STAT 557 - Project 2

Ben Straub, Hillary Koch, Jiawei Huang, Arif Masrur

3/27/2017

Data overview

Mining activity has long been associated to mining hazards, such as fire, flood, toxic contaminant and others. (Dozolme, P., 2016) Among these hazards, seismic hazard is the hardest detectable and predictable, in this respect it is comparable to an earthquake. (Sikora & Wr?bel, 2010) Minimizing loss from seismic hazard requires both advanced data gathering method and data analysis method. In recent years, more and more advanced seismic and seismoacoustic monitoring systems allow better and more timely data acquisition of rock mass processes. Still, the big disproportion between the number of low-energy seismic events and the number of high-energy phenomena (e.g. > 10^4J) makes traditional statistical analysis methods insufficient to make useful prediction. Machine learning are needed to achieve higher prediction accuracy within short time window.

In this project, we used seismic-bumps dataset provided by Sikora & Wr?bel (2010), found in the UCI Machine Learning Repository. This seismic-bumps dataset comes from two longwalls located in a coal mine in Poland and contains 2584 observations and 19 attributes. Each observation holds summary statement of about seismic activity in the rock mass within one shift (8 hours) (Sikora & Wr?bel, 2010). 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 (>104 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 > 10^4J was listed. The purpose is to find whether and how the other 18 variables can be used to determine the hazardous status of the mine.

Table 1. Attribute information of the seismic-bumps dataset (Sikora & Wróbel, 2010)

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
nbumps2	the number of seismic bumps (102-103 J) registered within previous shift	Continuous

Data Attributes	Description	Data Types
nbumps3	the number of seismic bumps (103-104 J) registered within previous shift	Continuous
nbumps4	the number of seismic bumps (104-105 J) registered within previous shift	Continuous
nbumps5	the number of seismic bumps (105-106 J) registered within previous shift	Continuous
nbumps6	the number of seismic bumps (106-107 J) registered within previous shift	Continuous
nbumps7	the number of seismic bumps (107-108 J) registered within previous shift	Continuous
nbumps8 and nbumps9	the number of seismic bumps (108-1010 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

2. Exploratory Data Analysis

The distribution of class variable suggests the complexity of seismic processes, which can be seen from the big disproportion between the number of low-energy seismic events and the number of high-energy phenomena (e.g. $>$ 10ˆ4J) : in 2584 records, only 170 has show the value 1. Each of the predictor variables (e.g., seismic hazard state, seismoacoustic hazard state, shift, seismic energy, puls, number of bumps) represent measurements of seismic activity during each shift. Some predictor variables are stored as categorical factors and some are continuous (see Table 1).

Since the seismic-bumps dataset involves predicting a qualitative response values, we want to employ widely-used classification techniques such as logistic regression and discriminant analysis methods for the prediction of future seismic hazards. In other words, we want to examine which observations of seismic activities in multiple shifts can potentially lead to the hazardous or non-hazardous states in the next shift. Both logistic regression and discriminant analysis methods assume predictors to be normally distributed. Therefore, we examine the distribution of predictor variables and found that data is right skewed. We also found the existence of severe multicollinearity (VIF>10) among several variables.

