Data Mining Project 3

Ben Straub, Hillary Koch, Jiawei Huang, Arif Masrur

April 18th, 2017

1 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 & Wrobel, 2010). Minimizing loss from seismic hazards requires advanced data collection and analysis. In recent years, more and more cutting-edge 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 data set provided by Sikora & Wrobel (2010), found in the UCI Machine Learning Repository. This seismic-bumps data set 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.

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 from the 18 variables. 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, meets 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 for logistic regression, LDA, QDA and RDA and use it as a benchmark to look at our advanced techniques. Below is our step-model:

$$class \sim \beta_0 + \beta_1 \cdot genergy + \beta_2 \cdot gpuls + \beta_3 \cdot nbumps + \beta_4 \cdot nbumps2 + \beta_5 \cdot nbumps4 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

In section 2, we report ROC curves, missclassification rates and computing time for logistic regression, LDA, QDA and RDA. In section 3, we report the same(ROC curves, missclassification rates and computing time) for boosting, random forest and support vector machines. As in project 2, effectiveness of a method will be determined by assessing the misclassification rate for each method, but we will also look at the receiver operating characteristic (ROC curve). In a ROC curve the true positive rate(TPR) is plotted in function of the false positive rate(FPR) for different cut-off points. Each point on the ROC curve represents a TPR/FPR pair corresponding to a particular decision threshold. The area under the curve (AUC) will be used to help judge the effectiveness of a method. Ideally, we would like an AUC value of 90% or higher to determine a "good fit" for the seismic data. In section 4, we provide concluding remarks and our project four proposal.

2 Logistic Regression, LDA, QDA, RDA

2.1 Logistic Regression

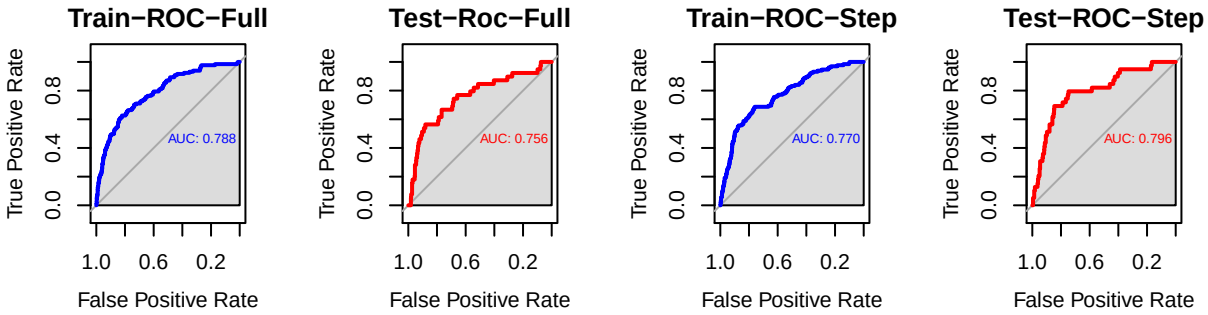


Table 1: Logistic Regression

	Full	Step
Computing Time	0.116	0.082
Train Error Rates	0.067	0.070
Test Error Rates	0.065	0.062

The first method we used was logistic regression. The full model was using all of the predictors in the training data set. Initially, we used a threshold probability of 0.5 to classify into state 0 or 1. This yields an overall error rate of 6.7% for the training data and 6.5% for the test data, with minimal improvement in sensitivity. The ROC curve for this model indicates that it is still not a great fit for the data. However, reducing the dimension of the full model through step-wise regression produces a 4% increase in AUC and a slight reduction in missclassification found in Table 1. Computing time in seconds is all provided in Table 1.

2.2 Linear Discriminant Analysis

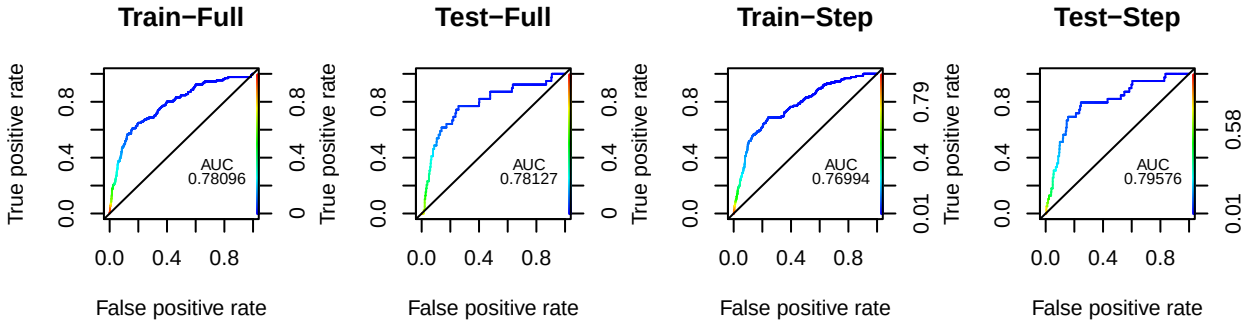


Table 2: Linear Discriminant Analysis

	Full	Step
Computing Time	0.472	1.371
Train Error Rates	0.074	0.081
Test Error Rates	0.077	0.076

