STAT 557 - Project 2

Ben Straub, Hillary Koch, Jiawei Huang, Arif Masrur

3/27/2017

Data Overview

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 both 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 this project, we used the seismic-bumps dataset provided by Sikora & Wrobel (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. According to Bukowska (2006), a number of factors having an effect on seismic hazard occurrence were proposed. Among other factors, the occurrence of tremors with energy $> 10^4$J was listed. The purpose is to find whether and how the other 18 variables can be used to determine the hazard status of the mine.

Table 1. Attribute information of the seismic-bumps dataset

Data Attributes	Description	Data Types
seismic	result of shift seismic hazard assessment: 'a' - lack of hazard, 'b' - low hazard, 'c' - high hazard, 'd' - danger state	Categorical
seismoacoustic	result of shift seismic hazard assessment	Categorical
shift	type of a shift: 'W' - coal-getting, 'N' - preparation shift	Categorical
genergy	seismic energy recorded within previous shift by active geophones (GMax) monitoring the longwall	Continuous
gpuls	number of pulses recorded within previous shift by GMax	Continuous
gdenergy	deviation of recorded energy within previous shift from average energy recorded during eight previous shifts	Continuous
gdpuls	deviation of recorded pulses within previous shift from average number of pulses recorded during eight previous shifts	Continuous
ghazard	result of shift seismic hazard assessment by the seismoacoustic method based on registration coming from GMax	Categorical
nbumps	the number of seismic bumps recorded within previous shift	Continuous
nbumps$i$, $i \in \{1, \ldots, 5\}$	the number of seismic bumps ($10^i - 10^{i+1}$ J) registered within previous shift	Continuous
energy	total energy of seismic bumps registered within previous shift	Continuous
maxenergy	maximum energy of the seismic bumps registered within previous shift	Continuous
class	the decision attribute: '1' - high energy seismic bump occurred in the next shift ('hazardous state'), '0' - no high energy seismic bumps occurred in th next shift ('non-hazardous state')	Categorical

Exploratory Data Analysis

The state of the mine was indeed deemed hazardous infrequently − only 170 shifts out of 2584 − a difficult problem in our analyses. We want to examine which observations of seismic activity can help in the prediction of the hazard state of the mine during the next shift. 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, as shown below.
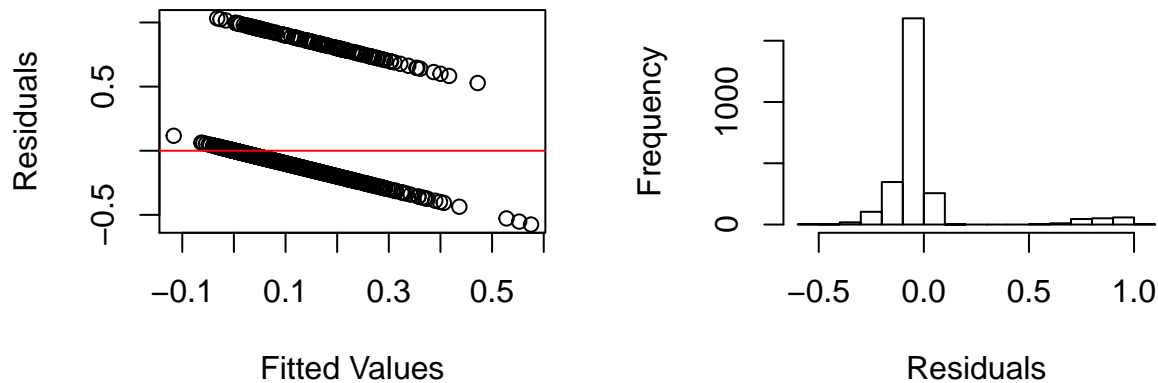

Table 2: VIFs of Linear Model

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

Table 3: 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

Classification before Variable Selection

We first take the seismic-bumps dataset and partition the data into training (75%) and test (25%) datasets. The next steps involve examining multiple classification methods on the training and test datasets separately. The goal is to examine which classification method outputs comparatively better prediction for seismic hazards based on available predictors.

Linear Regression of an Indicator Matrix:

We begin with linear regression of an indicator matrix as this method can approximate a linear decision boundary among observations belonging to one of two classes. Our model outputs an overall error rate of ~6.8% on the training data and ~6.3% on the test data. However, this comes with no sensitivity, which means we would almost never predict a hazardous event in the mine.

