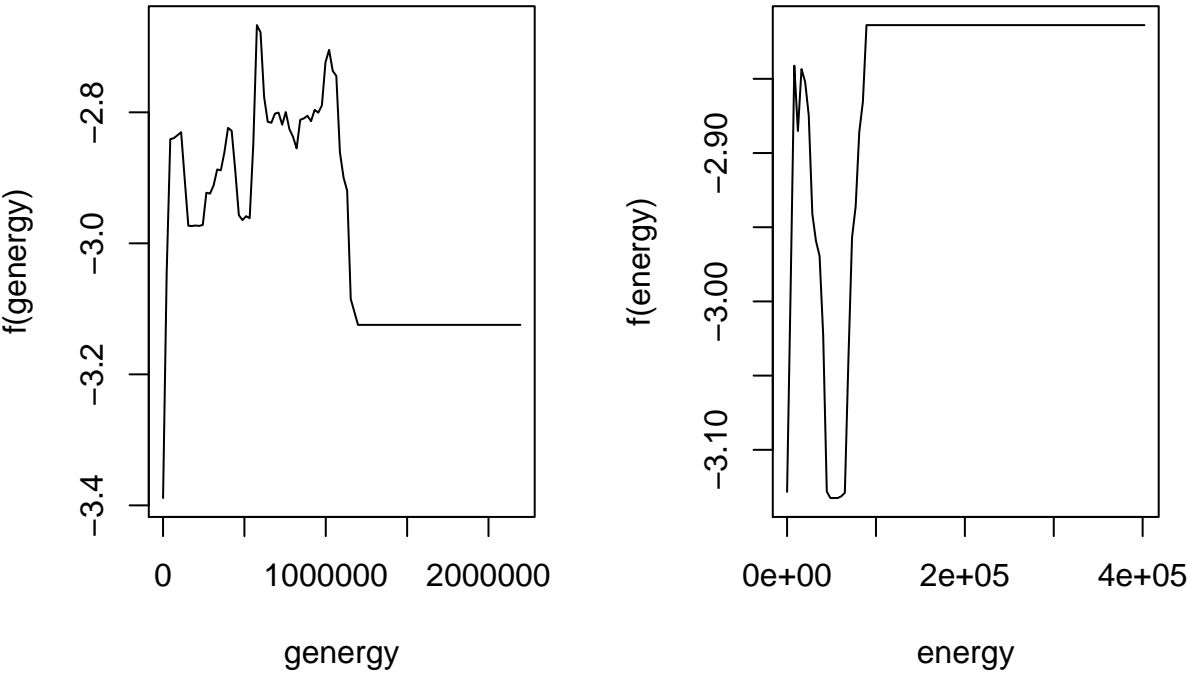


Table 4: Regularized Discriminant Analysis

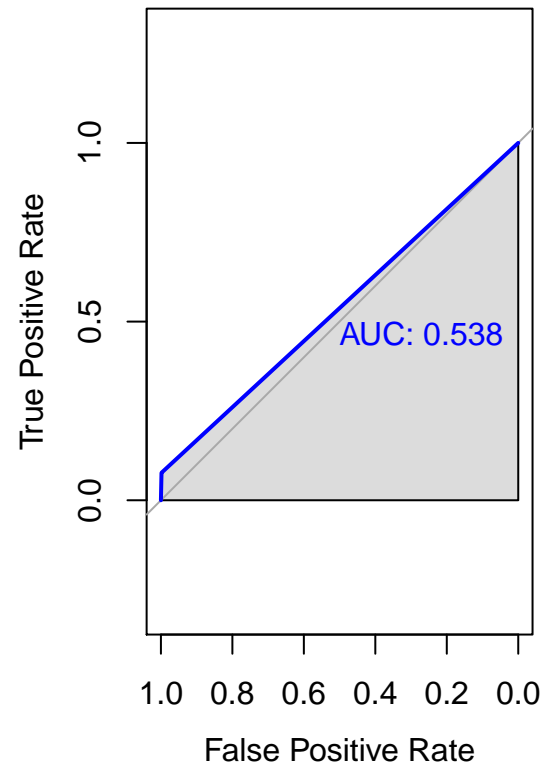
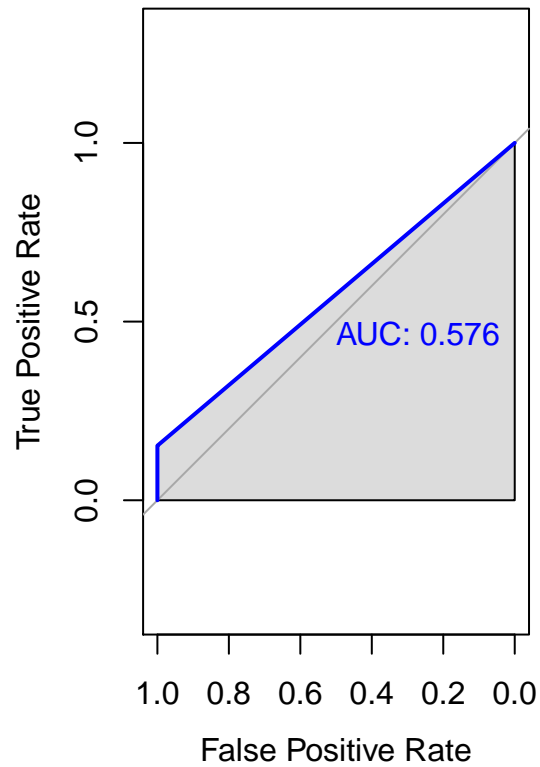
	Full	Step	Lasso
Computing Time	2.994	2.078	1.308
Train Error Rates	0.076	0.082	0.077
Test Error Rates	0.082	0.085	0.074

Boosting before variable selection

elapsed
7.851



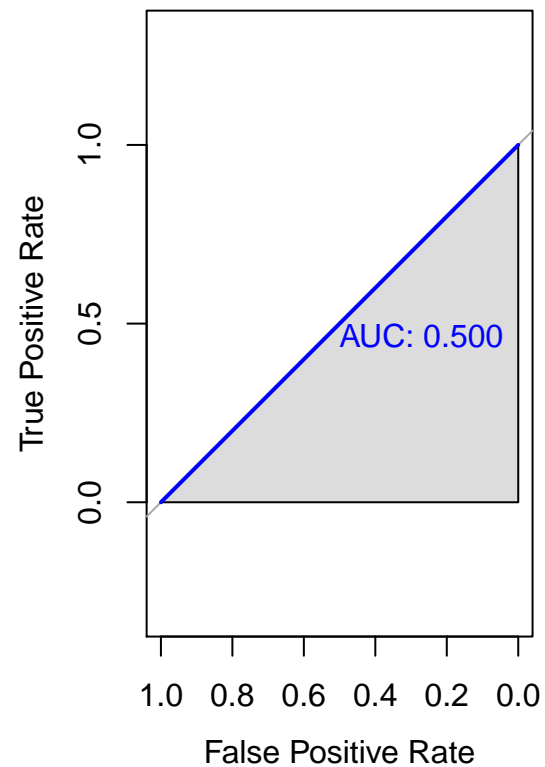
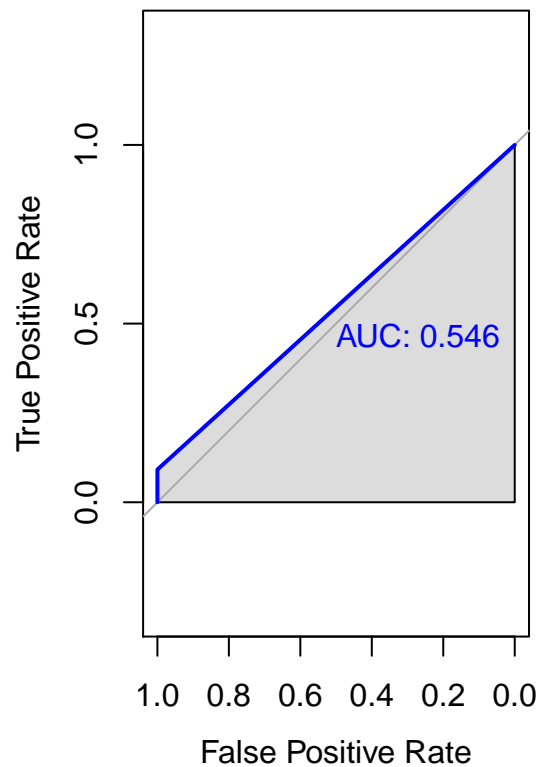
Test ROC for Boosting Classification



