Report on my own edx project

FUNG CHE HEI

8/19/2020

1. Introduction

1.1 Background

There are many valuable minerals and geological materials in the Earth that are stored in the form of some kinds of ores, lode, vein and reef. Mining is the human activities to extract those of them. Although mining activity seems interesting for human to discover different kinds of minerals in the Earth, it is a dangerous activities because a hazard of seismic bumps would occurs in many underground mines. Inaccurate prediction and detection would cause great damage to human life.

Therefore a good seismic hazard assessment is important and required for mining activities. With the aid of machine learning technologies, some research including clustering [1] and artificial neural networks [2] are used for prediction of seismic tremors in the past years.

1.2 Aim

Our main aim is:

- to forecast the high energy seismic bumps (higher than 10^4 J) in mine

With predicting the possibility of the occurrence of hazardous situation, appropriate risk assement and supervision service can be made. For example, reducing the risk of rockburst by the use of distressing shooting method and withdrawing workers from the threatened area.

1.3 Data set Information

The data set used is called "seismic-bumps Data Set" which is downloaded from UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/seismic-bumps#

Here we read the downloaded data set and call it 'seismic'.

```
seismic <- as.data.frame(read_csv("data/seismic_bumps.csv"))
```

An overview on the seismic-bumps Data Set

```
nrow(seismic)
```

```
## [1] 2584
```

```
ncol(seismic)
```

```
## [1] 20
```

```
head(seismic)
```

```
##   id seismic seismoacoustic shift genergy gpuls gdenergy gdpuls ghazard nbumps
## 1  1       a              a     N   15180    48      -72    -72       a      0
## 2  2       a              a     N   14720    33      -70    -79       a      1
## 3  3       a              a     N    8050    30      -81    -78       a      0
## 4  4       a              a     N   28820   171      -23     40       a      1
## 5  5       a              a     N   12640    57      -63    -52       a      0
## 6  6       a              a     W   63760   195      -73    -65       a      0
##   nbumps2 nbumps3 nbumps4 nbumps5 nbumps6 nbumps7 nbumps89 energy maxenergy
## 1       0       0       0       0       0       0        0      0         0
## 2       0       1       0       0       0       0        0   2000      2000
## 3       0       0       0       0       0       0        0      0         0
## 4       0       1       0       0       0       0        0   3000      3000
## 5       0       0       0       0       0       0        0      0         0
## 6       0       0       0       0       0       0        0      0         0
##   class
## 1     0
## 2     0
## 3     0
## 4     0
## 5     0
## 6     0
```

After having a quick look on the data set, there are 2584 rows (observations) and 20 columns (attributes). Each observation contains a summary statement about seismic activity in the rock mass within one shift (8 hours) which will be described in section 1.4, to predict 'hazardous' (positive class with value = 1) and 'non-hazardous' (negative class with value = 0) states. If 'hazardous' is predicted, it is possibly that seismic bump with an energy higher than 10^4 J would occur in the next shift.

Here note the there is unbalanced distribution of positive and negative class. Among 2584 observations, only 170 of them are positive class.

```
sum(seismic$class == 1)
```

```
## [1] 170
```

1.4 Arributes Information

- 1. seismic: result of shift seismic hazard assessment in the mine working obtained by the seismic method (a - lack of hazard, b - low hazard, c - high hazard, d - danger state);
- 2. seismoacoustic: result of shift seismic hazard assessment in the mine working obtained by the seismoacoustic method;
- 3. shift: information about type of a shift (W - coal-getting, N -preparation shift);

- 4. genergy: seismic energy recorded within previous shift by the most active geophone (GMax) out of geophones monitoring the longwall;
- 5. gpuls: a number of pulses recorded within previous shift by GMax;
- 6. gdenergy: a deviation of energy recorded within previous shift by GMax from average energy recorded during eight previous shifts;
- 7. gdpuls: a deviation of a number of pulses recorded within previous shift by GMax from average number of pulses recorded during eight previous shifts;
- 8. ghazard: result of shift seismic hazard assessment in the mine working obtained by the seismoacoustic method based on registration coming from GMax only;
- 9. nbumps: the number of seismic bumps recorded within previous shift;
- 10. nbumps2: the number of seismic bumps (in energy range [10^2, 10^3)) registered within previous shift;
- 11. nbumps3: the number of seismic bumps (in energy range [10^3, 10^4)) registered within previous shift;
- 12. nbumps4: the number of seismic bumps (in energy range [10^4, 10^5)) registered within previous shift;
- 13. nbumps5: the number of seismic bumps (in energy range [10^5, 10^6)) registered within the last shift;
- 14. nbumps6: the number of seismic bumps (in energy range [10^6, 10^7)) registered within previous shift;
- 15. nbumps7: the number of seismic bumps (in energy range [10^7, 10^8)) registered within previous shift;
- 16. nbumps89: the number of seismic bumps (in energy range [10^8, 10^10)) registered within previous shift;
- 17. energy: total energy of seismic bumps registered within previous shift;
- 18. maxenergy: the maximum energy of the seismic bumps registered within previous shift;
- 19. class: the decision attribute - '1' means that high energy seismic bump occurred in the next shift ('hazardous state'), '0' means that no high energy seismic bumps occurred in the next shift ('non-hazardous state').

1.5 Variables Information

There are totally 18 input variables (attributes) and 1 binary output variable (class) in the data set. The below table summarize some information of the variables.