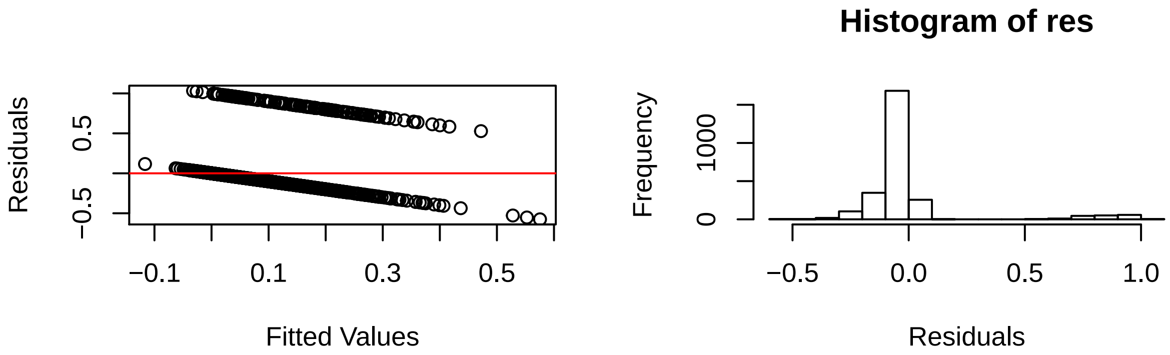


Table 2: VIFs of Linear Model

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.

Table 4: Training and Test Dimensions

	Training	Test
Obs	1938	646
Varialbes	16	16

	Training	Test

Linear Regression of an Indicator Matrix

Since our response variable has two classes (e.g., hazardous vs non-hazardous states), we start with linear regression of an indicator matrix as this method approximates a linear decision boundary among observations belonging to these classes. Our model outputs overall error rate of 6.7% with sensitivity 0% and specificity 100%. That essentially means, while the model has lower overall error rate, it has only 1% chance of predicting hazard state in the next shift, whereas in 99% of the cases it successfully predicted non-hazardous state.

Table 5: Training vs Test

	A	B	C	D
A	1804	130	605	39
B	3	1	2	0

Logistic Regression:

We first fit a logistic regression model to predict “class” using all the predictors in the training dataset. It appeasers that seismic, shift, and gpuls have positive coefficients and statistically significant p-values. Therefore, seismic, shift, gpuls have clear positive association with the hazardous/non-hazardous seismic activity. Then we compute the probabilities for the training observations to predict which observations correspond to the hazardous or non-hazardous seismic activity. For that, we used a threshold probability of 0.5. The confusion matrix on training data shows that while this logistic model slightly outperformed LDA (training) model in terms of overall error rate (6.7%), it has substantially lower sensitivity (4.6%) than that of LDA (17.6%). Our fitted logistic model on the training dataset has even poorer performance (see Figure 4) in making prediction on the test dataset (overall error rate: 6.6%; sensitivity: 0%,; specificity: 100%).