Table 4: Training vs. Test for regression of indicator matrix

	Train 0	Train 1	Test 0	Test 1
Predict 0	1804	130	605	39
Predict 1	3	1	2	0

Logistic Regression:

We then fit a logistic regression model to predict the response using all the predictors in the training dataset. Initially, we used a threshhold 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 dest data, with minimal improvement in sensitivity. The ROC curve for this model indicates that it is still not a great fit for the data.

Table 5: Training vs. Test for logistic regression

	Train 0	Train 1	Test 0	Test 1
Predict 0	1802	125	604	39
Predict 1	5	6	3	0

ROC Curve

AUC: 0.756

True Positive Rate (y-axis), *False Positive Rate* (x-axis)
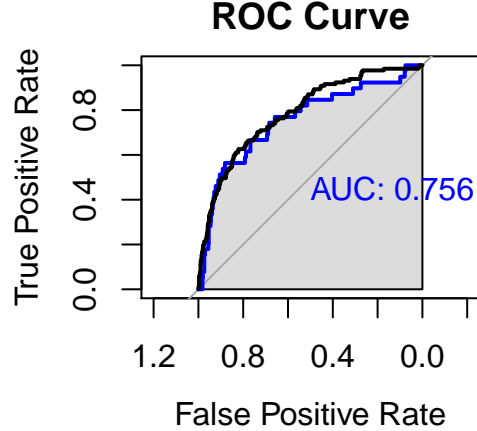
Linear Discriminant Analysis:

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 threshhold 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.

Table 6: Training vs Test

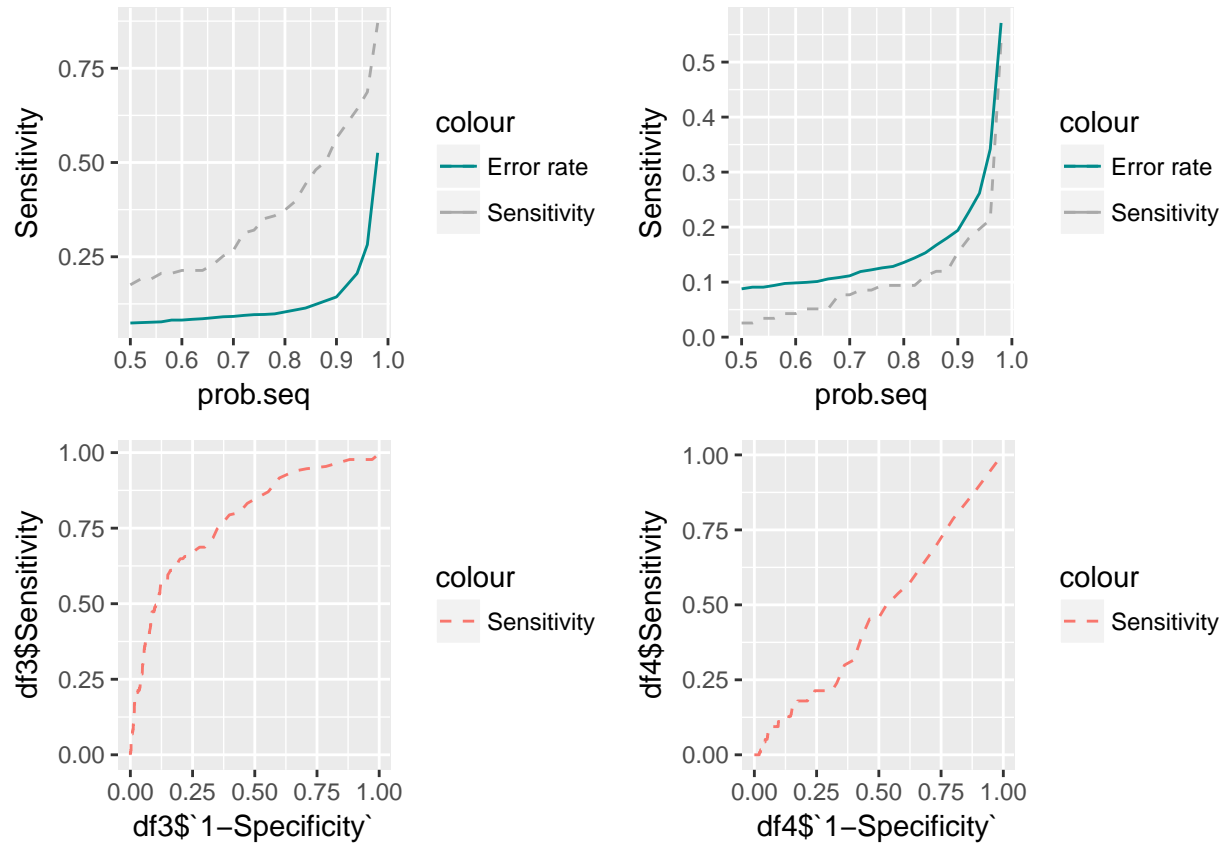|	Train 0	Train 1	Test 0	Test 1
Predict 0	1771	108	591	34
Predict 1	36	23	16	5