Table 1: Variable Summary Table

variable	Cardinality	Filled	Nulls	Total	Uniqueness
class	2	2584	0	2584	0.0
energy	242	2584	0	2584	0.1
gdenergy	334	2584	0	2584	0.1
gdpuls	292	2584	0	2584	0.1
genergy	2212	2584	0	2584	0.9

variable	Cardinality	Filled	Nulls	Total	Uniqueness
ghazard	3	2584	0	2584	0.0
gpuls	1128	2584	0	2584	0.4
id	2584	2584	0	2584	1.0
maxenergy	33	2584	0	2584	0.0
nbumps	10	2584	0	2584	0.0
nbumps2	7	2584	0	2584	0.0
nbumps3	7	2584	0	2584	0.0
nbumps4	4	2584	0	2584	0.0
nbumps5	2	2584	0	2584	0.0
nbumps6	1	2584	0	2584	0.0
nbumps7	1	2584	0	2584	0.0
nbumps89	1	2584	0	2584	0.0
seismic	2	2584	0	2584	0.0
seismoacoustic	3	2584	0	2584	0.0
shift	2	2584	0	2584	0.0

Although 'maxenergy' and 'nbumps' are numeric data representing the magnitude of energy and number of bumps respectively, they have a relatively small cardinality which result in zero Uniqueness (defined by the ratio of Cardinality to Total). Therefore they are classified into categorical variables. For those variables with uniqueness greater than zero are then classified as numeric variables. Below is the table that summarizing the variable type of class and each attributes.

Table 2: Variable Type

Variable	Type
class	binary
energy	numeric
gdenergy	numeric
gdpuls	numeric
genergy	numeric
ghazard	catagorical
gpuls	numeric
maxenergy	catagorical
nbumps	catagorical
nbumps2	catagorical
nbumps3	catagorical
nbumps4	catagorical
nbumps5	catagorical
nbumps6	catagorical
nbumps7	catagorical
nbumps89	catagorical
seismic	catagorical
seismoacoustic	catagorical
shift	catagorical

1.6 Key Steps

2. Data Analysis

2.1 Data Cleaning

2.1.1 Present of Nulls

Refer to Table 1 in section 1.5, there is no Null value in the data set therefore removing of those null values is not required.

2.1.2 Statistic

Statistic of attributes is presented as follow:

```
summary(seismic)
```

```
##        id           seismic          seismoacoustic        shift
##  Min.   :   1.0   Length:2584        Length:2584        Length:2584
##  1st Qu.: 646.8   Class :character   Class :character   Class :character
##  Median :1292.5   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1292.5
##  3rd Qu.:1938.2
##  Max.   :2584.0
##     genergy            gpuls           gdenergy           gdpuls
##  Min.   :    100   Min.   :   2.0   Min.   : -96.00   Min.   :-96.000
##  1st Qu.:  11660   1st Qu.: 190.0   1st Qu.: -37.00   1st Qu.:-36.000
##  Median :  25485   Median : 379.0   Median :  -6.00   Median : -6.000
##  Mean   :  90242   Mean   : 538.6   Mean   :  12.38   Mean   :  4.509
##  3rd Qu.:  52832   3rd Qu.: 669.0   3rd Qu.:  38.00   3rd Qu.: 30.250
##  Max.   :2595650   Max.   :4518.0   Max.   :1245.00   Max.   :838.000
##    ghazard             nbumps          nbumps2          nbumps3
##  Length:2584        Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  Class :character   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Mode  :character   Median :0.0000   Median :0.0000   Median :0.0000
##                     Mean   :0.8595   Mean   :0.3936   Mean   :0.3928
##                     3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##                     Max.   :9.0000   Max.   :8.0000   Max.   :7.0000
##     nbumps4           nbumps5            nbumps6     nbumps7     nbumps89
##  Min.   :0.00000   Min.   :0.000000   Min.   :0   Min.   :0   Min.   :0
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0   1st Qu.:0   1st Qu.:0
##  Median :0.00000   Median :0.000000   Median :0   Median :0   Median :0
##  Mean   :0.06772   Mean   :0.004644   Mean   :0   Mean   :0   Mean   :0
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0   3rd Qu.:0   3rd Qu.:0
##  Max.   :3.00000   Max.   :1.000000   Max.   :0   Max.   :0   Max.   :0
##      energy         maxenergy          class
##  Min.   :     0   Min.   :     0   Min.   :0.00000
##  1st Qu.:     0   1st Qu.:     0   1st Qu.:0.00000
##  Median :     0   Median :     0   Median :0.00000
##  Mean   :  4975   Mean   :  4279   Mean   :0.06579
##  3rd Qu.:  2600   3rd Qu.:  2000   3rd Qu.:0.00000
##  Max.   :402000   Max.   :400000   Max.   :1.00000
```

Refer to the above summary and looking at attributes 'nbumps6', 'nbumps7' and 'nbumps89', it is observed all the values are zero which means those of them do not provide any information for classifying positive and

negative class. Therefore, we remove 'nbumps6', 'nbumps7' and 'nbumps89' from the entire data set. For the attribute 'id', it can be regarded as primary key of the data set and do not use for binary classification.

2.1.3 Correctness

It is obvious that the total number of seismic bumps (nbumps) equals to the sum of seismic bumps with different energy levels (nbumps2 + nbumps3 + ... + nbumps7 + nbumps89) and they should have no difference. The below code test this fact to ensure the correctness of the data set.

```
#test the correctness of the data set
seismic %>%
  mutate(total = nbumps2+nbumps3+nbumps4+nbumps5+nbumps6+nbumps7+nbumps89) %>%
  mutate(diff = total - nbumps) %>%
  filter(diff!=0) %>%
  summarize(n=n()) %>%
  pull(n)
```

```
## [1] 2
```

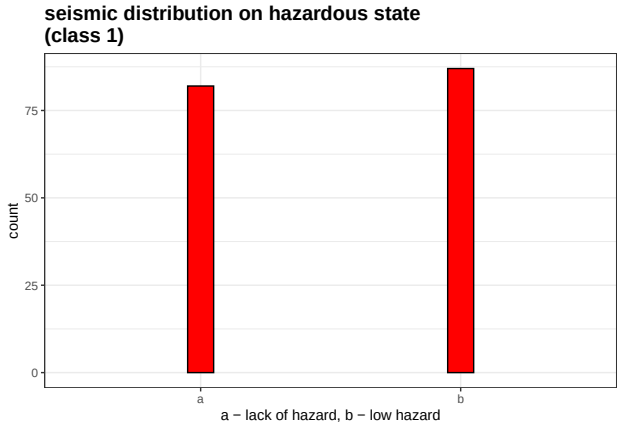
From the above result we can see that two observations suffer from the problem of inconsistency of the number of seismic bumps. Therefore these two observations will be removed from the entire data set and we call the corrected data set 'corrected_seismic'.