Boosting after variable selection

elapsed
3.483

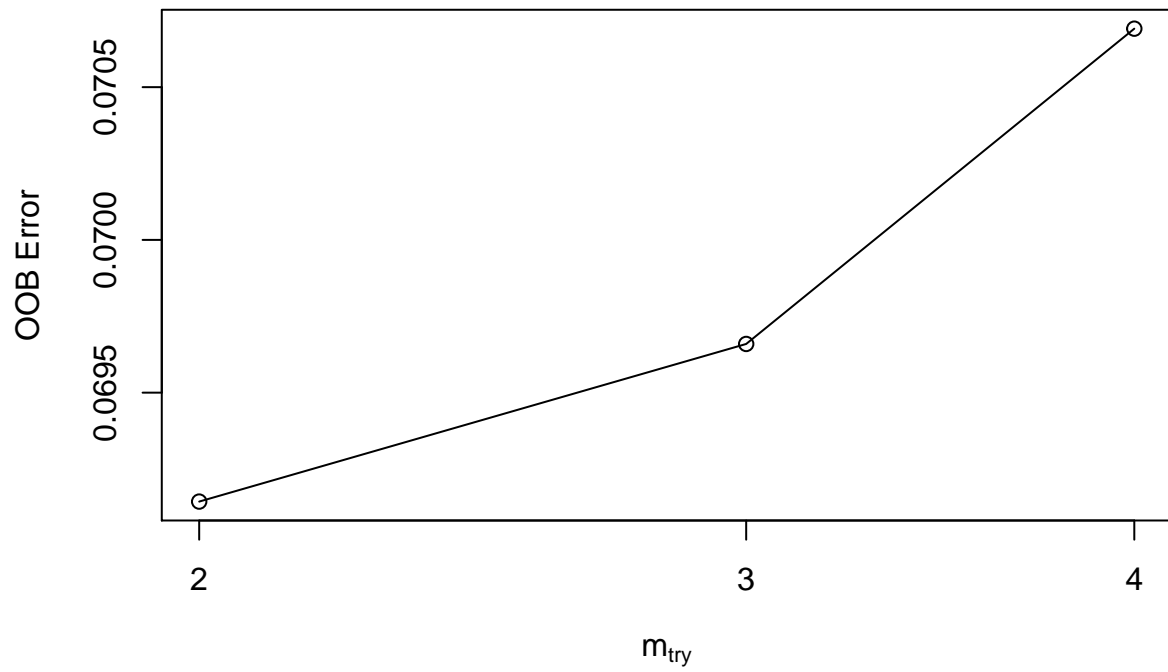
Test ROC for Boosting Classification



Random Forests Classification

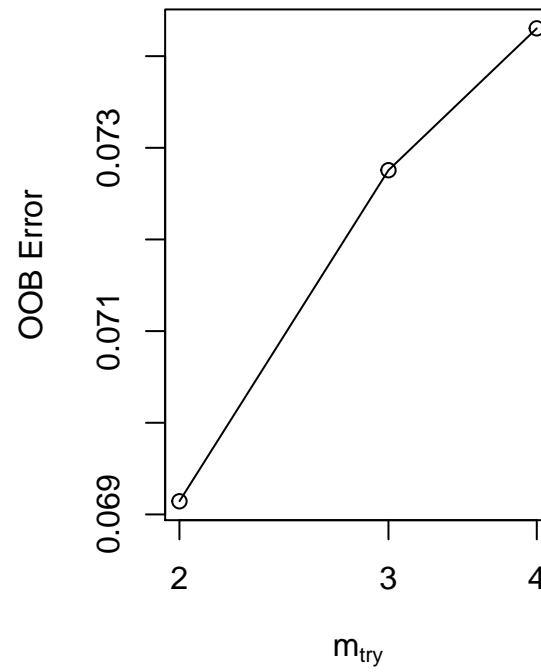
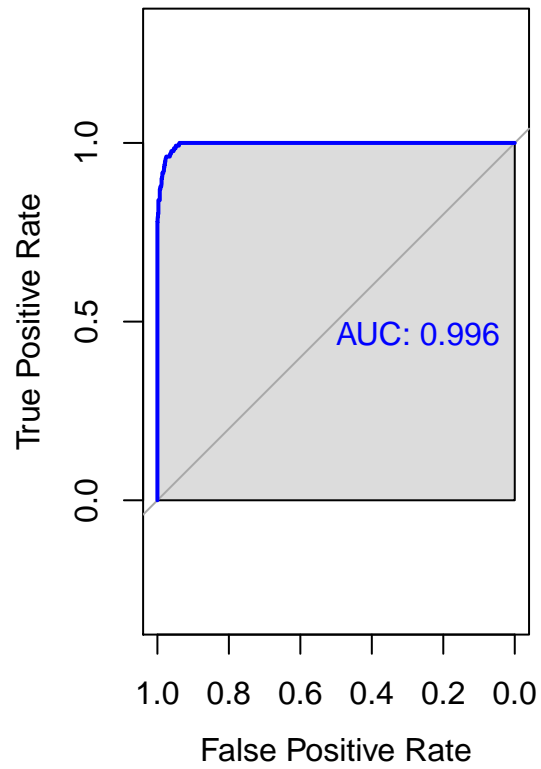
RF Classification BEFORE Variable Selection

```
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
```



```
mtry = 3  OOB error = 7.28%  
Searching left ...  
mtry = 2    OOB error = 6.91%  
0.04964539 0.01  
Searching right ...  
mtry = 4    OOB error = 7.43%  
-0.07462687 0.01
```


Train ROC for RF Classification

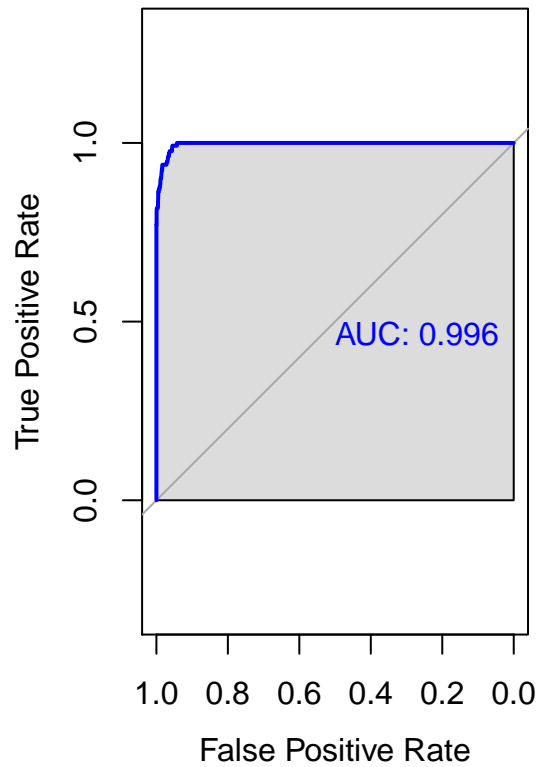


[1] 0

[1] 0.2363033

[1] 0.2203302

Train ROC for RF Classification



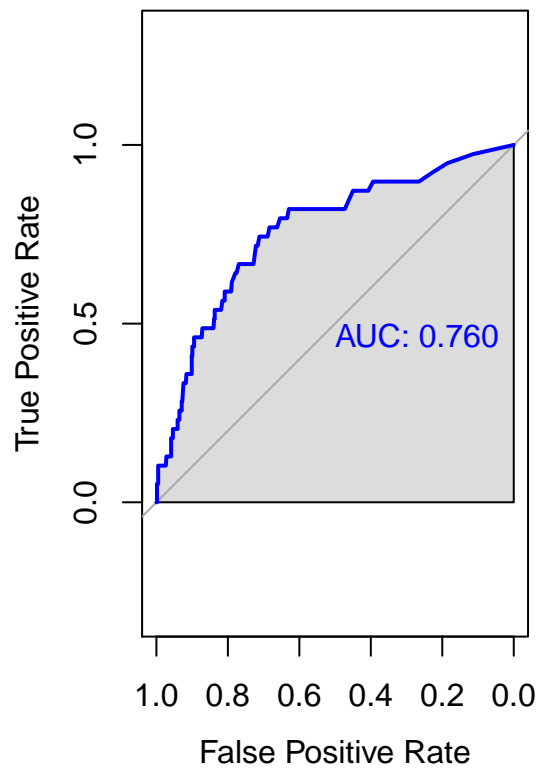
[1] 0.02564103

[1] 0.1136738

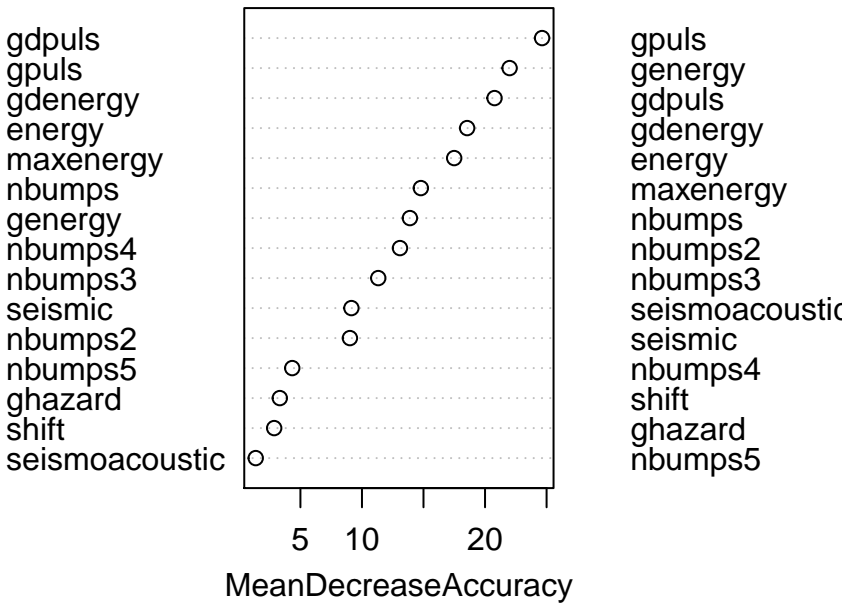
[1] 0.1083591

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
seismic	7.639190	5.1409941	9.142981	4.2665367
seismoacoustic	1.581907	-0.2409790	1.368947	4.5015275
shift	2.486616	0.7266865	2.869017	2.4176743
genergy	12.086539	2.4284640	13.895572	25.1355203
gpuls	18.476810	13.6828191	21.994591	26.7512211
gdenergy	22.120246	-8.0739569	20.771536	20.7055737
gdpuls	25.688347	-7.5341551	24.634248	20.8810289
ghazard	4.587309	-2.7149240	3.327211	1.9414849
nbumps	13.977076	5.3373298	14.784089	11.5360046
nbumps2	6.668420	8.5245738	9.021708	8.5027181
nbumps3	9.531100	5.9025441	11.324696	7.3784317
nbumps4	14.878958	-10.0707066	13.088679	2.7869821
nbumps5	4.832149	-2.6126517	4.336337	0.3214691
energy	17.725076	-1.2777291	18.544822	18.4047305
maxenergy	17.086692	-5.1894493	17.493649	13.4764157