Figure 1: Results from the training data are presented on the left, and for the test data on the right. Top: Sensitivity and overall error rate as a function of the threshhold probability used to classify an event as 0 or 1. Bottom: ROC curves show much worse performance on the test data.

Regularized Discriminant Analysis:

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 threshhold of 0.5 yields an overall error rate of ~7.6% for the training data and ~8.2% for the test data. Poor performance on the test data might be indicative of RDA overfitting for this model.

Table 7: Training vs Test

	Train 0	Train 1	Test 0	Test 1
Predict 0	1785	126	593	39
Predict 1	22	5	14	0

Variable Selection and Refitting

Based on high error rate and low prediction accuracy (e.g., low sensitivity) estimates both in the training and test datasets, it is evident that the fitted regression and classification models in the preceding sections are not able to approximate the relationship between response and predictor variables. The strong multicollinearity among some of the predictor variables may have contributed to the high error rate. Therefore,in order tp improve prediction accuracy, we employ some commonly used variable selection methods: stepwise selection, LASSO, and principle component analysis.

Stepwise Variable Selection:

We performed forward and backward selection as a starting point. We did further manual selection, and chose the model that produced the lowest AIC score (AIC = -83.097). We denote the model with these variables as Model 1, which can be written as

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

LASSO:

We also performed Least Absolute Shrinkage and Selection Operator (LASSO) regression, which fits a model by shrinking some of the coefficients toward exactly zero using the $\mathcal{L}_1$ penalty. According to LASSO, 11 of the 15 covariates should be removed from the model. The 4 remaining covariates are seismic, shift, gpuls, and nbumps. These variables are incorporated into a model, denoted Model 2.

Principal Component Analysis:

From the result of PCA we find that the first 11 components explain 97.2% of variance. Looking at variable loading from PCA we can see that the variable load is small (<0.6 for the first four component).

```
# Needs code
```

Logistic Regression after Variable Selection

In logistic regression, we first fit the model onto the training data. From the confusion matrix, we can see that the sensitivity (2.3%) is still very low and the specificity is very high (99.8%), with 6.8% overall error
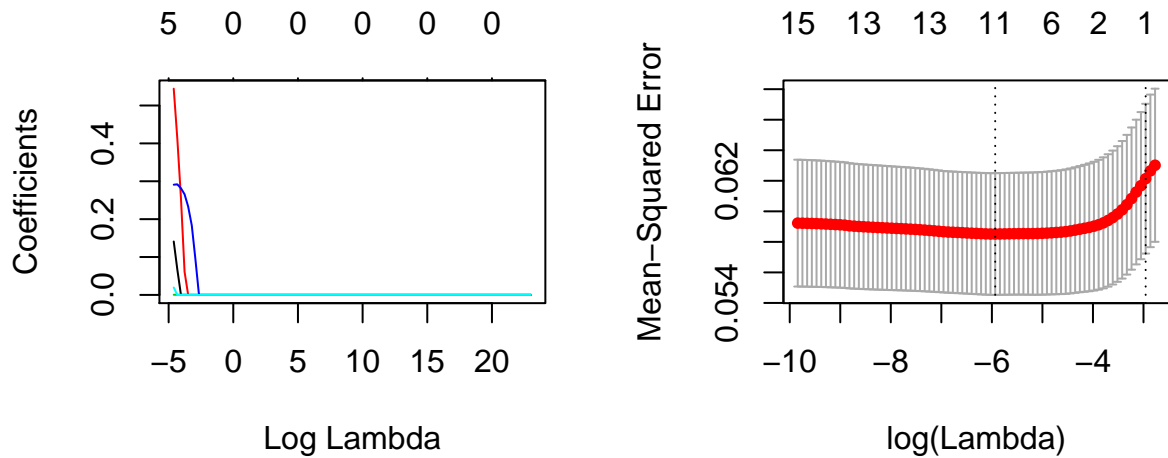
Figure 2: Plots of covariate coefficients (left) and lambda selection (right) for LASSO.