It is reasonable to believe that each class is distributed normally with some different mean vector. Thus, we

implemented an LDA approach. Using a classification threshold of 0.5 yields an overall error rate of 7.4% for the training data and 7.7% for the test data, but with much higher sensitivity than in the previous models. The group means suggest that a mining shift with a higher number of seismic bumps and associated higher released energy (measured in Joules) is correlated with hazard status of the mine in the subsequent shift. We performed linear discriminant analysis on the data after variable selection. For the reduced model we observed an error rate of 8.10% on the train data an error rate of 7.6% for the rest data. The AUC's for the full model and step model are roughly the same as the logistic regression model. Computing time is available in Table 2.

2.3 Quadratic Discriminant Analysis

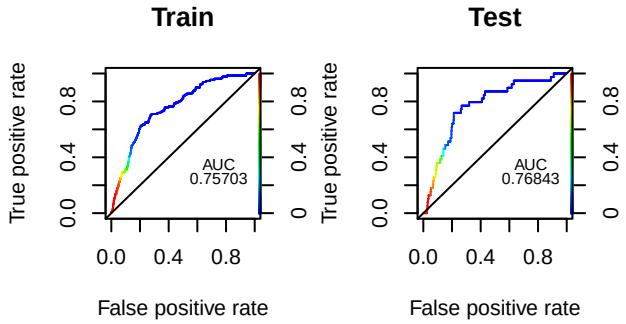


Table 3: Quadratic Discriminant Analysis

	Full	Step
Computing Time	NA	0.46
Train Error Rates	NA	0.149
Test Error Rates	NA	0.159

Quadratic discriminant analysis (QDA) provides an alternative approach to LDA in that QDA assumes each class has its own covariance matrix. This allows the method to be more flexible in some ways, although the singularity of the covariance matrices resulting from multicollinearity prohibited us from fitting a QDA model prior to variable selection. We observed an error rate of 15% for Training set and 16% for the Test set using the step-model. AUC's for the Step model are lower when compared to the previous method. Computing times are available in table 3.

2.4 Regularized Discriminant Analysis

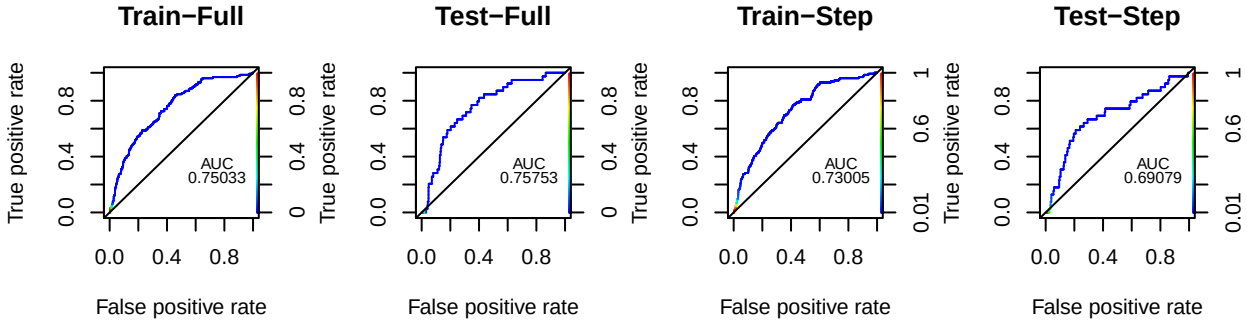


Table 4: Regularized Discriminant Analysis

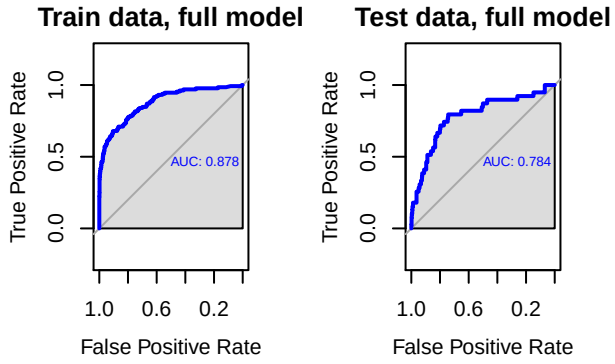
	Full	Step
Computing Time	3.265	2.245
Train Error Rates	0.076	0.082
Test Error Rates	0.082	0.085

Similar to LDA, it is reasonable to believe that each class is distributed normally with some different mean vector and covariance matrix, as seen in quadratic discriminant analysis (QDA). RDA allows for a compromise between LDA and QDA. We implemented an RDA approach. Using a classification threshold of 0.5 yields an overall error rate of 7.6% for the training data and 8.2% for the test data. Similar results were obtained for the step model. The AUC's are also observed as quite poor. Poor performance on the test data might be indicative of RDA over fitting for this model.