Test ROC for RF Classification



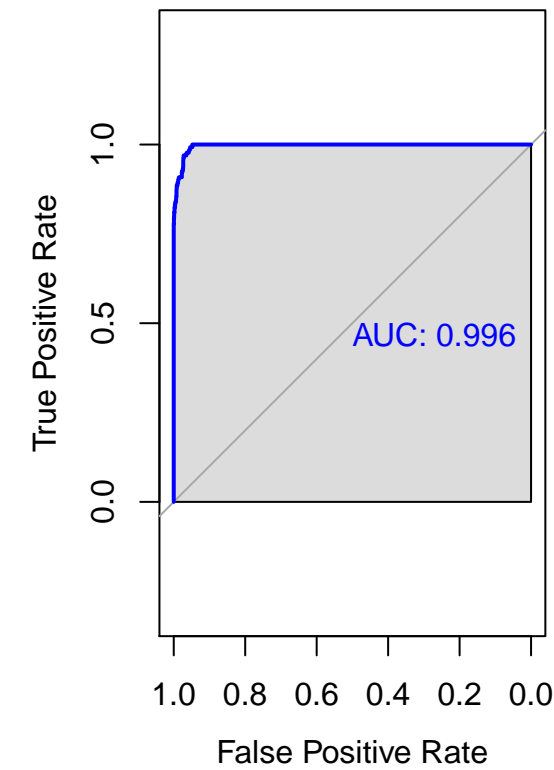
rf.seismic



RF Classification AFTER Variable Selection

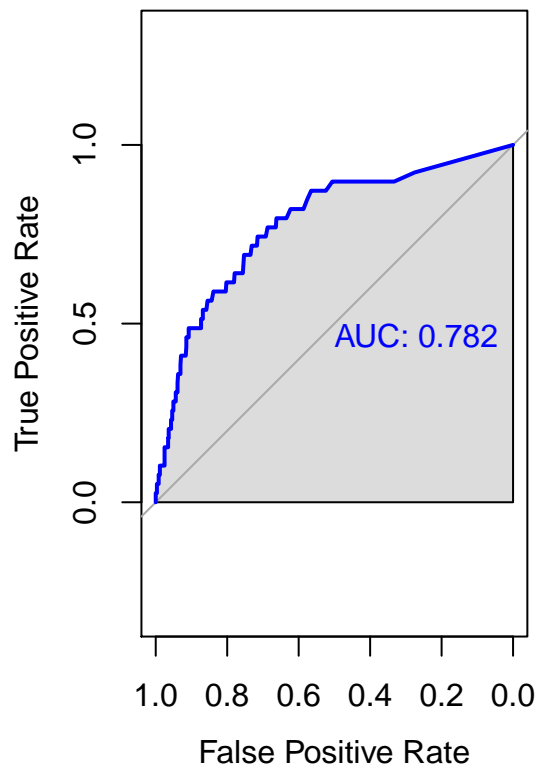
RF Classification AFTER Variable Selection

Model 1: Train ROC for RF Classific:

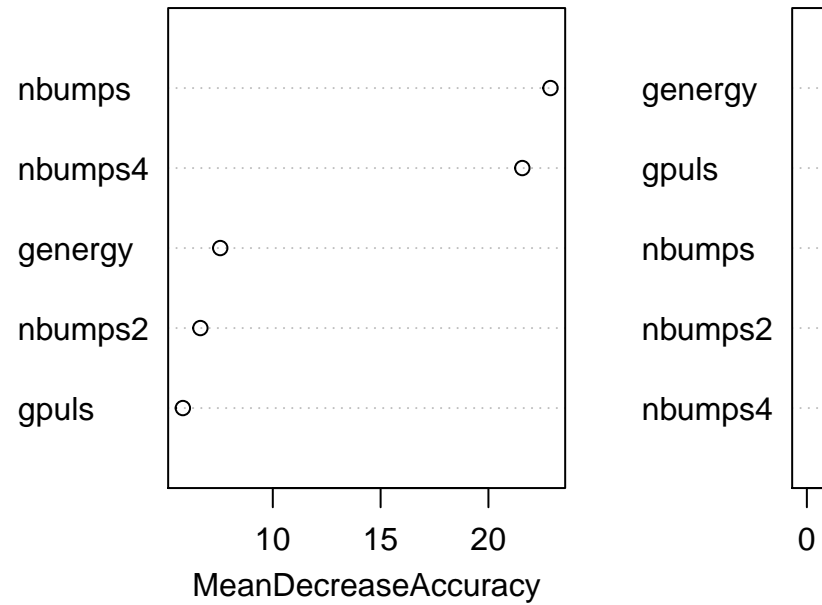


	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
genergy	5.1300375	5.586820	7.563177	68.556399
gpuls	0.0700258	19.539698	5.829128	66.551265
nbumps	14.9000236	32.868947	22.877694	26.059296
nbumps2	3.0991937	9.011757	6.641027	14.385561
nbumps4	24.1165725	-8.627134	21.572093	7.176123

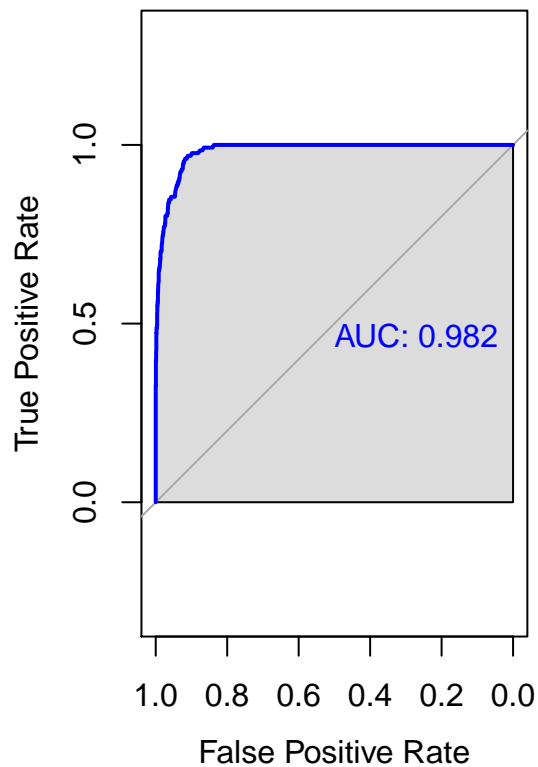
Model 1: Test ROC for RF Classification



rf.seismic



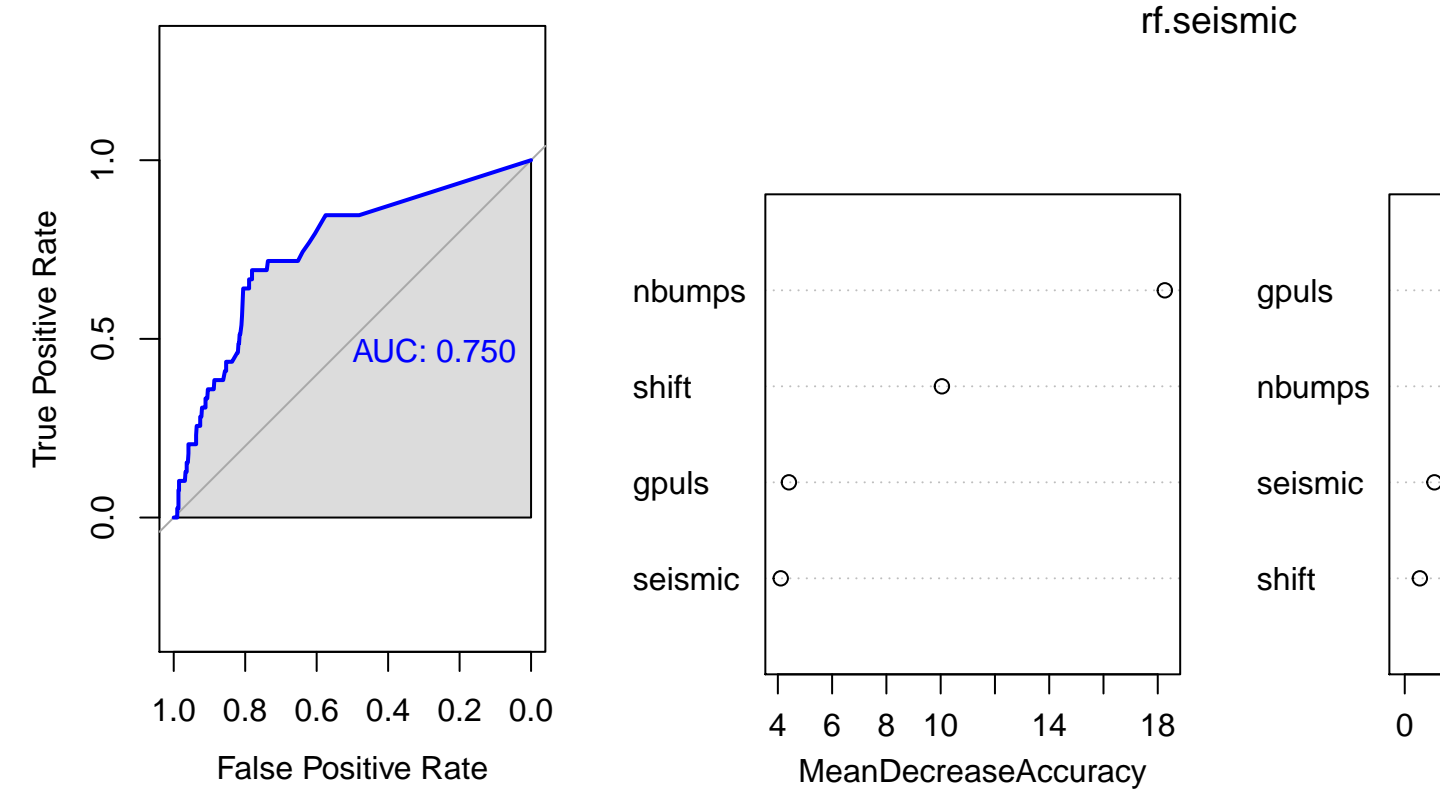
Model 2: Train ROC for RF Classification