rate. Our fitted model on training observations has peformed poorly on test observations by further reducing sensitivity score to 0%. Therefore, it is evident that our logistic regression model doesn't perform well on the seismic-bumps dataset to provide reliable estimates for the prediction of seismic hazards.

**** I think, below table shows Model 1 vs Model 2, but that actually is Train vs Test on Model 2 only.

Table 8: Model-1 vs Model-2

	0	1	0	1
0	1803	128	606	39
1	4	3	1	0

Table 9: Model-1 vs Model-2

	0	1	0	1
0	1799	127	606	39
1	8	4	1	0

Quadratic Discriminant Analysis after Variable Selection

Quadratic Discriminant Analysis (QDA) provides an alternative approach to the LDA in that QDA assumes each class (hazardous vs non-hazardous) has its own covariance matrix. Although LDA classifier is less flexible than RDA, and therefore has lower variance, however, QDA is recommended when training dataset is very large so that variance becomes less of a concern.

Using predictor variables from Model 1 and Model 2, we first fit two QDA models onto training observations. Then we use these fitted models to predict seismic activity from test observations. In case of the predictor variables from Model 1, we have achieved significantly higher estimates of sensitivity (30.7%) than model outputs discussed before, with higher specificity (93.1%) and overall error rate 10.7%. However, Model 2 outputs better sensitivty (41%), slightly decreased specificity (86.8%), and slightly increased overall error rate (15.9%).

Table 10: Model-1 vs Model-2

	0	1	0	1
0	565	27	527	23

Table 11: Model-1 vs Model-2

	0	1	0	1
0	593	35	527	23
1	14	4	80	16
# Con	clusio	n and	Futur	e Work

At first we were not able to perform QDA and RDA because of the existence of multicollinearity in the data, which is partly(?) overcome by performing stepwise variable selection. Our final models (Model 1 nad Model 2) show that seismic hazard is related to previous mining shift's hazard status, seismic energy, number of pulses, number of seismic bumps recorded within previous shift, the number of seismic bumps (102-103 J) registered within previous shift, as well as the number of seismic bumps (102-103 J) registered within previous shift.

Our QDA results show that we have roughly 87% of chance to correctly predict that seismic hazard will not occur in the future. We have 41% of chance to predict seismic hazard ahead of time. Its evident that, among all the classification methods we have used on seismic-bumps dataset, the QDA method performs best in making prediction with relatively higher accuracy, which can be effective to prevent possible loss from seismic hazard, although future improvements in the model selection can be made. Further, research shows that in the seismic hazard assessment, data clustering techniques can be applied (Lesniak et al. 2009), and for the space-time prediction of seismic tremors, artificial neural networks are used (Kabiesz, 2005).

Summary of Models

Pre-Variable Selection

Model	Test Specificity	Test Sensitivity	Training Specificity	Training Sensitivity
Indicator	99.67%	0%	99.83%	0.76%
LDA	97.36%	12.8%	98.01%	17.56%
QDA	**	**	**	**
RDA	**	**	**	**
Log Regression	99.51%	0%	99.72%	4.58%

** QDA and RDA weren't performed on seismic-bumps data having high multi-collinearity.

Post-Variable Selection

Model 1	Test Specificity	Test Sensitivity	Training Specificity	Training Sensitivity
Indicator	**	**	**	**
LDA	**	**	**	**
QDA	93.08%	30.77%	***	***
RDA	97.69%	10.26%	***	***
Log Regression	99.83%	0%	99.78%	2.29%