Table 6: Training vs Test

	0	1	0	1
0	1802	125	604	39
1	5	6	3	0

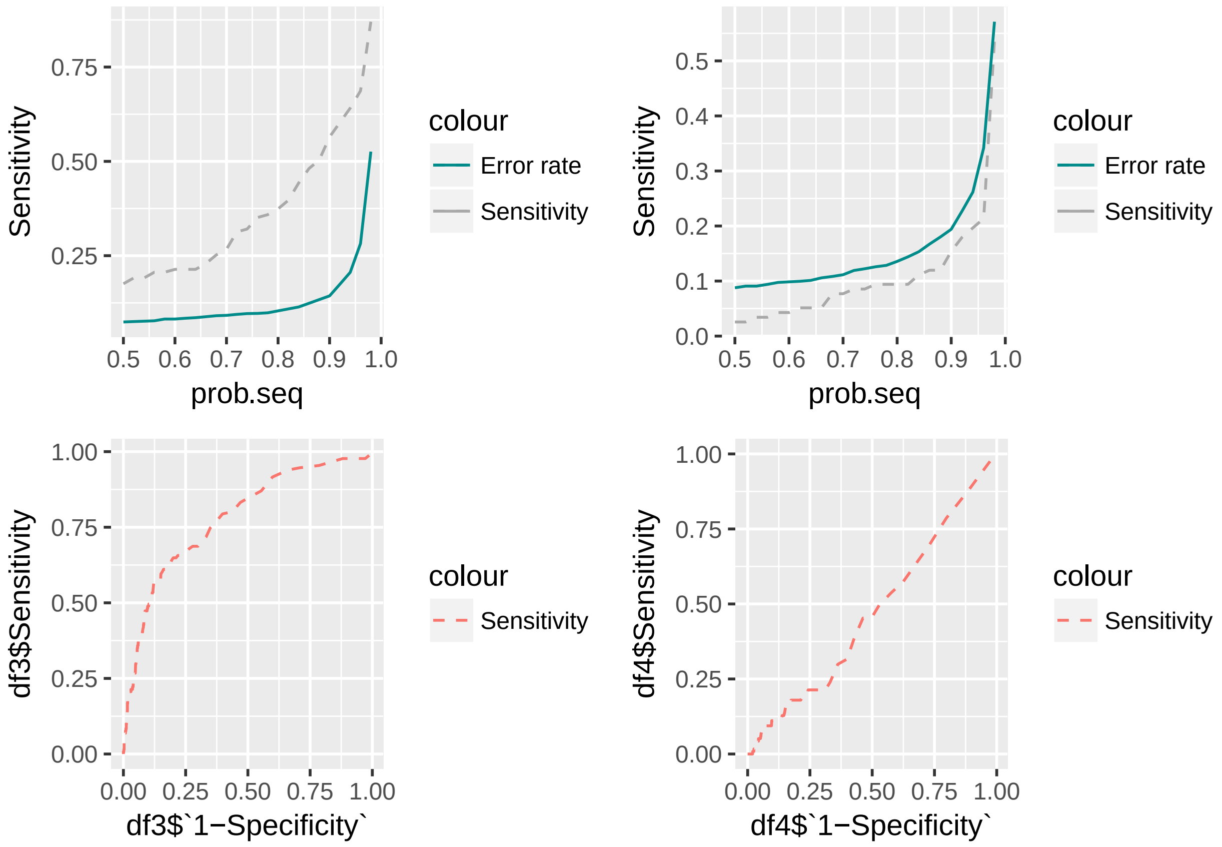
Linear Discriminant Analysis

We perform LDA on both training and test datasets. First we fit a LDA model using observations from training dataset. The LDA outputs indicate that 93.2% of the training observations correspond to the non-hazardous state in the next shift of mining activity. The group means for the estimation of ?k suggests that previous mining shift’s higher number of seismic bumps and associated higher released energy (measured in Joule) from seismic bumps can be related to the hazardous state in the later mining activity. Using this fitted model parameter we obtain two level of predictions, first onto training data, and then onto test data.

The confusion matrices obtained from training and test data show slightly different outputs in terms of sensitivity, specificity, and overall error rate (see Figure 2 and 3). It appears that training dataset produced slightly improved LDA model (overall error rate: 7.4%; sensitivity: 17.6%; specificity: 98%) than that of the test dataset (error rate: 7.8%; sensitivity: 12.8%; specificity: 97%).

Table 7: Training vs Test

	0	1	0	1
0	1771	108	591	34
1	36	23	16	5



Regularized Discriminant Analysis

Table 8: Training vs Test

	0	1	0	1
0	1785	126	593	39
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 section have not been able to approximate better the relationship between response and predictor variables. The strong multicollinearity among some of the predictor variables found in the EDA (see section 2) may have contributed to the high error rate and lower interpretability in the resulting models. Therefore, in this section, for the improvement of prediction accuracy and model interpretability, we employ some of the commonly used variable selection methods: stepwise subset selection, shrinkage, and dimensionality reduction.

Stepwise Variable Selection

We use forward and backward selection. We did further manual selection, and chose the model that produced the lowest AIC score. The model with least AIC score resulted following subset of 5 predictor variables: genergy, gpuls, nbumps, nbump2, and nbumps4. We denote the model with these variables as Model 1.

LASSO

We perform Least Absolute Shrinkage and Selection Operator (LASSO) regression, which fits a model by shrinking some of the coefficients toward exactly zero. LASSO is expected to perform well for the seismic-bumps dataset as some of the predictors aren't related to the response, as we found in the EDA process described in section 2. After performing LASSO, we found that, 11 of total 15 coefficient estimates of predictor variables are exactly 0. The 4 variables with substantial coefficients are: seismic, shift, gpuls, and nbumps. These variables are incorporated in the model, which we denote as Model 2.