0 1 MeanDecreaseAccuracy MeanDecreaseGini

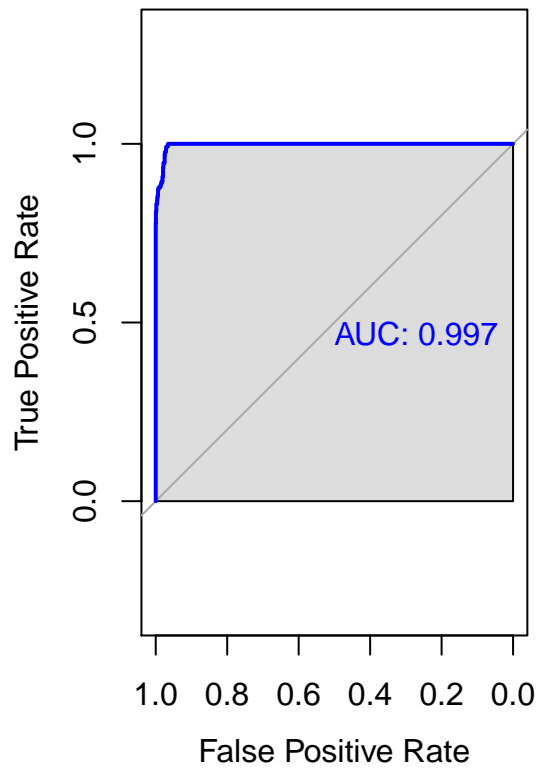
seismic	0.7571748	9.478397	4.105877	5.821832
shift	11.4969471	-9.940458	10.045574	2.959784
gpuls	1.6136431	7.619120	4.407885	75.308097
nbumps	12.7174931	21.298900	18.259825	27.149236

Model 2: Test ROC for RF Classification



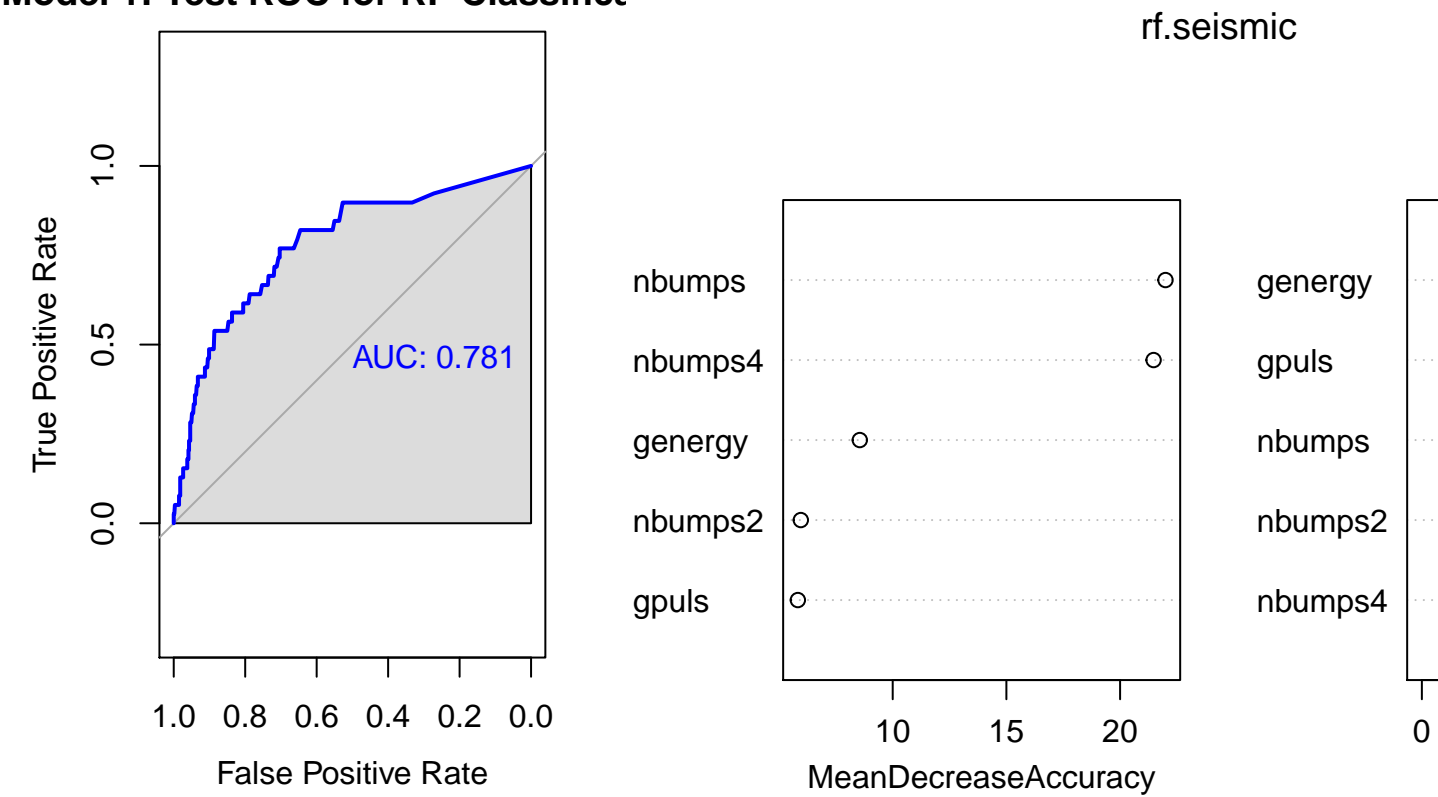
Boosting

Model 1: Train ROC for RF Classifier:

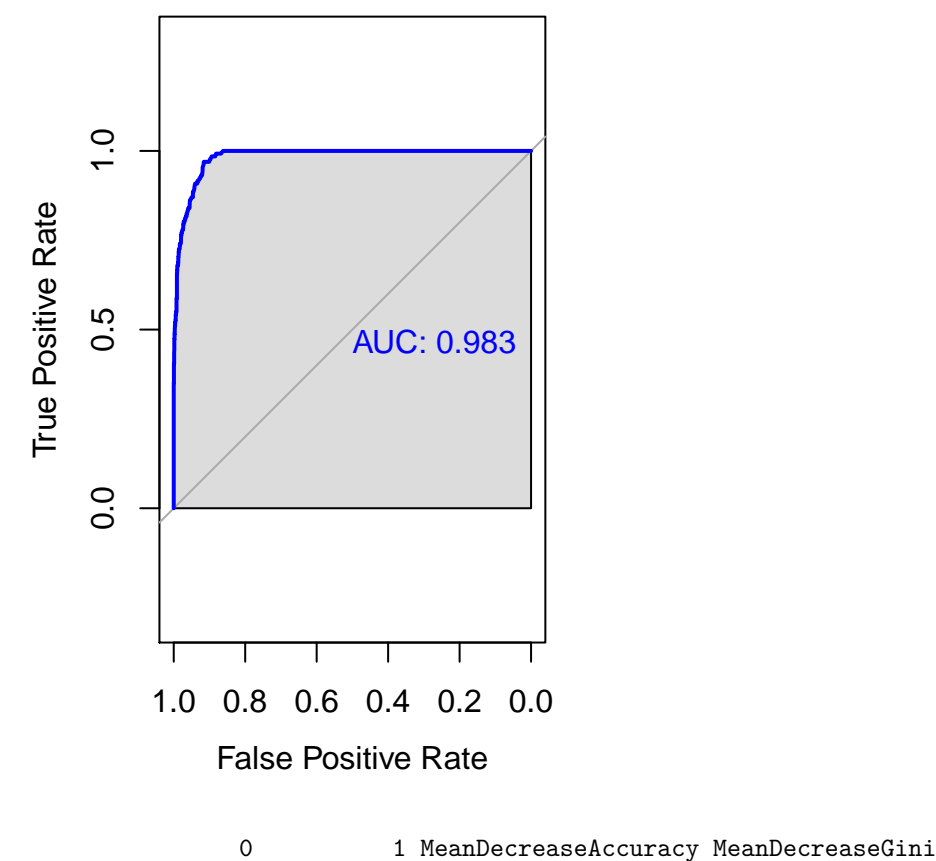


	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
genergy	5.9551277	6.344774	8.553180	68.108834
gpuls	0.2512005	19.107813	5.829010	66.967880
nbumps	13.6588559	34.168577	21.995502	25.976946
nbumps2	1.7909806	11.436422	5.958585	14.314461
nbumps4	23.5704107	-7.592203	21.472542	7.171765