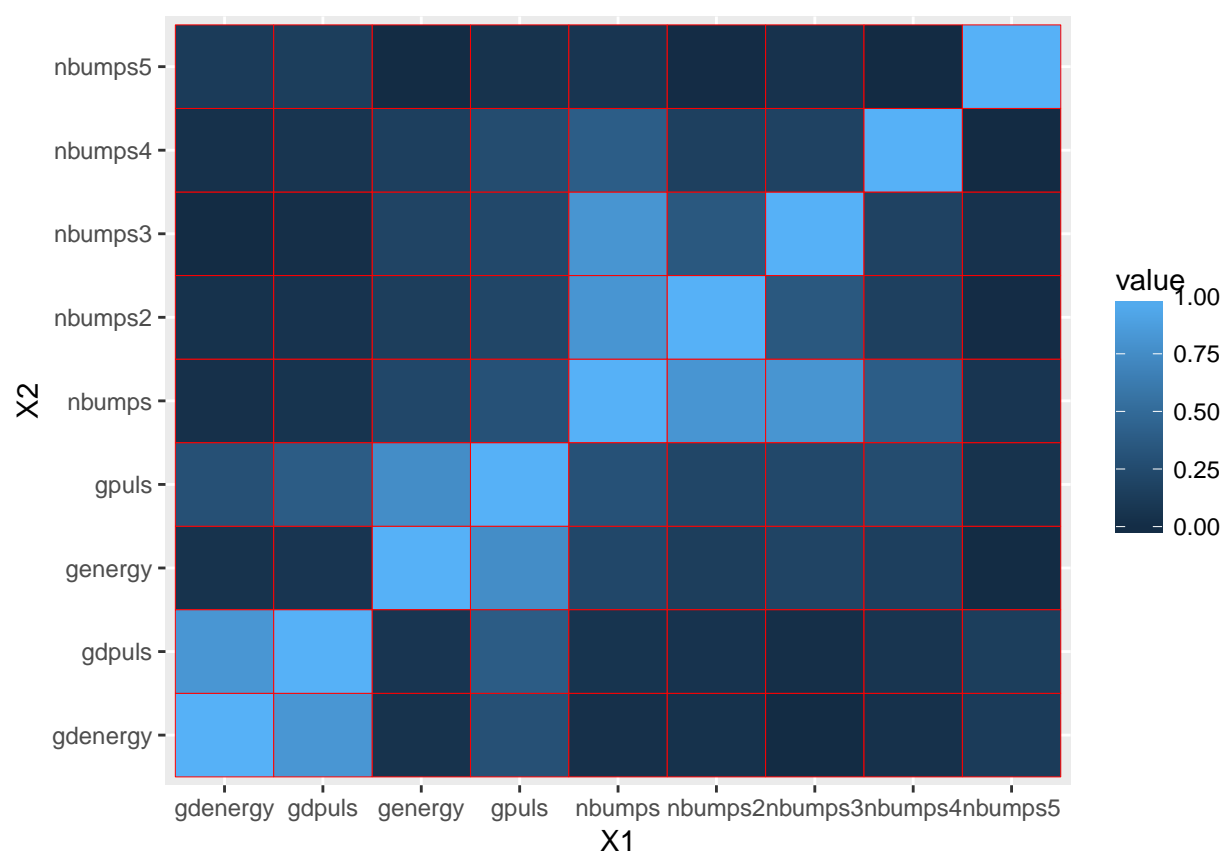| Model 2	Test Specificity	Test Sensitivity	Training Specificity	Training Sensitivity
Indicator	**	**	**	**
LDA	**	**	**	**
QDA	86.8%	41%	***	***
RDA	86.8%	41%	***	***
Log Regression	99.83%	0%	99.56%	3.05%

** Analysis wasn't performed after variable selection.
*** No Analysis was performed on Training observations.

References

1. Dozolme, P. (November 21,). Specific and Non-Specific Hazards in Underground Mines. Retrieved March 26, 2017, from https://www.thebalance.com/specific-and-non-specific-hazards-in-underground-mines-2367338
2. Sikora M., Wrobel L.: Application of rule induction algorithms for analysis of data collected by seismic hazard monitoring systems in coal mines. Archives of Mining Sciences, 55(1), 2010, 91-114.
3. Lesniak A., Isakow Z.: Space-time clustering of seismic events and hazard assessment in the Zabrze-Bielszowice coal mine, Poland. Int. Journal of Rock Mechanics and Mining Sciences, 46(5), 2009, 918-928.
4. Kabiesz, J.: Effect of the form of data on the quality of mine tremors hazard forecasting using neural networks. Geotechnical and Geological Engineering, 24(5), 2005, 1131-1147.

Appendix

Contribution of Group Members