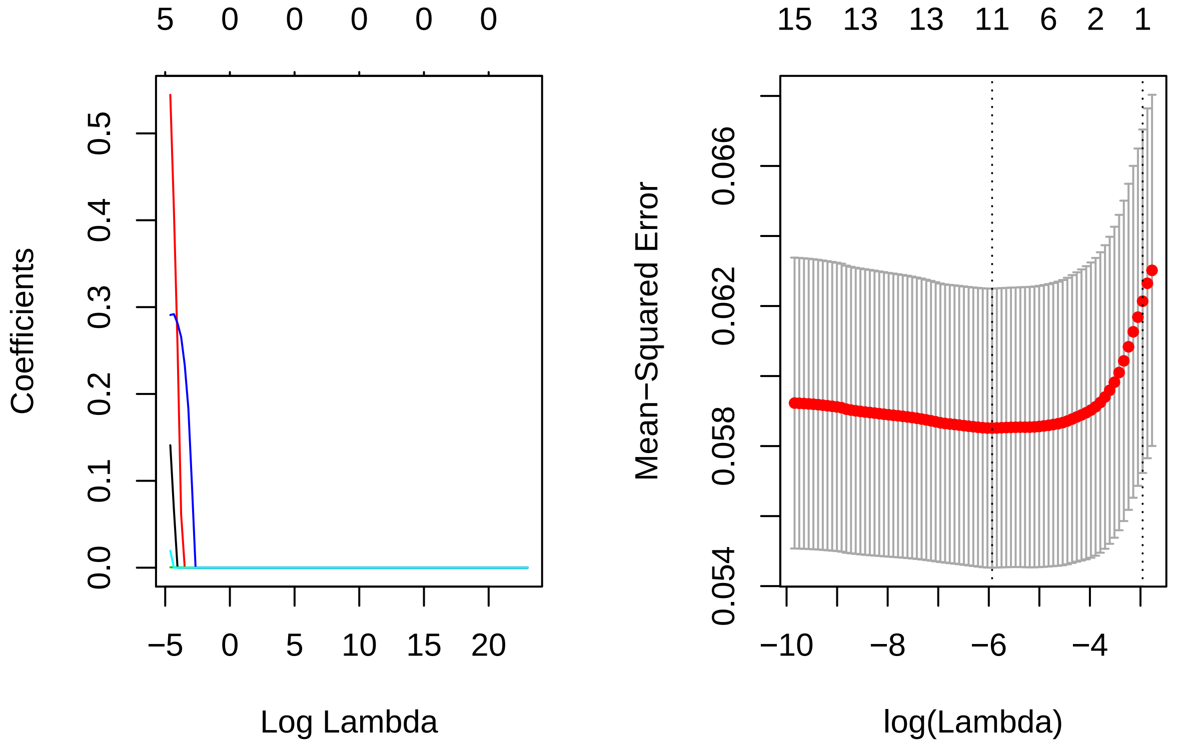


Table 9: Model 2 through Lasso

Intercept	Seismic	Shift	Gpuls	nbumps
-0.00814	0.0088	0.00798	5e-05	0.03118

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 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 10: Model-1 vs Model-2

	0	1	0	1
0	1803	128	606	39
1	4	3	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 sensitvity (41%), slightly decreased specificity (86.8%), and slightly increased overall error rate (15.9%).

Table 11: Model-1 vs Model-2

	0	1	0	1
0	565	27	527	23
1	42	12	80	16

Regularized Discriminant Analysis after variable selection

RDA makes trade-off between LDA and QDA by shrinking the separate covariances of QDA toward a common covariance as in LDA. We performed RDA for Model 1 on the test dataset with varying set of lambda values. It appears that with lambda values between 0-1, sensitivity estimates resulting from RDA varies inversely with increasing lambda values. That means, if lambda = 0, RDA results in similar higher sensitivity value as QDA does, whereas lambda = 1 results in lower sensitivity value similar to LDA.