Model 1: Test ROC for RF Classification

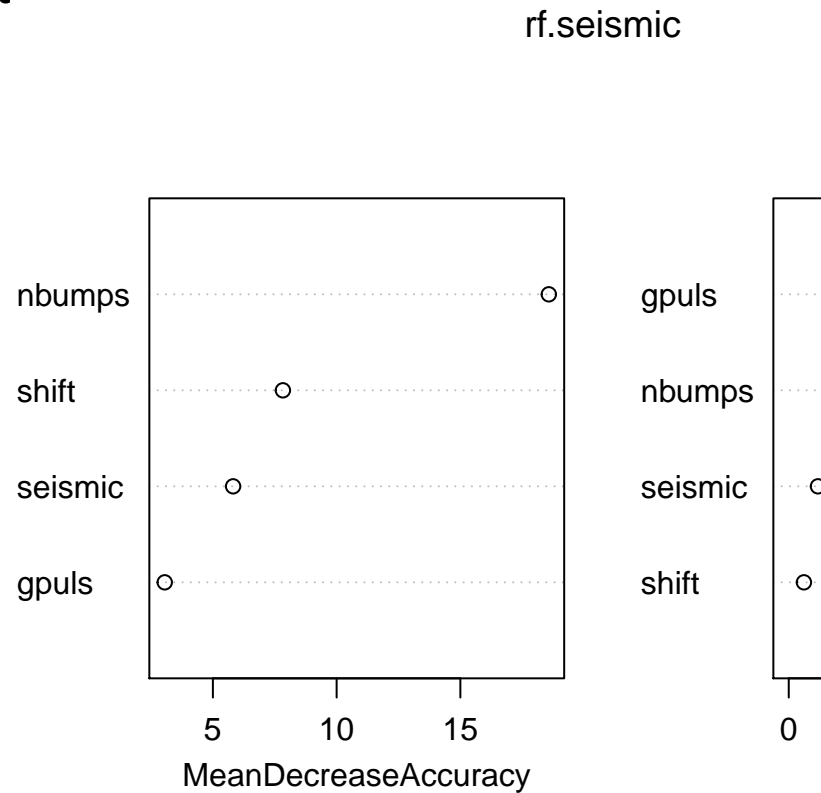
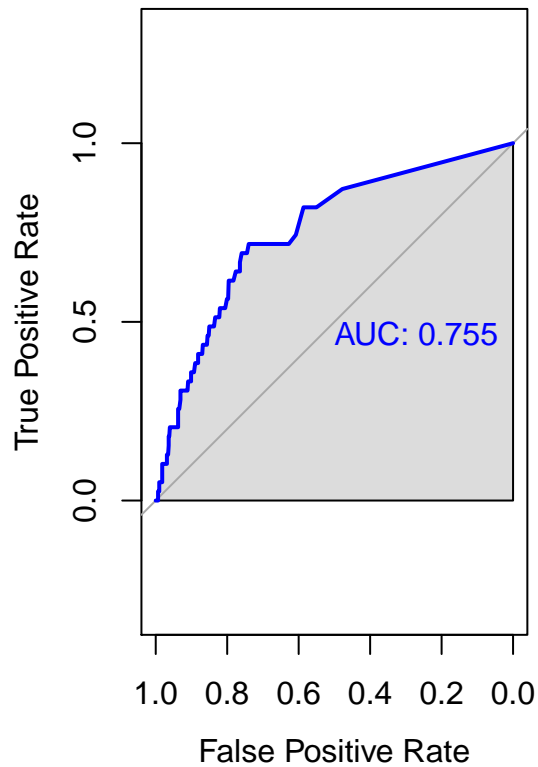


Model 2: Train ROC for RF Classification



seismic	1.9780394	11.987044	5.816797	5.702709
shift	9.8636851	-10.596828	7.828376	2.964649
gpuls	0.3244053	7.259524	3.053310	74.992297
nbumps	12.5548977	20.947972	18.574913	27.469977

Model 2: Test ROC for RF Classification



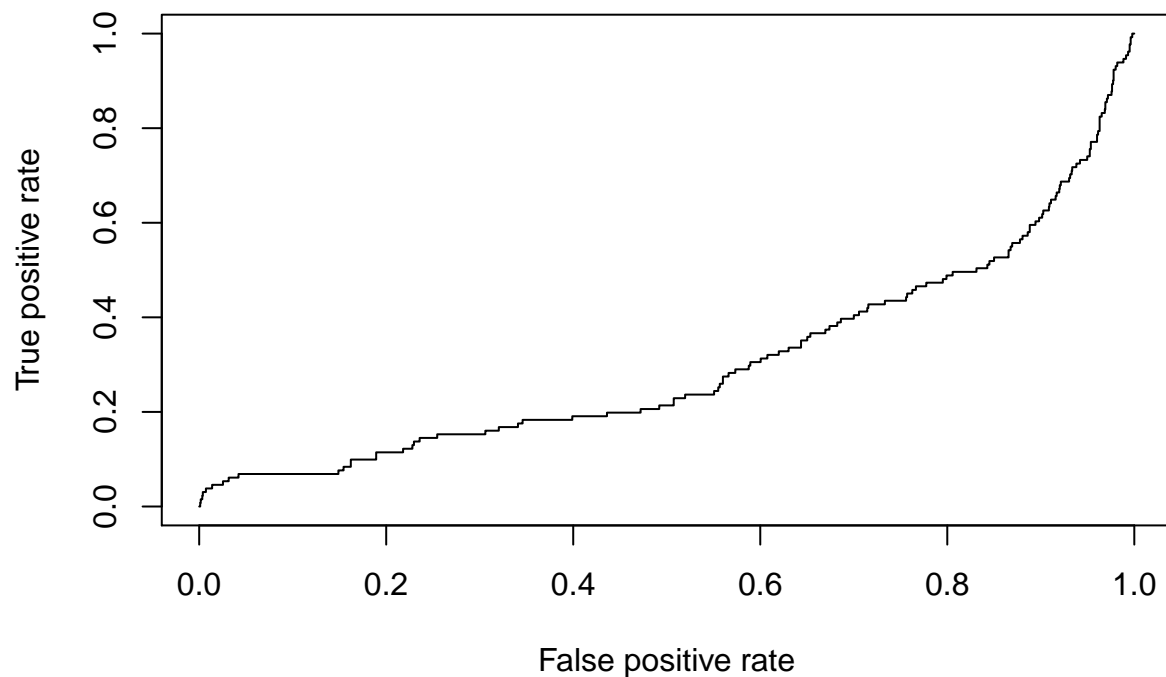
Support vector classifier and support vector machine

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##   cost
##   0.001
##
## - best performance: 0.06759521
##
## - Detailed performance results:
##   cost      error dispersion
## 1 0.001 0.06759521 0.01071376
## 2 0.010 0.06759521 0.01071376
## 3 0.100 0.06759521 0.01071376
## 4 1.000 0.06759521 0.01071376
## 5 5.000 0.06759521 0.01071376
```

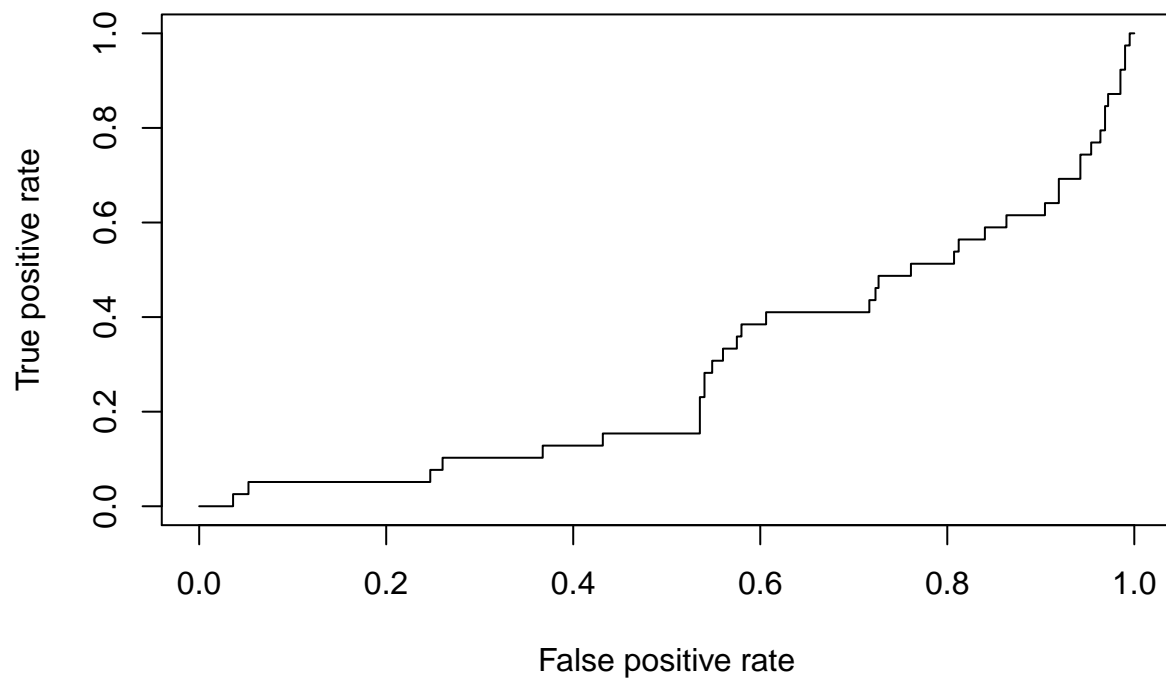
```
##
## Call:
## best.tune(method = svm, train.x = factor(class) ~ genergy + gpuls +
##      nbumps + nbumps2 + nbumps4, data = seismic[train, ], ranges = list(cost = c(0.001,
##      0.01, 0.1, 1, 5)), kernel = "linear")
##
##
## Parameters:
##   SVM-Type:  C-classification
##   SVM-Kernel: linear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

