Hazard Prediction using Machine Learning Models

Background

Mining activity was and is always connected with the occurrence of dangers which are commonly called mining hazards.

A special case of such threat is a seismic hazard which frequently occurs in many underground mines.

Therefore, it is essential to search for new opportunities of hazard prediction using machine learning models.

Content

Exploratory Data Analysis

Getting to know our data

Pre-Processing

Briefly explanation about data processing

Model Result

How well the model performed?

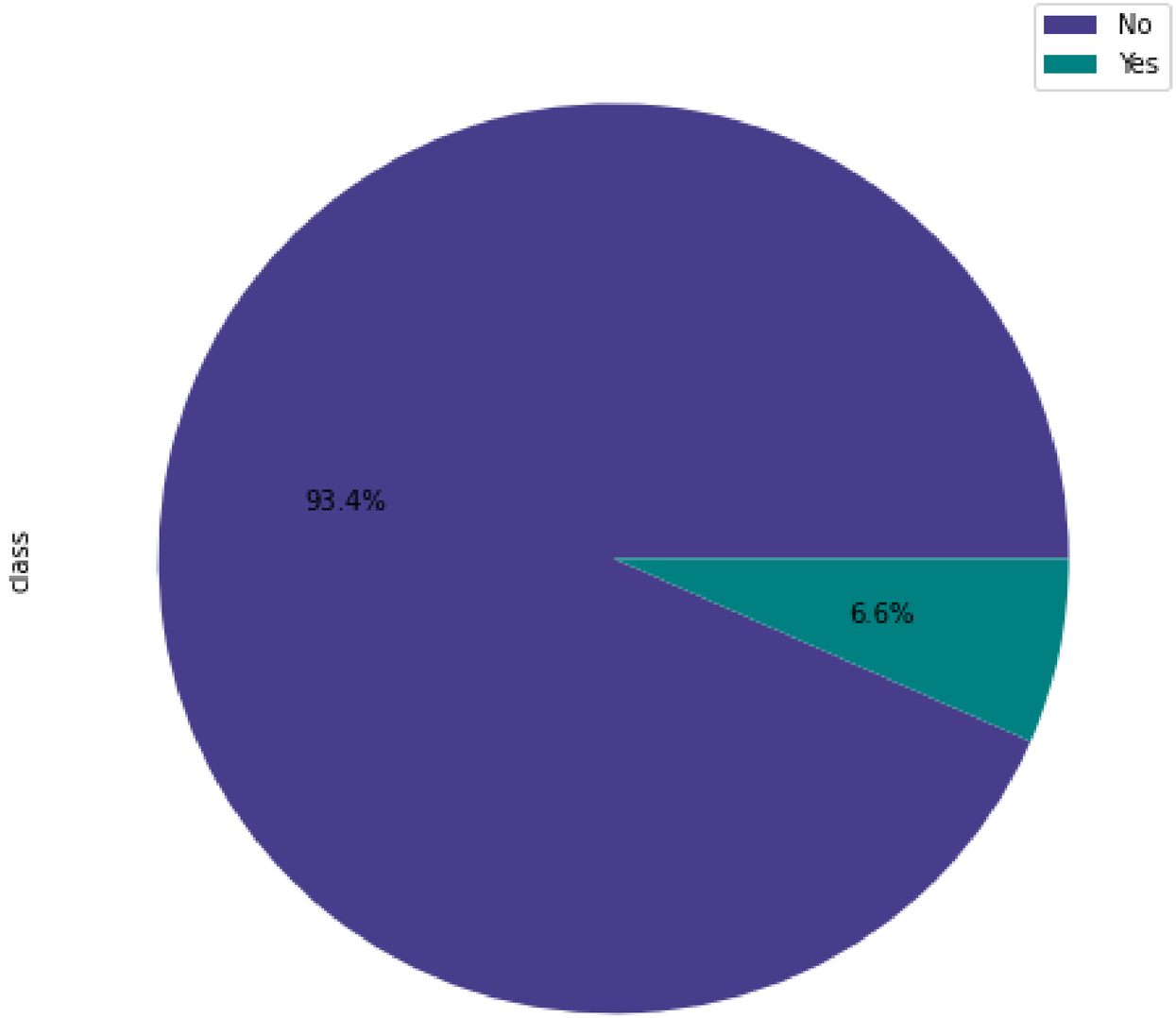
Next Steps

Our plan and suggestion for improvement