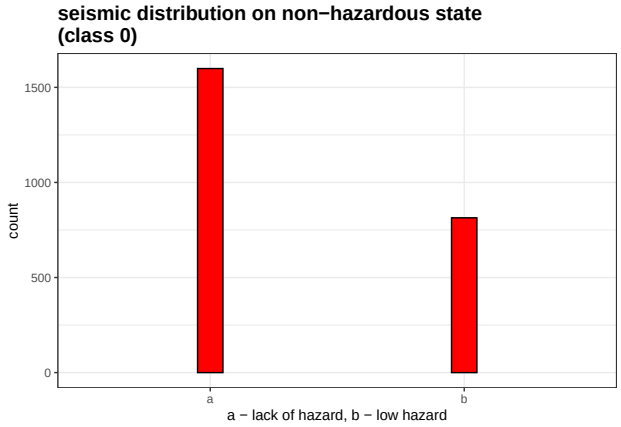
```
#extract the index of incorrect data set
incorrect_index<-
  seismic %>%
  mutate(total = nbumps2+nbumps3+nbumps4+nbumps5+nbumps6+nbumps7+nbumps89) %>%
  mutate(diff = total - nbumps) %>%
  filter(diff!=0) %>%
  pull(id)

#filter out the incorrect observations
#correct the data set
corrected_seismic<-
  seismic %>%
  filter(id!=incorrect_index) %>%
  select(-nbumps6,-nbumps7,-nbumps89,-id)
```

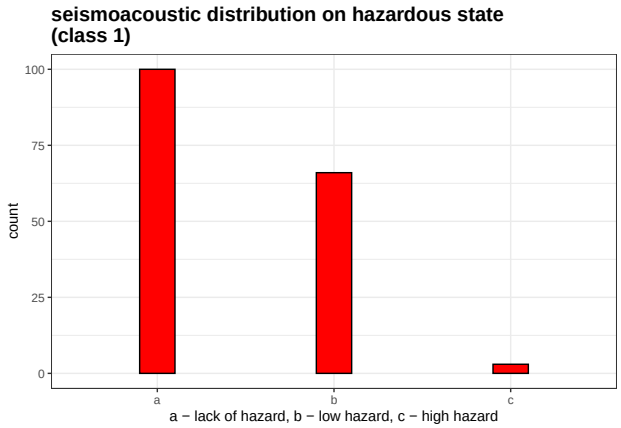
2.2 Data Exploration and Visualization

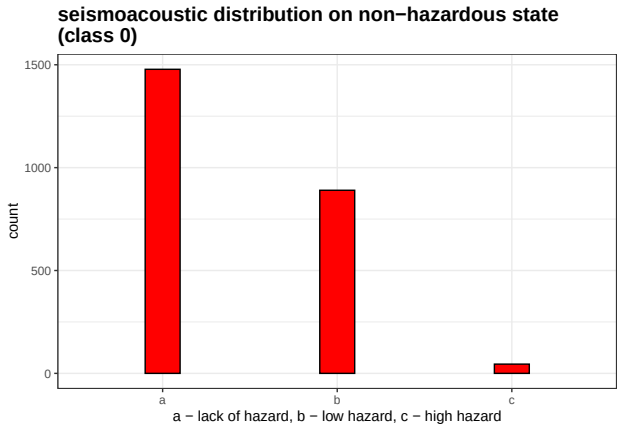
2.2.1 Ratio between seismic (result of shift seismic hazard assessment) on positive and negative class



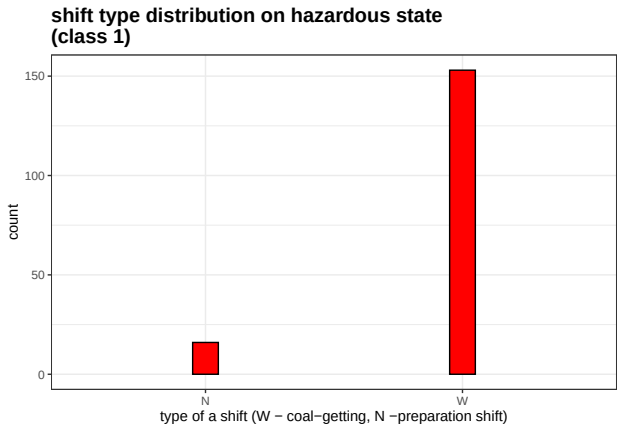


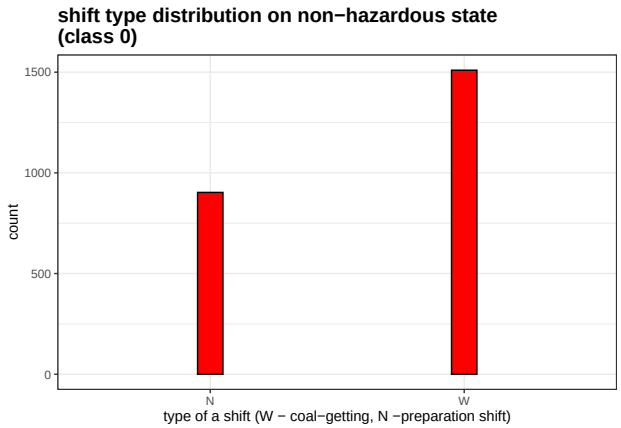
2.2.2 Ratio between seismoacoustic (result of shift seismic hazard assessment) on positive and negative class