Training data



Test data

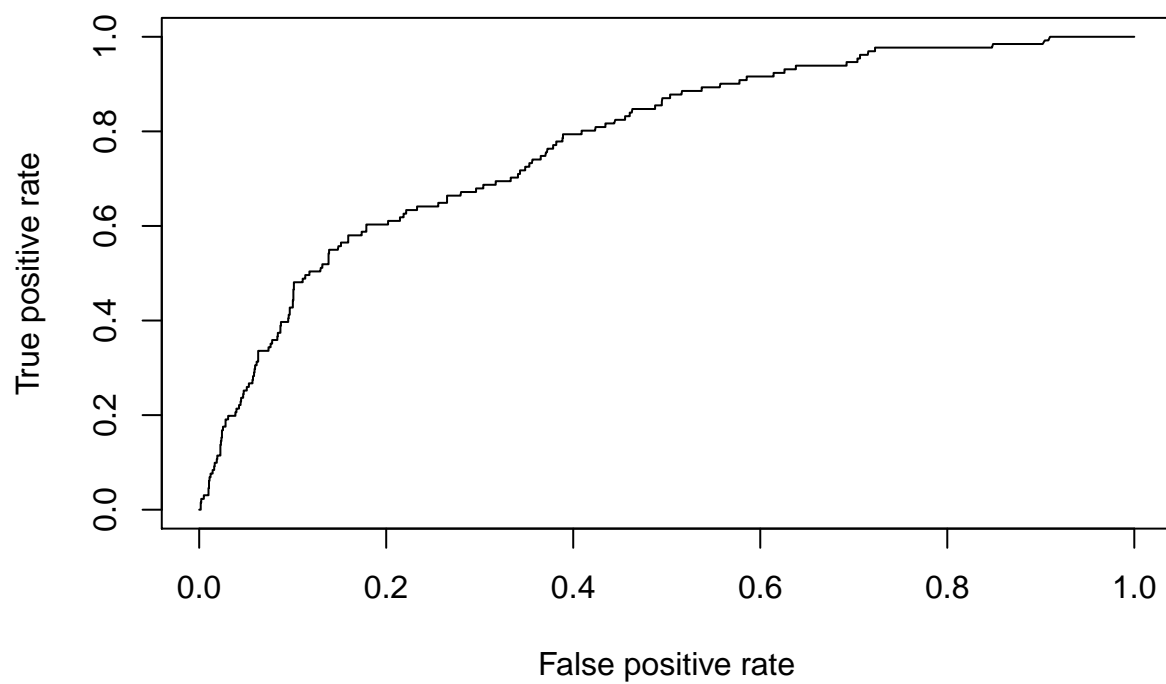


```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##   cost
##   0.001
##
## - best performance: 0.06759521
##
## - Detailed performance results:
##   cost      error dispersion
## 1 0.001 0.06759521 0.01638741
## 2 0.010 0.06759521 0.01638741
## 3 0.100 0.06759521 0.01638741
## 4 1.000 0.06759521 0.01638741
## 5 5.000 0.06759521 0.01638741
##
## Call:
## best.tune(method = svm, train.x = factor(class) ~ seismic + shift +
##   gpuls + nbumps, 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
```

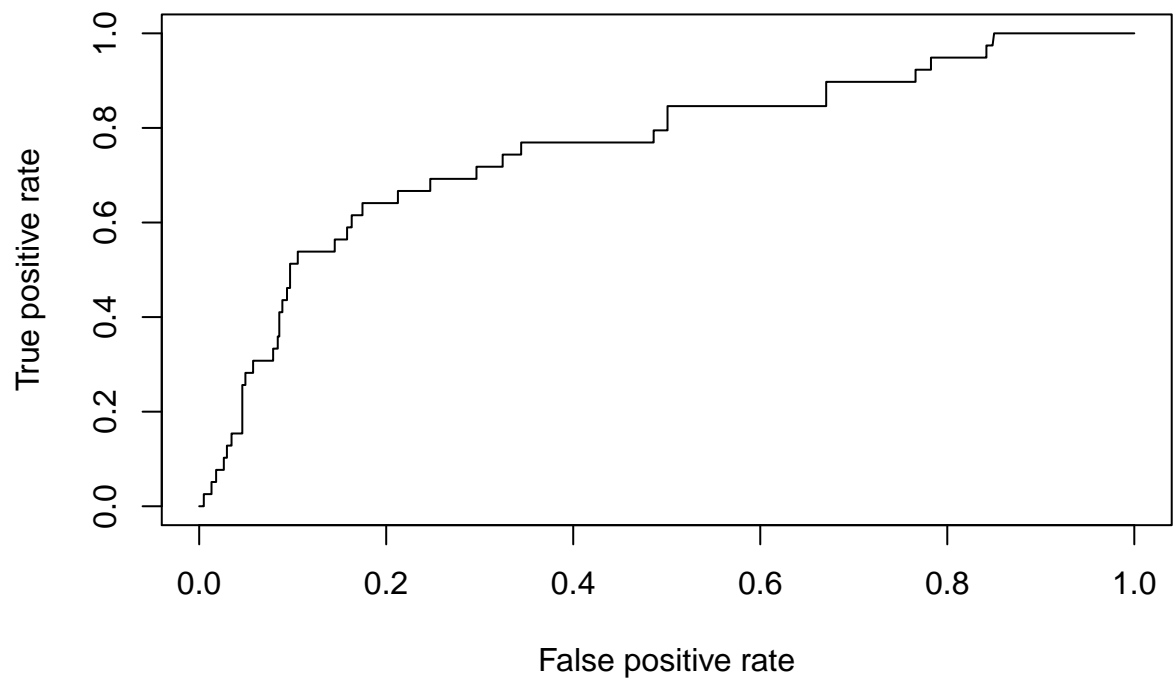
```
##      cost: 0.001
##      gamma: 0.25
##
## Number of Support Vectors: 265
##
## ( 134 131 )
##
##
## Number of Classes: 2
##
## Levels:
## 0 1
```

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

Training data

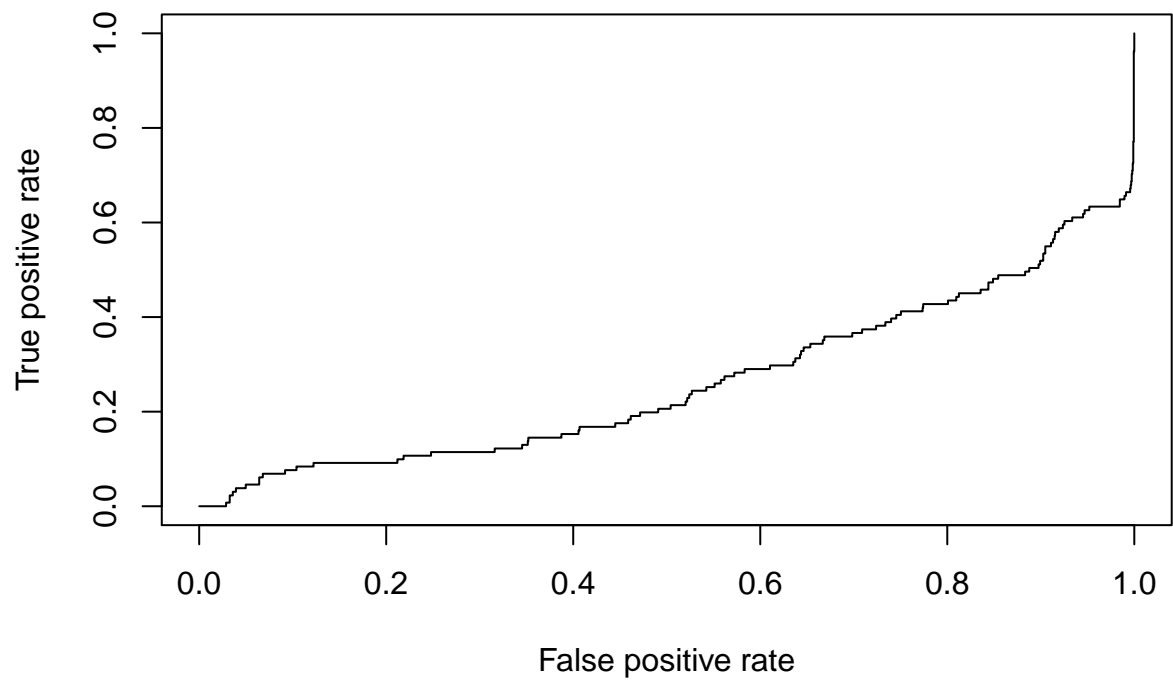


Test data

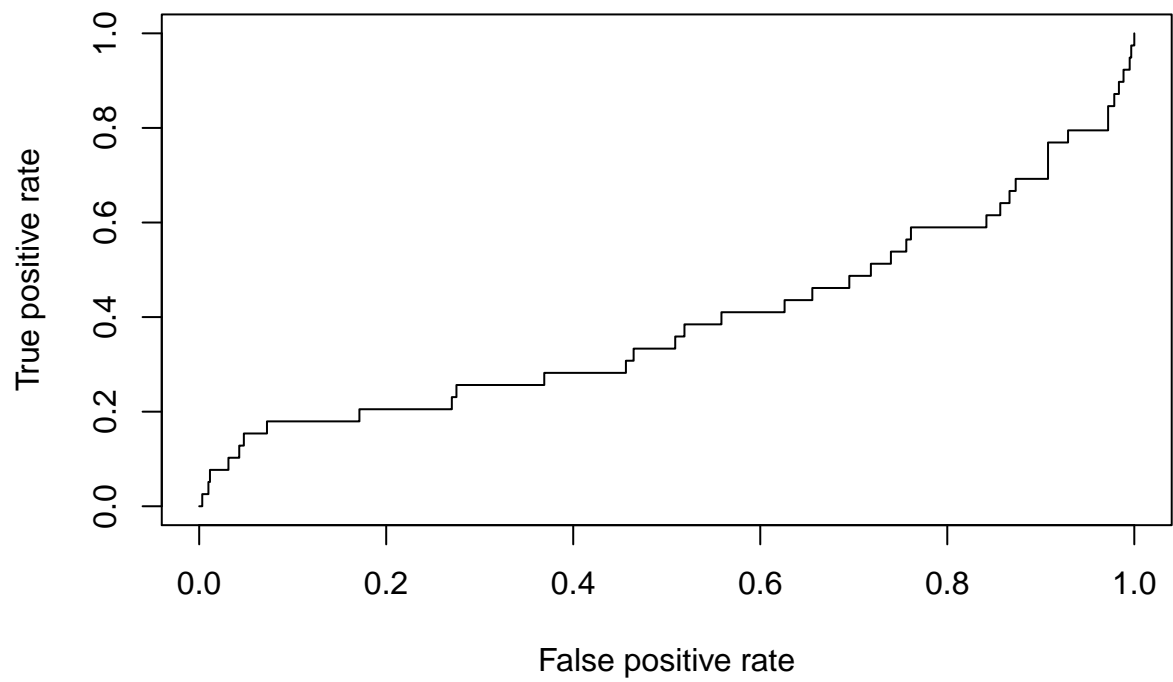


```
##      truth
## predict  0   1
##      0 607  39
##      1   0   0
```