About The Data

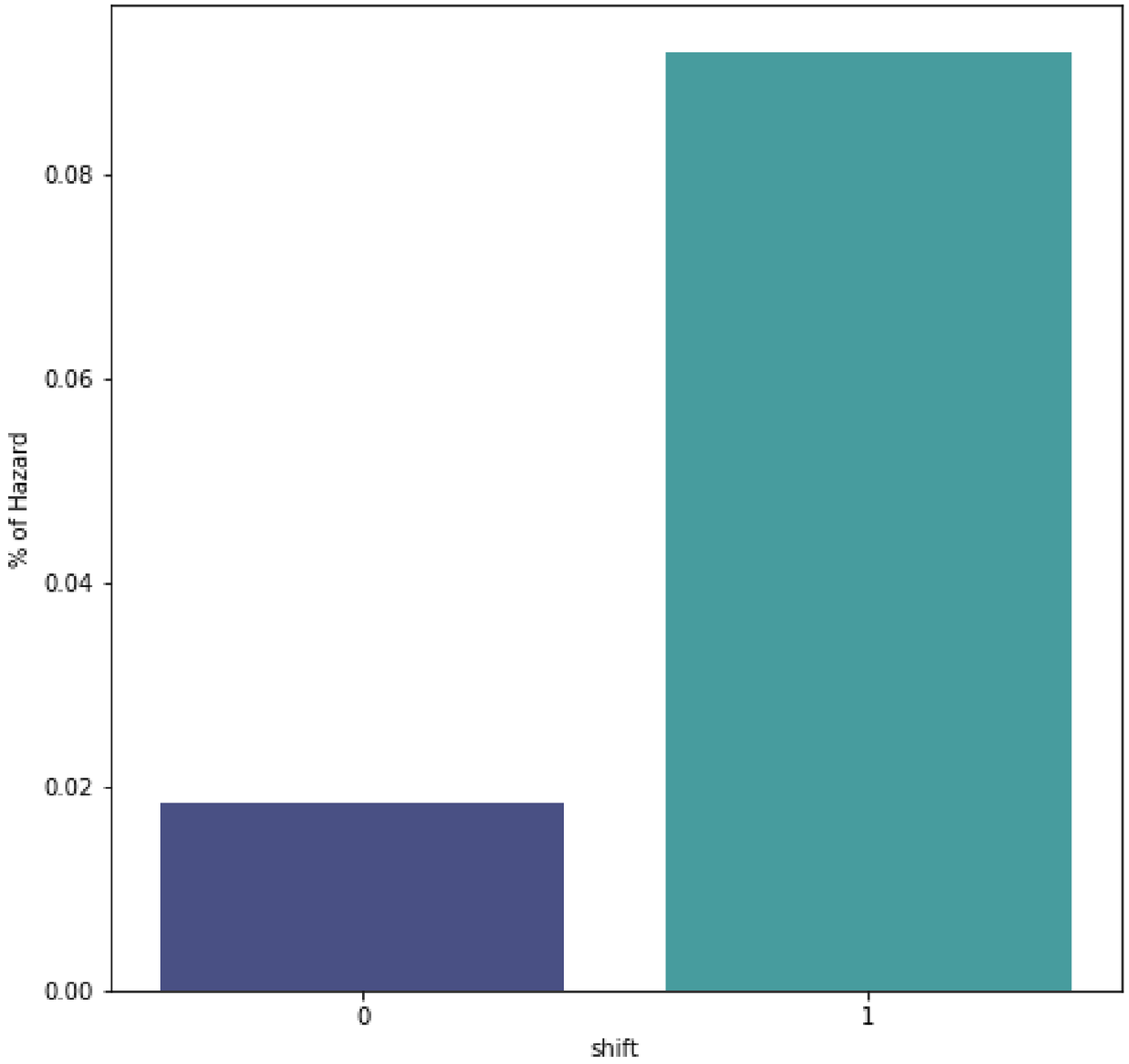
seismic	seismoacoustic	shift	genergy	gpuls	gdenergy	gdpuls	hazard	nbumps	nbumps2	nbumps3
a	a	N	15180	48	-72	-72	a	0	0	0
a	a	N	14720	33	-70	-79	a	1	0	1
a	a	N	8050	30	-81	-78	a	0	0	0
a	a	N	28820	171	-23	40	a	1	0	1
a	a	N	12640	57	-63	-52	a	0	0	0
a	a	W	63760	195	-73	-65	a	0	0	0
a	a	W	207930	614	-6	18	a	2	2	0
a	a	N	48990	194	-27	-3	a	1	0	1
a	a	N	100190	303	54	52	a	0	0	0

nbumps4	nbumps5	nbumps6	nbumps7	nbumps89	energy	maxenergy	class
0	0	0	0	0	0	0	0
0	0	0	0	0	2000	2000	0
0	0	0	0	0	0	0	0
0	0	0	0	0	3000	3000	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	1000	700	0
0	0	0	0	0	4000	4000	0



