557_markdown_draft

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. $> 10^4$J) 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 $> 10^4$J 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.

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. $> 10^4$J) : 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.

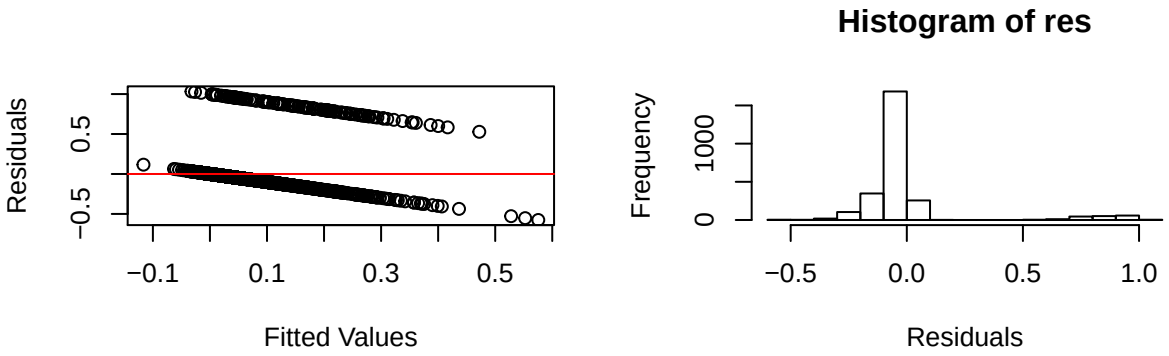


Table 1: VIFs of Linear Model

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

Table 2: 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 3: Training and Test Dimensions