Training data

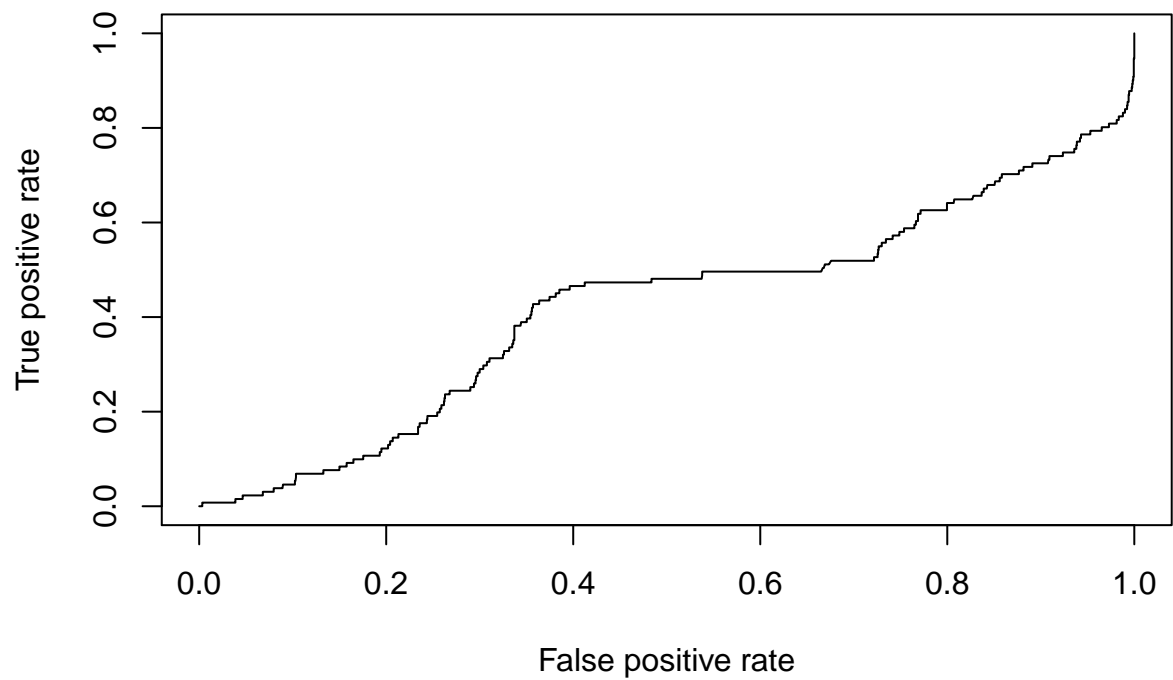


Test data

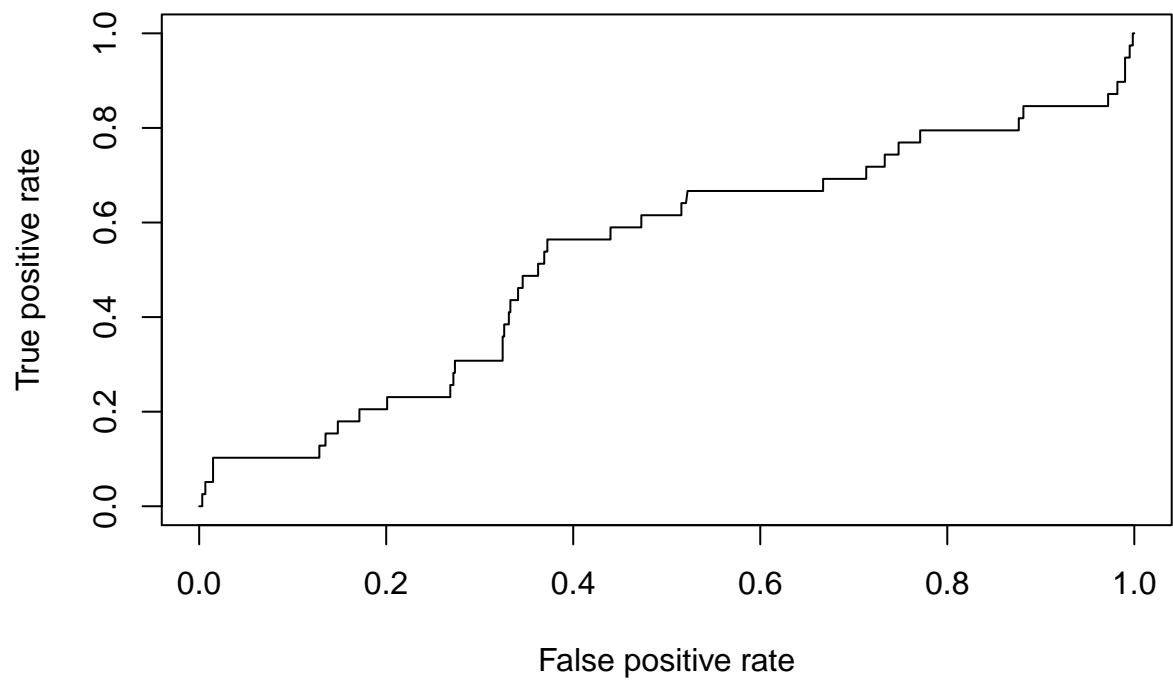


```
##      truth
## predict  0   1
##         0 607 39
##         1   0  0
```

Training data

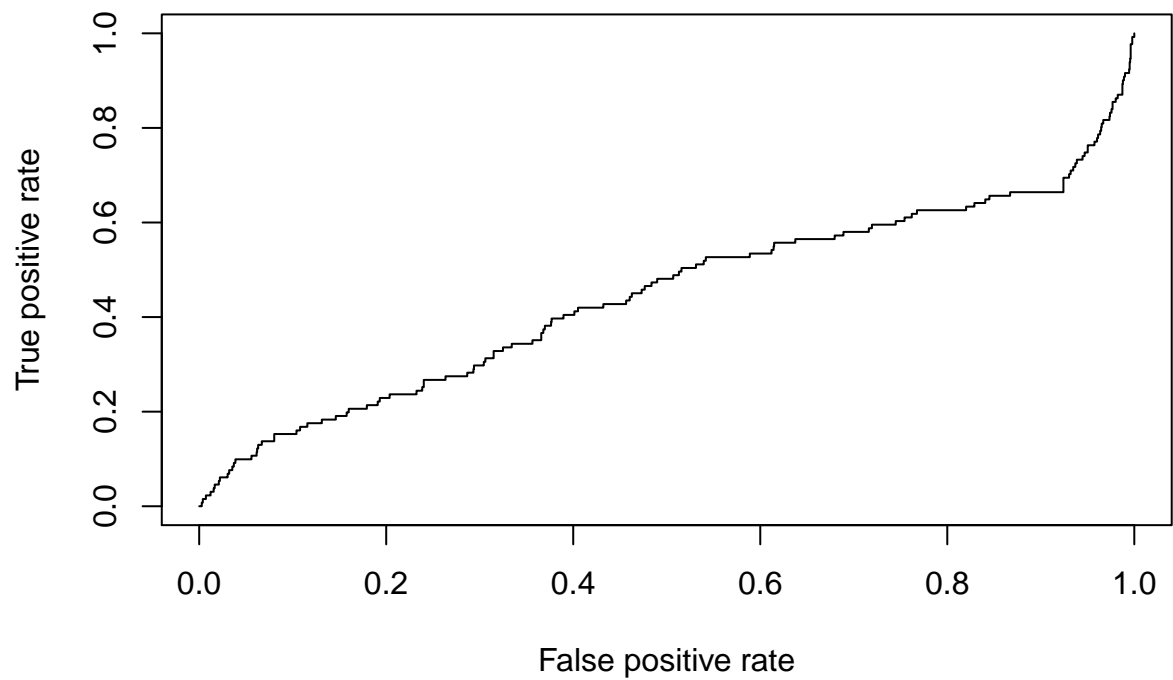


Test data

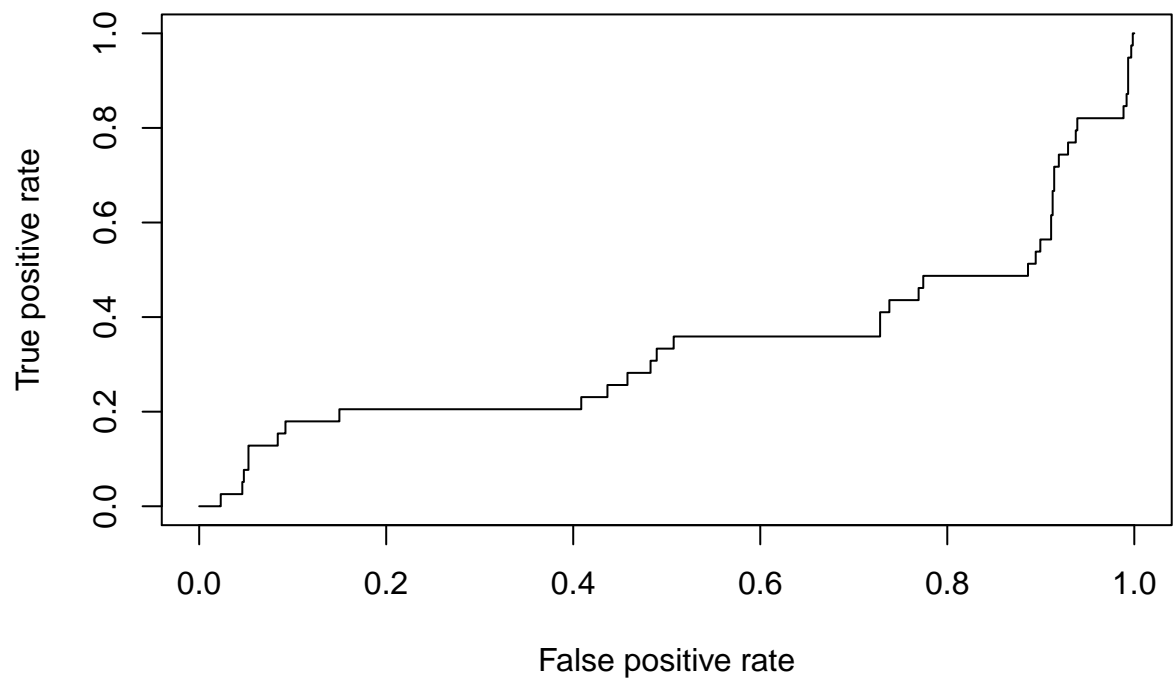


```
##      truth
## predict  0   1
##      0 607  39
##      1   0   0
```