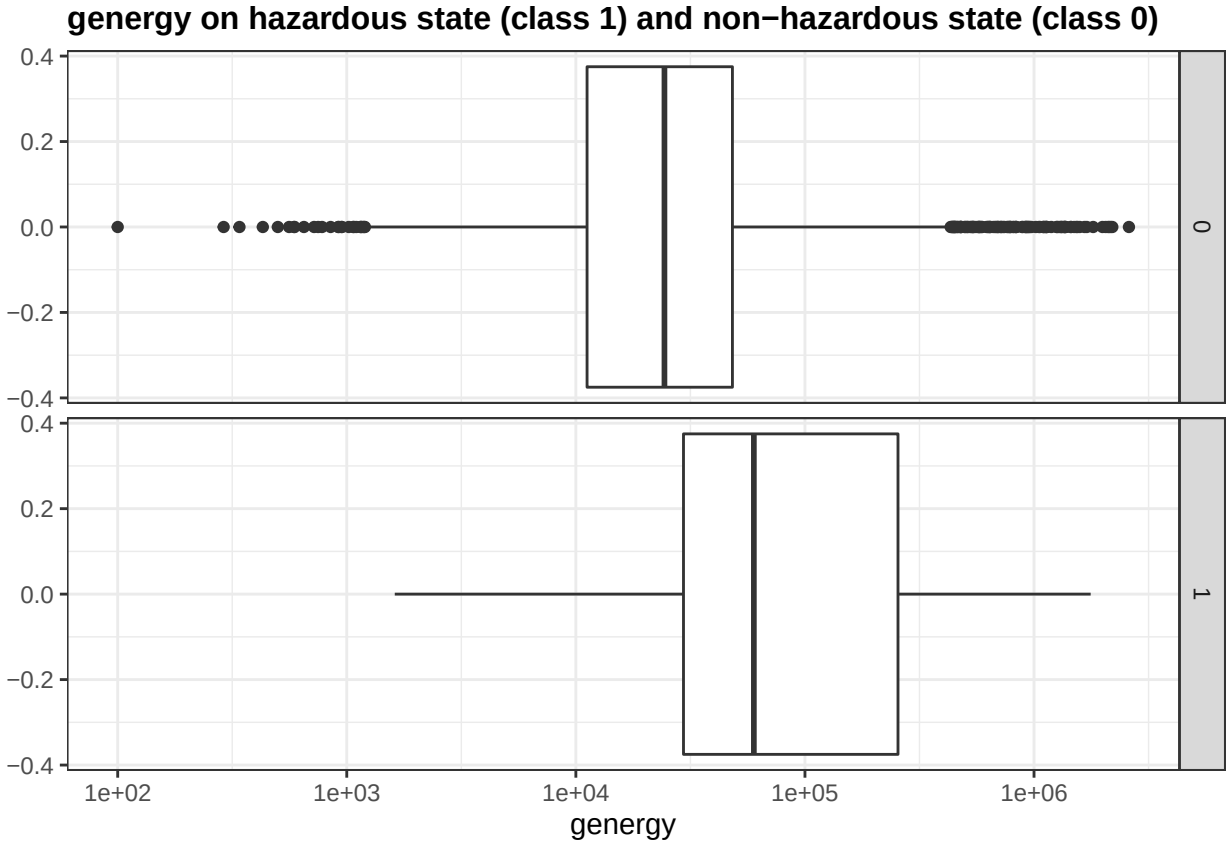


2.2.3 Ratio between shift on positive and negative class





2.2.4 genergy on positive class and negative class

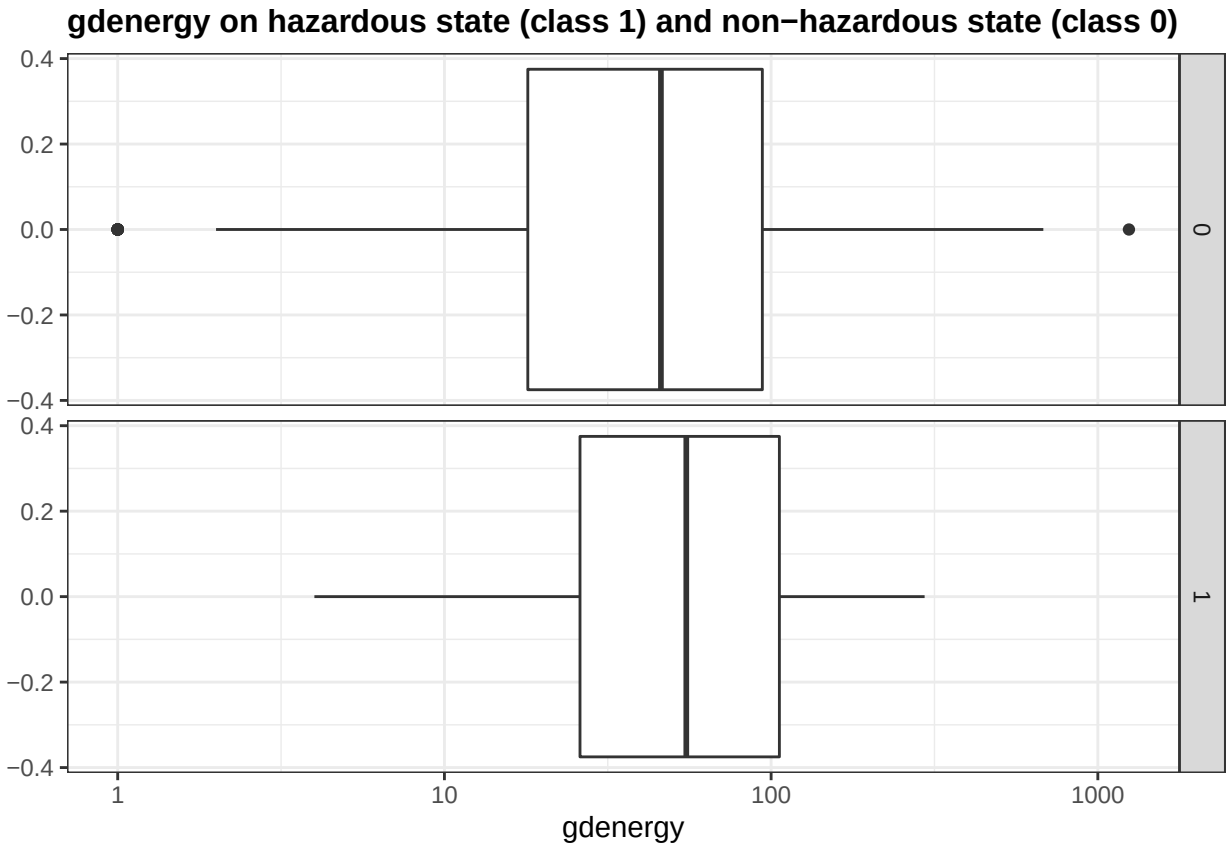


2.2.5 gdenergy on positive class and negative class

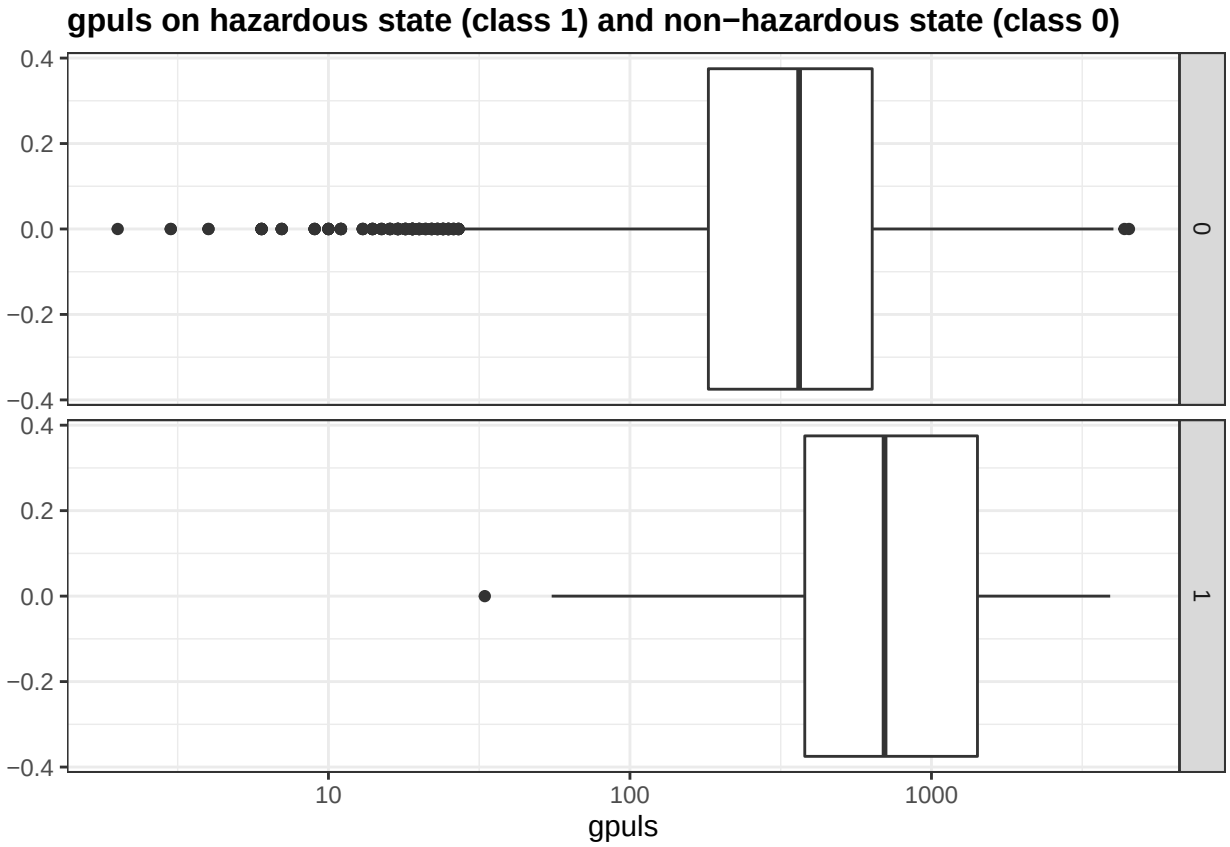
```
## Warning in self$trans$transform(x): NaNs produced
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 1412 rows containing non-finite values (stat_boxplot).
```


gdenergy on hazardous state (class 1) and non−hazardous state (class 0)
0.4
0.2
0.0
−0.2
−0.4
0
1
1
10
100
1000
gdenergy