3 Boosting, Random Forest Classification, Support Vector Machines

3.1 Boosting

Next we performed boosting to our data set to 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)



We also did boosting on model 1 (from stepwise variable selection), but the original model performs best, when evaluating by misclassification rate (shown in the table below). Therefore, in the interest of space, we only report the original model's ROC curves. The AUC for the train data is significantly better while the test data's AUC is roughly equivalent to logistic regression and LDA.

Table 5 Boosting

model	Full model	Stepwise model
time	10.38	5.64
misclassification rate	.057	.060

3.2 Random Forest Classification

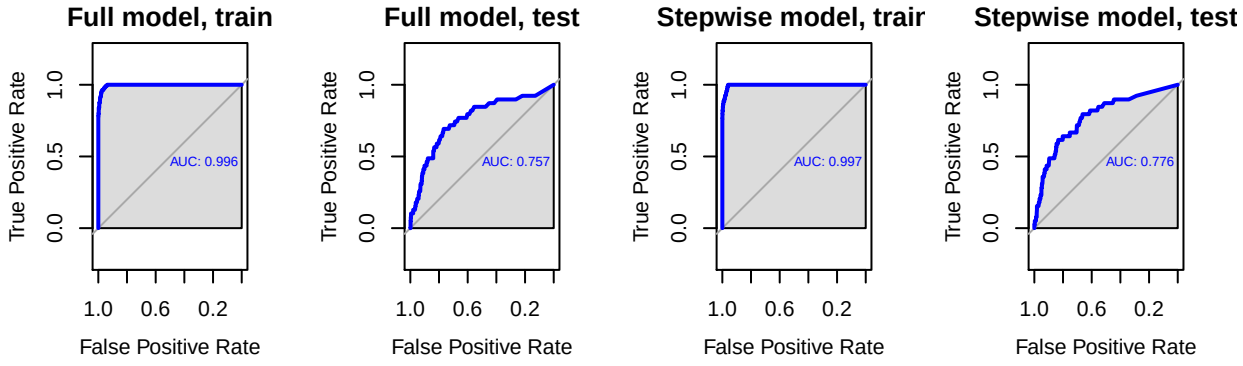
Next, we use Random Forest classification method as it yields relatively better classification results among all tree-based methods. As opposed to growing a single decision tree (as in CART), random forest grows multiple trees, with having each split to consider only a subset of all predictors.Then it takes average of all trees to make final tree. In this way, random forest can reduce amount of potential 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.

Then, we applied Random Forest formula on both train and test data sets 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 classification process.

Random forest on full model

Here we performed random forest classification on training and test data sets 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 data set are less accurate than predictions on train data set. We also 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.



Random forest on stepwise model

Here we performed random forest classifications on resulting model from stepwise variable selection. Missclassificaion rates for stepwise model's training and test data sets are 6.9% and 5.4%, respectively. This time the ROC curves revealed slightly improved test AUC. According to the variable importance plot, most important variables 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.

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 variables for predicting potential seismic events in future mining shift(s).

RF on Models	Error Rate	Important Variable(s)
Full Model (Train)	9.4%	nbumps2,3,4 , genergy, nbumps, maxenergy, gdenergy, gpuls, and energy
Full Model (Test)	5.7%	nbumps2,3,4 , genergy, nbumps, maxenergy, gdenergy, gpuls, and energy
Stepwise model (Train)	10%	nbumps and nbumps4
Stepwise model (Test)	7.8%	nbumps and nbumps4

Time elasped	Full Model: 5.09	Stepwise Model: 1.90

3.3 Support vector classifier and support vector machine

A support vector machine allows for the classification of data with, if desired, non-linear boundaries. Essentially, the idea is to project the data into a space of higher dimension, determine decision boundaries, and then project back into the original space. We attempted to fit this approach to our data, using a linear, radial, and polynomial kernel.