Training data

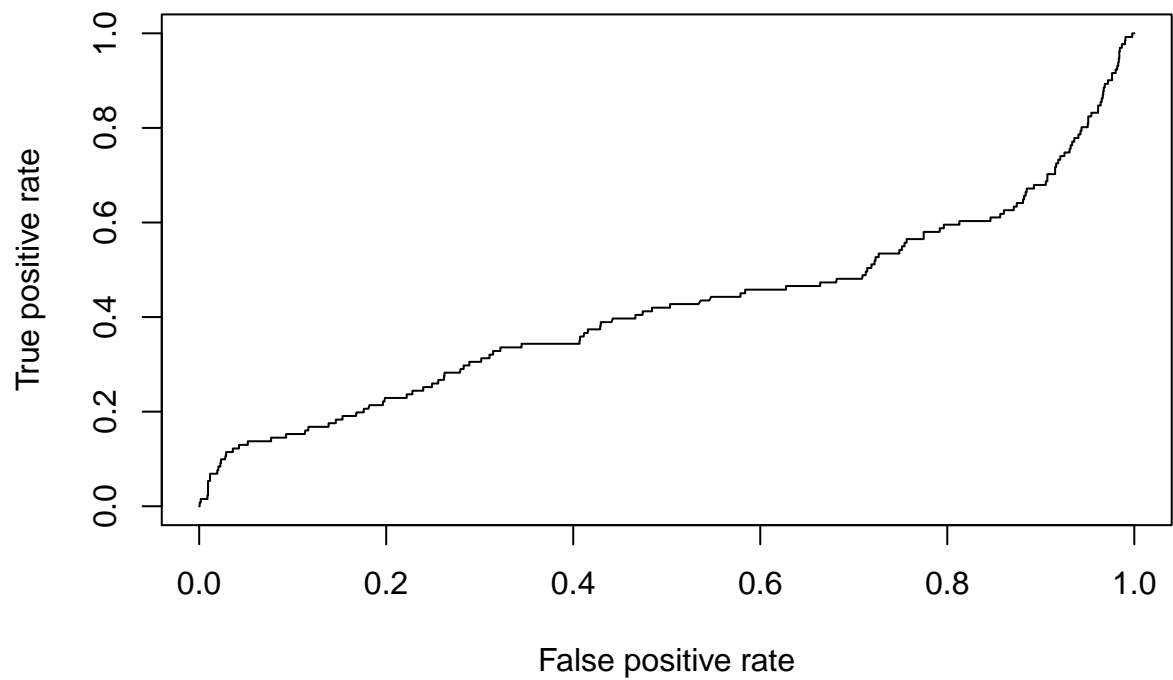


Test data

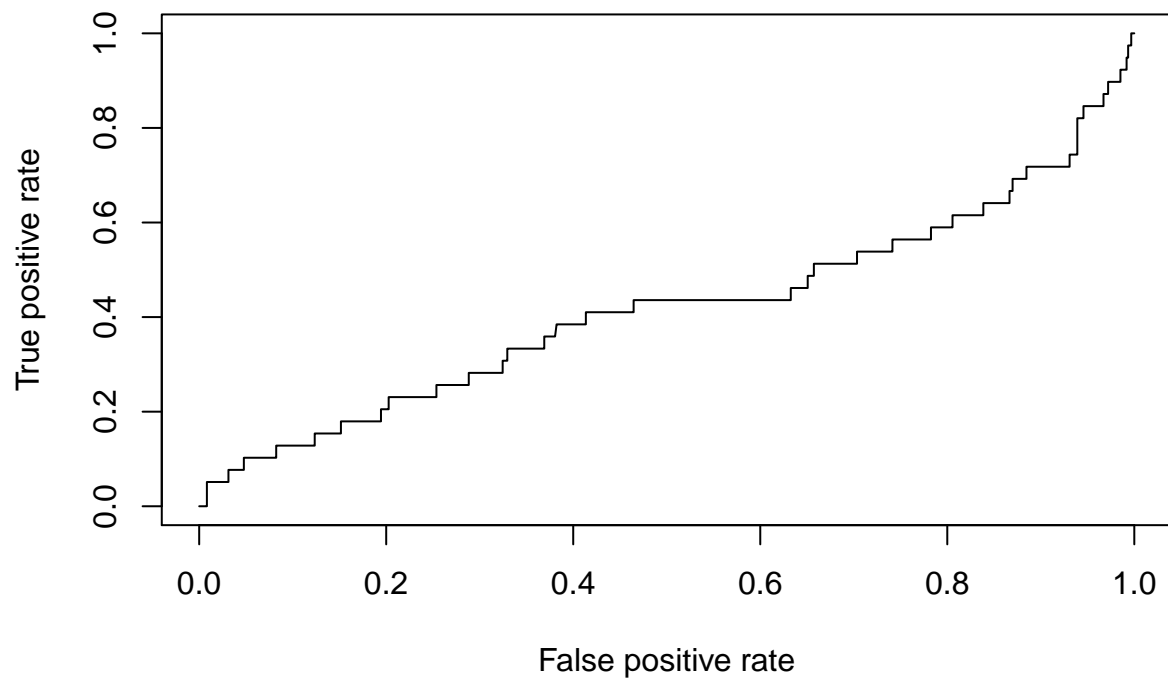


```
##      truth
## predict  0   1
##        0 607  39
##        1   0   0
```

Training data



Test data



	<i>Model 1</i>			<i>Model 2</i>		
<i>Kernel</i>	linear	radial	polynomial	linear	radial	polynomial
<i>cost</i>	.001	.001	.001	.001	.001	.001
<i>gamma</i>	.2	1	.2	.25	1	.25
<i>degree</i>	N/A	N/A	2	N/A	N/A	2
<i>time</i>	5.4	26.47	25.31	2.44	24.84	7.51
<i>misclassification rate</i>	.06	.06	.06	.06	.06	.06

How to time your code!!!

```
#-----
# How to time your method
#-----

# Put this before your method
start.time <- proc.time()

## the thing you are computing, like random forest or SVM goes here ##

total.time <- proc.time() - start.time

total.time[3] # the elapsed time

## elapsed
## 0.001
```

Appendix

Table I. Attribute information of the seismic-bumps dataset

Data Attributes	Description
seismic	result of shift seismic hazard assessment: ‘a’ - lack of hazard, ‘b’ - low hazard, ‘c’ - high hazard, ‘d’ - very high hazard
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 longwall
gpuls	number of pulses recorded within previous shift by GMax
gdenenergy	deviation of recorded energy within previous shift from average energy recorded during eight previous shifts
gdpuls	deviation of recorded pulses within previous shift from average number of pulses recorded during eight previous shifts
ghazard	result of shift seismic hazard assessment by the seismoacoustic method based on registration coming from the previous shift
nbumps	the number of seismic bumps recorded within previous shift
nbumps i , $i \in \{1, \dots, 5\}$	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’), ‘0’ - otherwise

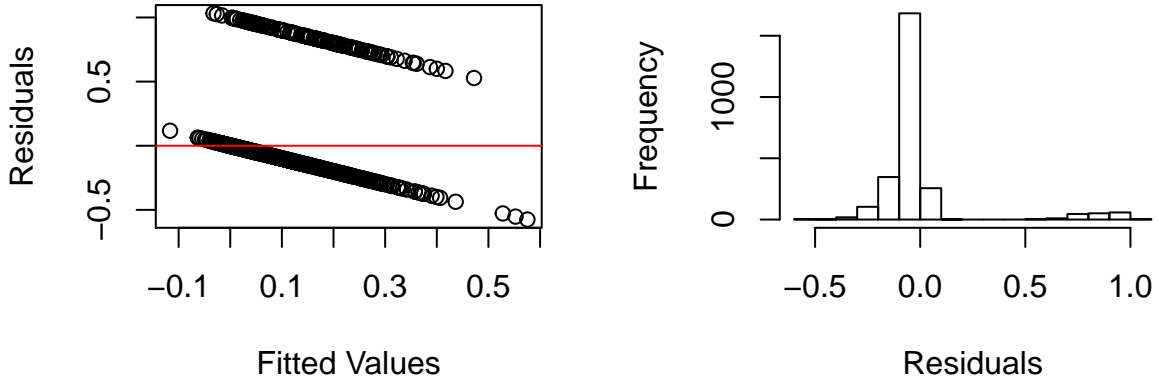


Table 7: Table II-VIFs of Linear Model

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

Table 8: Table II-VIFs of Linear Model

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