2.2.6 gpuls on positive class and negative class

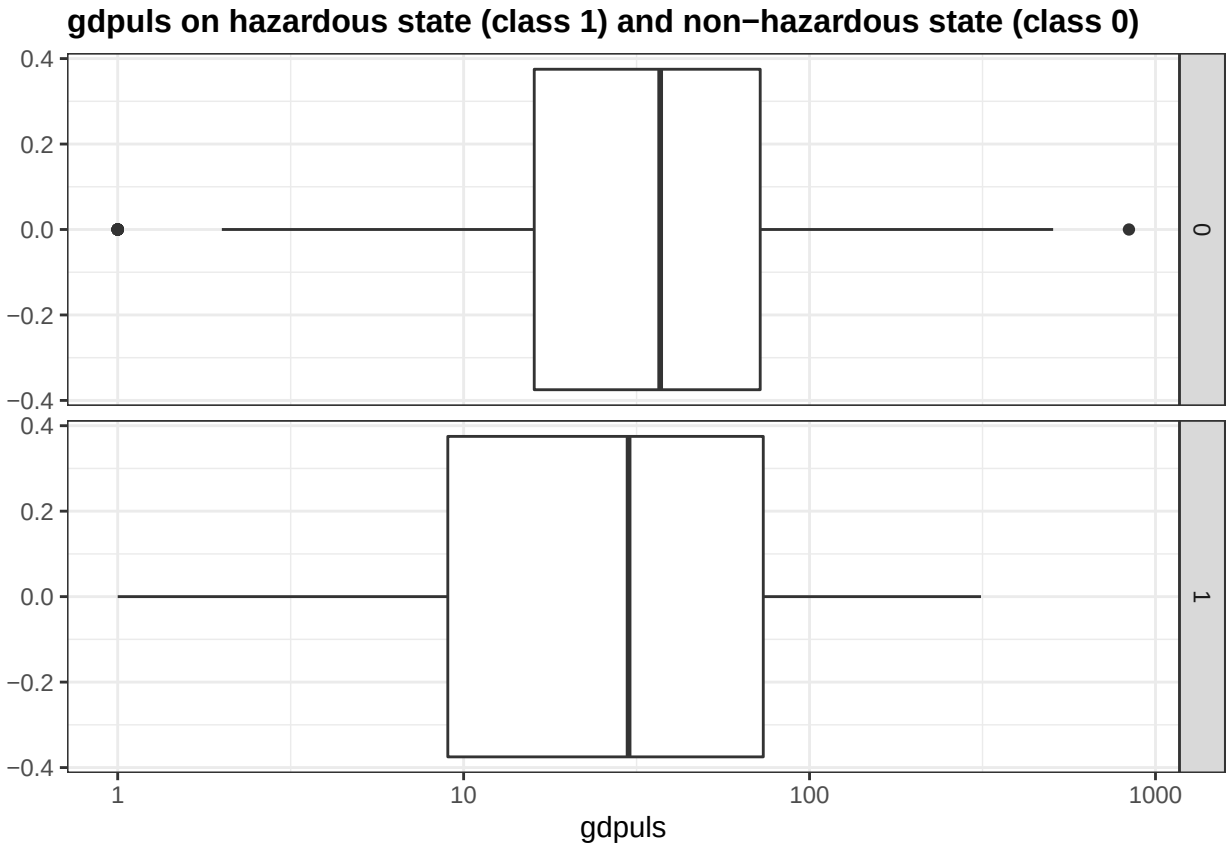


2.2.7 gdpuls on positive class and negative class

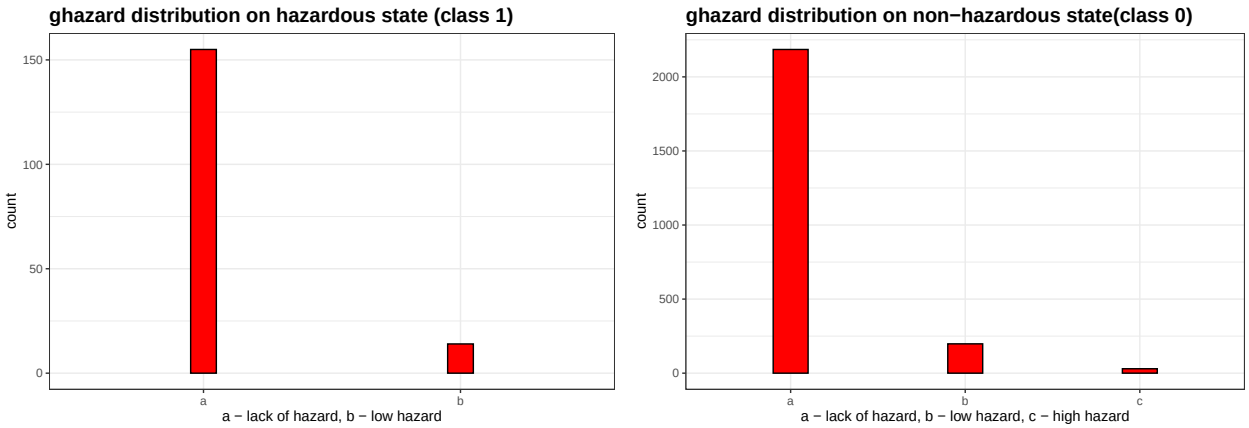
```
## Warning in self$trans$transform(x): NaNs produced
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

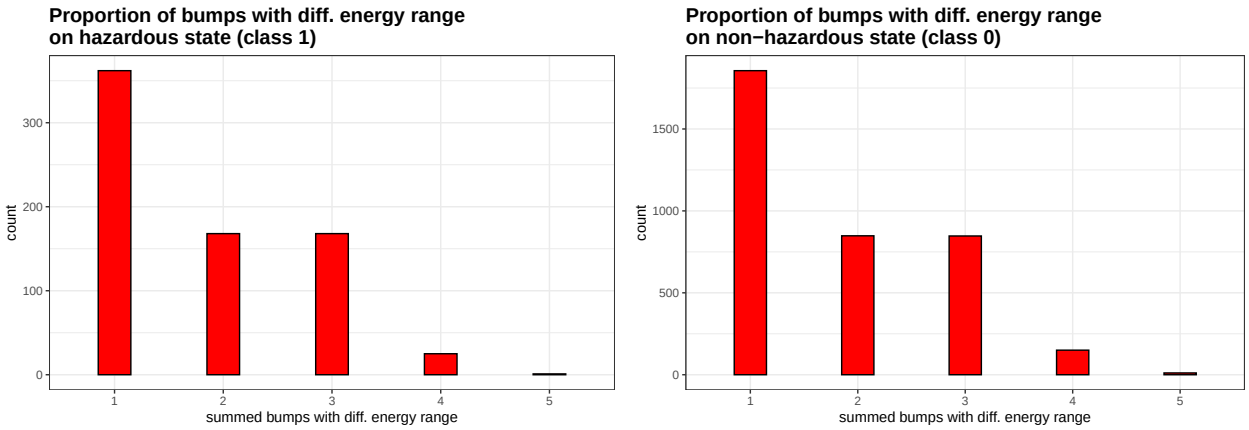
```
## Warning: Removed 1437 rows containing non-finite values (stat_boxplot).
```



2.2.8 Ratio between ghazard on positive class and negative class



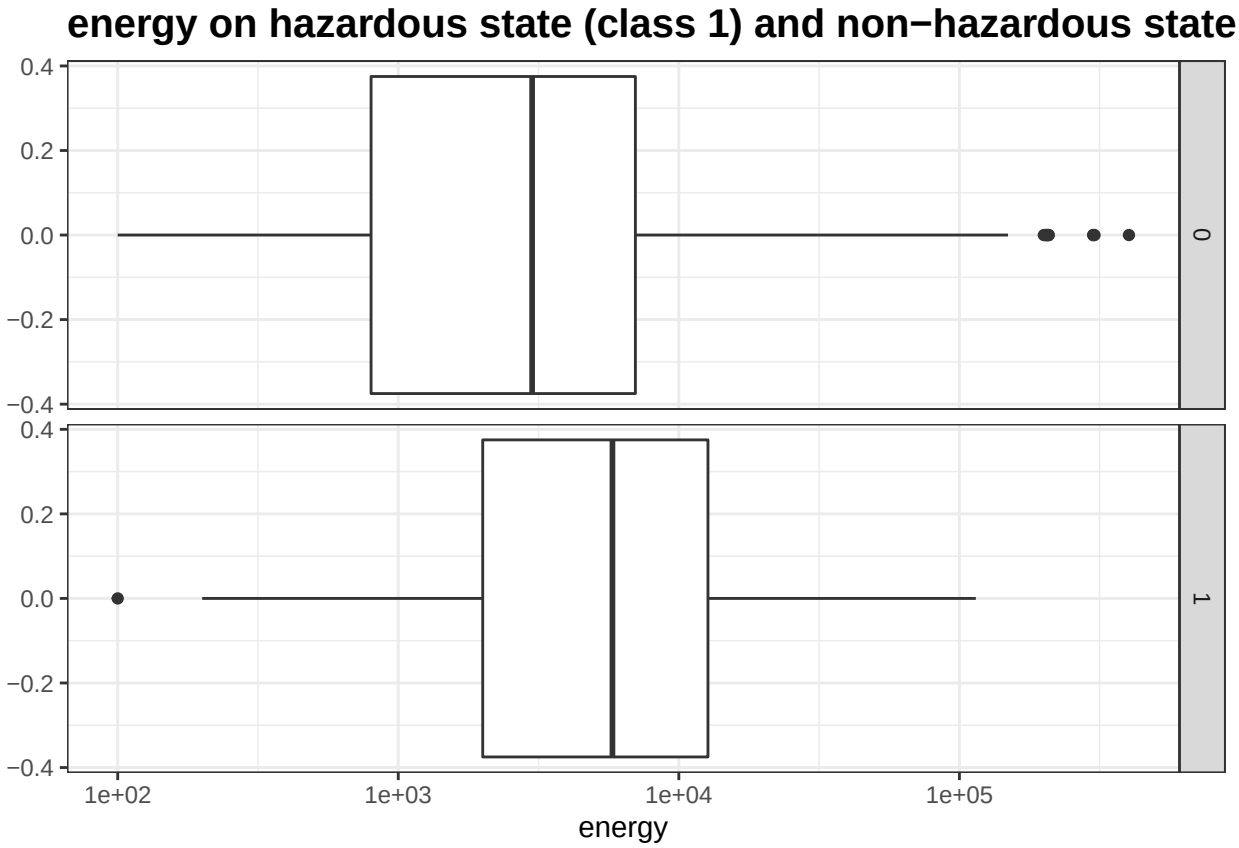
2.2.9 Effects by number of bumps with diff. energy range



2.2.10 energy on positive class and negative class

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

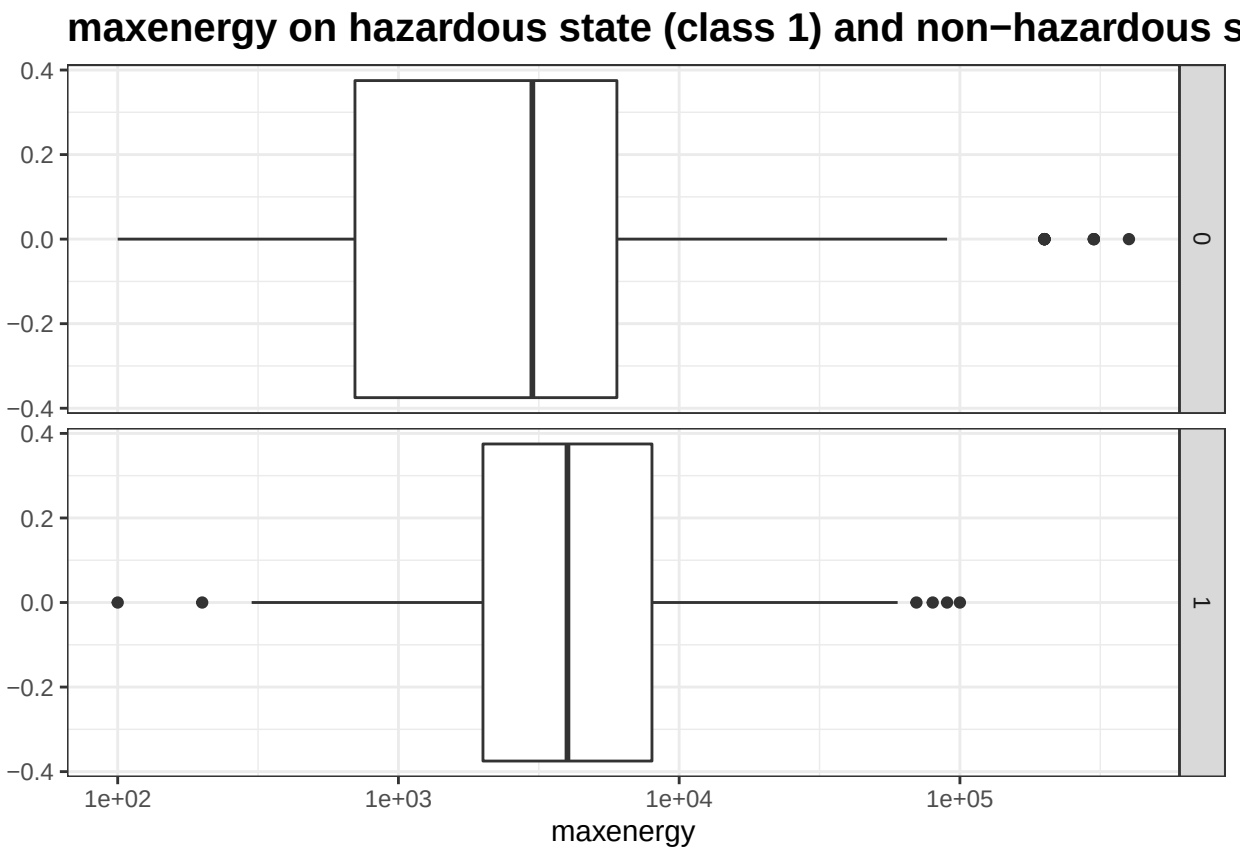
```
## Warning: Removed 1464 rows containing non-finite values (stat_boxplot).
```