	Training	Test
Obs	1938	646
Varialbes	16	16

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 4: Training vs Test

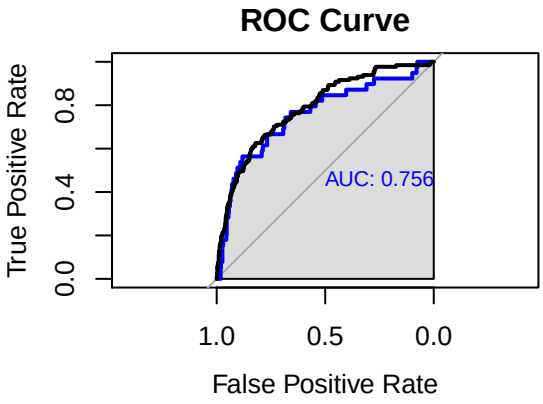
	A	B	C	D
A	1804	130	605	39
B	3	1	2	0

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 5: Training vs Test

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



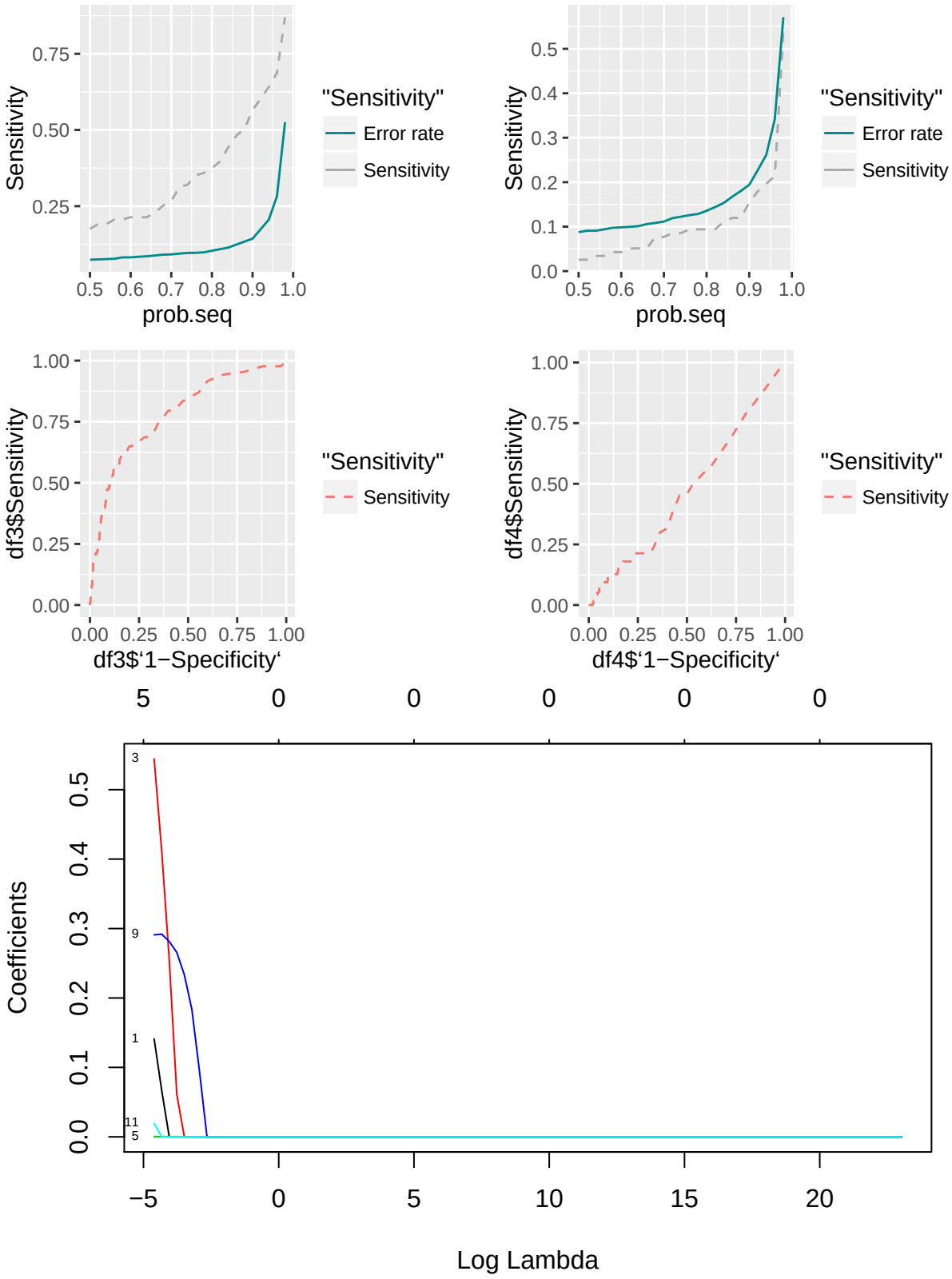
Linear Discriminant Analysis

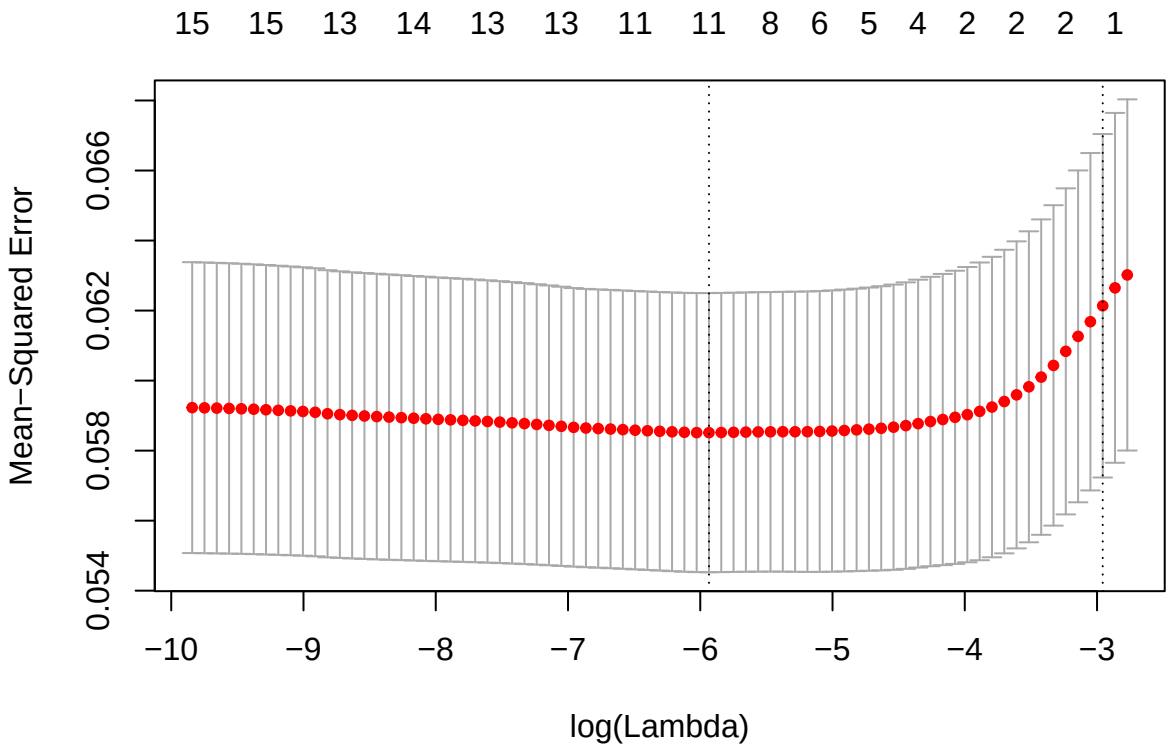
We perform LDA on both training and test datasets. First we fit a LDA model using observations from training dataset. The LDA outputs indicate that 93.2% of the training observations correspond to the non-hazardous state in the next shift of mining activity. The group means for the estimation of ?k suggests 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 6: Training vs Test

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


Sensitivity
prob.seq
"Sensitivity"
Error rate
Sensitivity
df3$Sensitivity
df3$‘1−Specificity‘
df4$Sensitivity
df4$‘1−Specificity‘
Coefficients
Log Lambda




```
[1] 8.670049
```

```
  (Intercept) seismic   shift gpuls  nbumps
1    -0.00814  0.0088 0.00798 5e-05 0.03118
```

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: shift, gpuls, and nbumps. These variables are incorporated in the model, which we denote as 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
```

Appendix