Fitting a model with a linear kernel is a support vector classifier with an $n \times p$ data set, and can be computed by solving the optimization problem

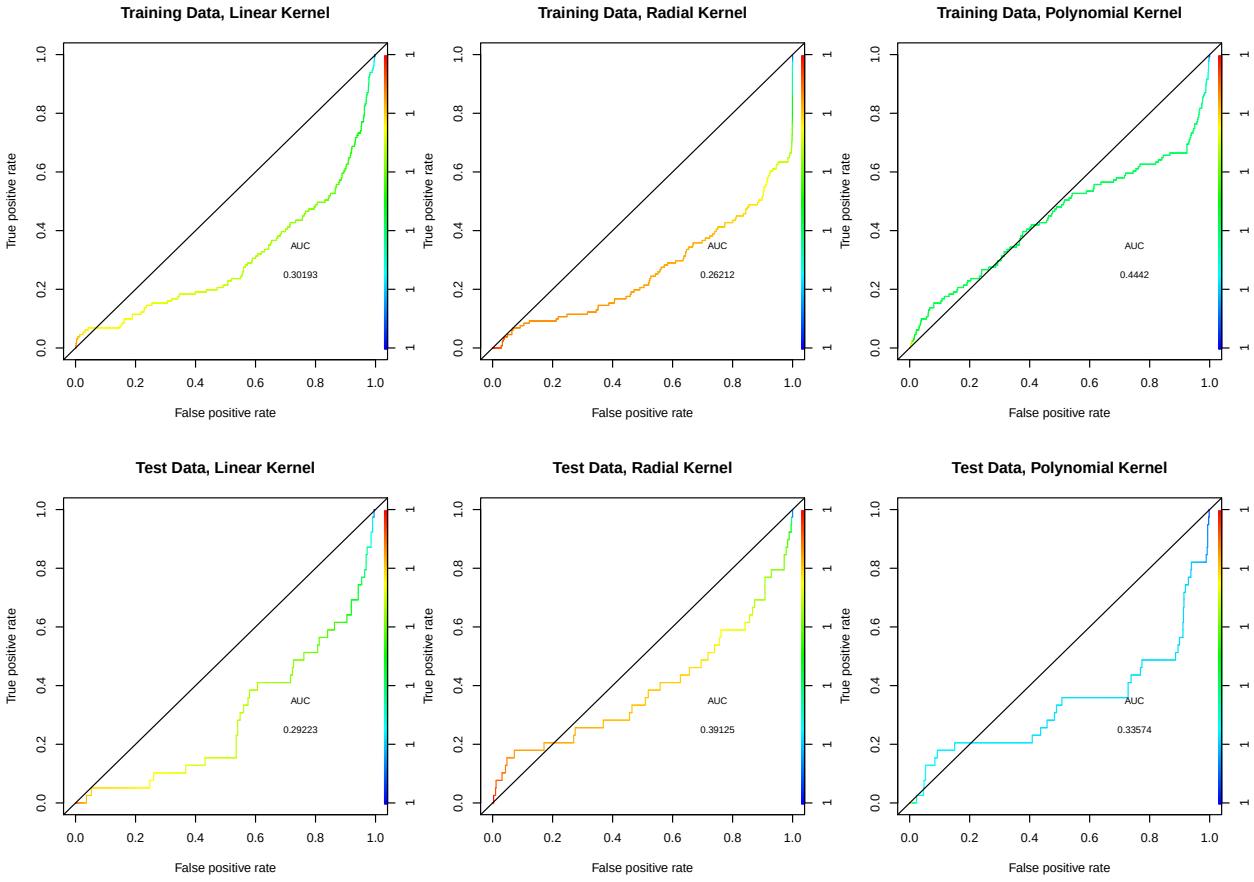
$$\text{maximize}_{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n} \quad M$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon),$$

However, replacing the inner product with a different, non-linear, kernel function permits different, more sophisticated, implementations of the support vector machine.

With an increased cost, 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.



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

SVM missclassification rates are identical across the three kernels. However, the AUC's from the ROC curve are quite poor, it would be better to randomly guess than use the SVM method.

4 Conclusions and future work

It is clear that variable selection in general improves the ability to predict hazardous events in the mine, according to our data. However, the improvements afforded are often minimal. In addition, while dimension reduction tends to dramatically improve sensitivity, this is often done at the cost of increasing overall error rate. Using the summaries of our findings, we recommend still using logistic regression to predict seismic events. LDA is a close second. However, random forest classification or boosting with a better understanding of its "fine-tuning" might achieve better results. Especially if more data was available.

For our final project using cluster analysis we will continue our exploration of seismic data. The Symposium on Advances in Artificial Intelligence and Applications had a competition in 2016 using a seismic data set produced from the same authors as this current data set. The original data set involves over 70,000 observations, but we will investigate a training set put forth in the competition that only has 13,000 observations.

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, ‘c
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 10: 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