*** Look at gamma and lambda values. Make changes according to what make senses. Writing for this section will follow those changes.

Table 12: 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%

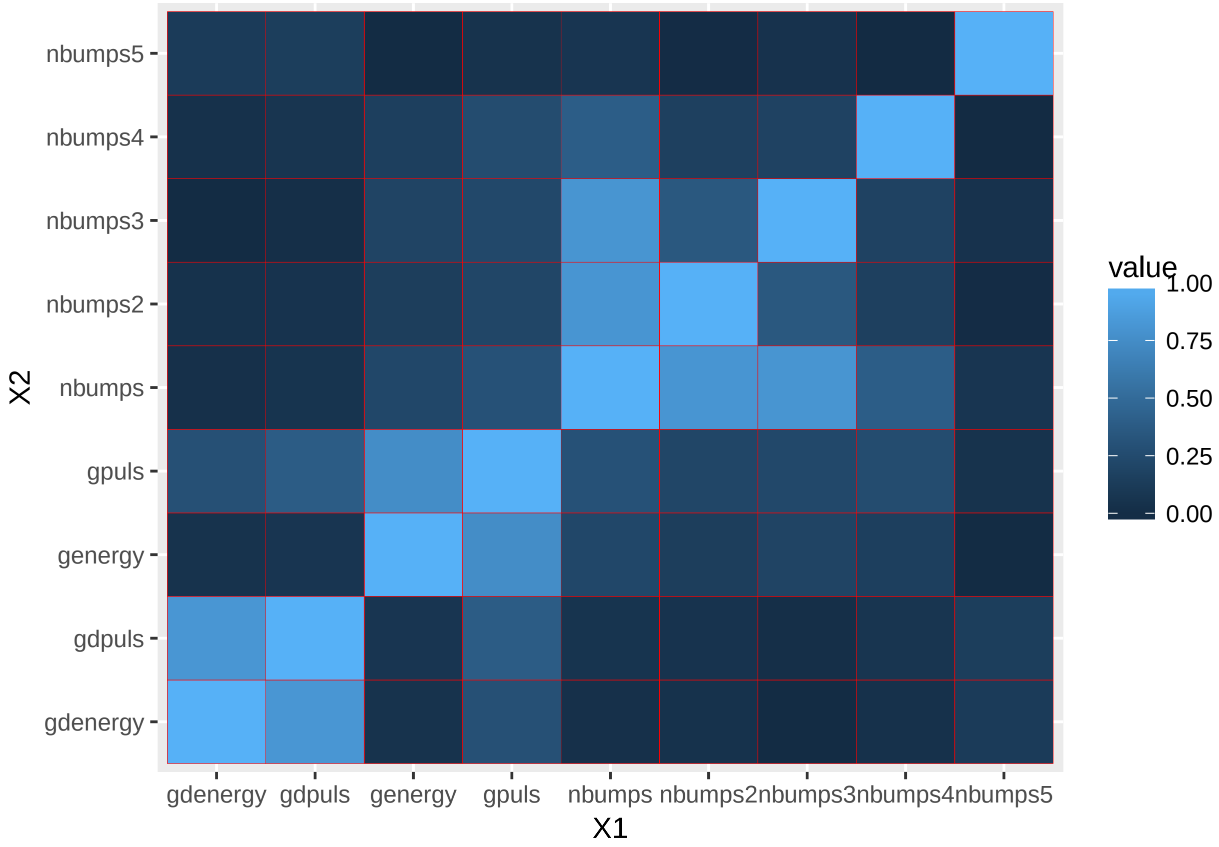
- QDA and RDA weren't performed on seismic-bumps data having high multi-collinearity.

Post-Variable Selection

Model	Test Specificity	Test Sensitivity	Training Specificity	Training Sensitivity
Indicator	*	*	*	*
LDA	123	*	*	*
QDA	86.8%	41%	**	**
RDA	86.8%	41%	**	**
Log Regression	99.83%	0%	99.78%	2.29%

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

Appendix



Contribution of Group Members