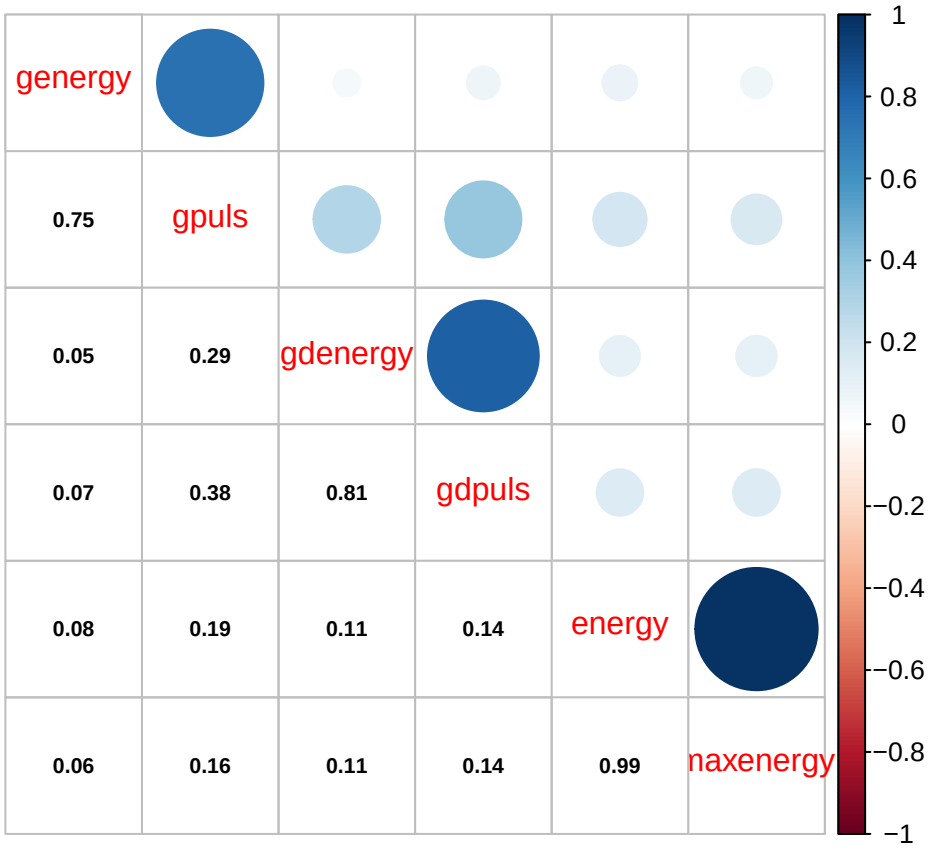
2.2.11 maxenergy on positive class and negative class

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 1464 rows containing non-finite values (stat_boxplot).
```



2.2.12 explore correlation between numeric variables



2.3 Modeling Approach

3. Results and Discussion

4. Conclusion and Future Work

4.1 Limitation

More and more advanced seismic and seismoacoustic monitoring systems allow a better understanding rock mass processes and definition of seismic hazard prediction methods. Accuracy of so far created methods is however far from perfect. Complexity of seismic processes and big disproportion between the number of low-energy seismic events and the number of high-energy phenomena (e.g. $>$ 10ˆ4J) causes the statistical techniques to be insufficient to predict seismic hazard Therefore, it is essential to search for new opportunities of better hazard prediction, also using machine learning methods. Unbalanced distribution of positive ('hazardous state') and negative ('non-hazardous state') examples is a serious problem in seismic hazard prediction. Currently used methods are still insufficient to achieve good sensitivity and specificity of predictions.

Aknowledgement

Marek Sikoraˆ{1,2} (marek.sikora ‘@’ polsl.pl), Lukasz Wrobelˆ{1} (lukasz.wrobel ‘@’ polsl.pl) (1) Institute of Computer Science, Silesian University of Technology, 44-100 Gliwice, Poland (2) Institute of Innovative Technologies EMAG, 40-189 Katowice, Poland

Reference

- 1. 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
- 2. 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
- 3.