Hazard State (Target)

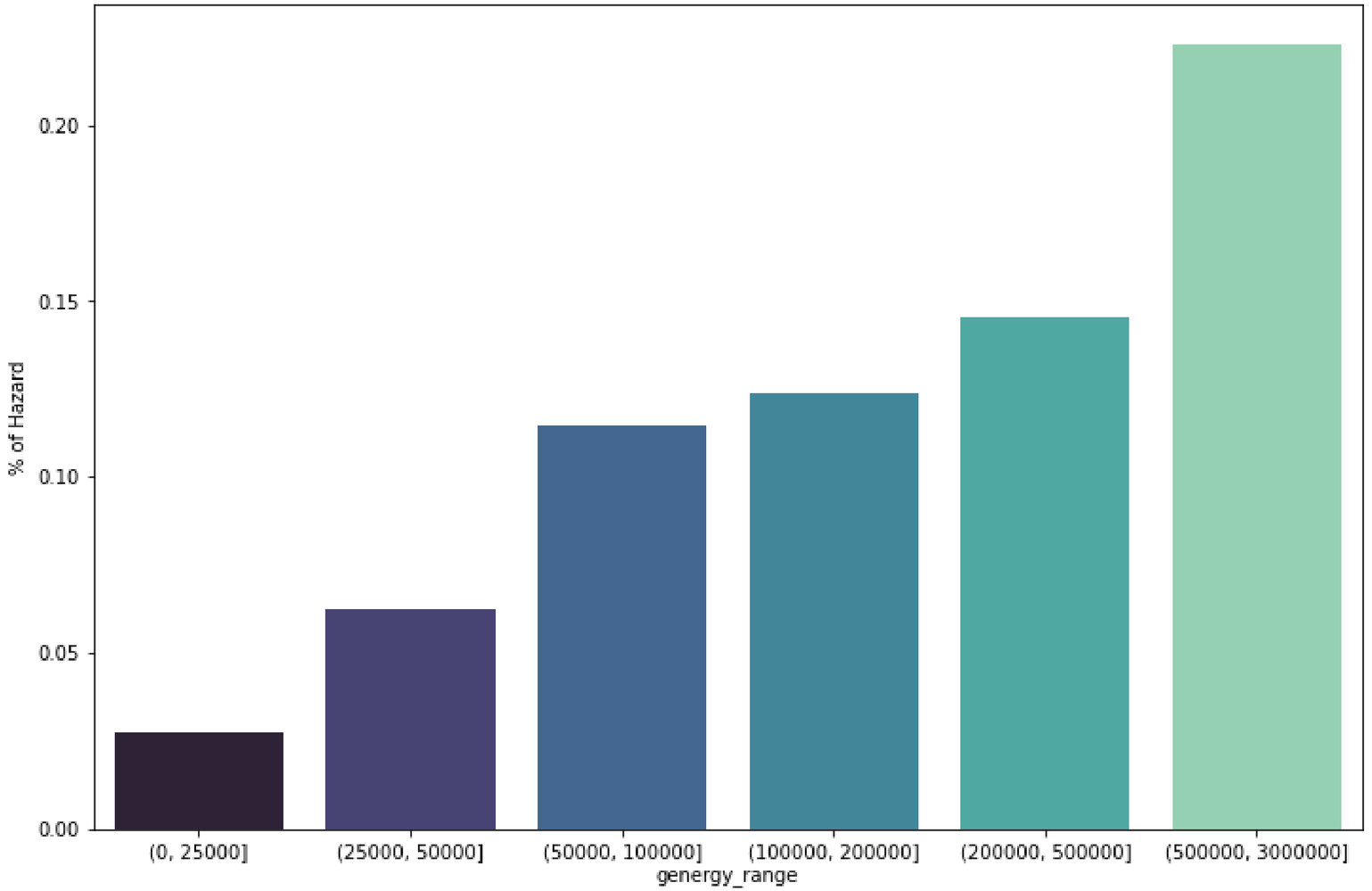
There are 170 seismic hazard happened in the past.
Taking only 6.6% portion of our target data.



Based on Shift

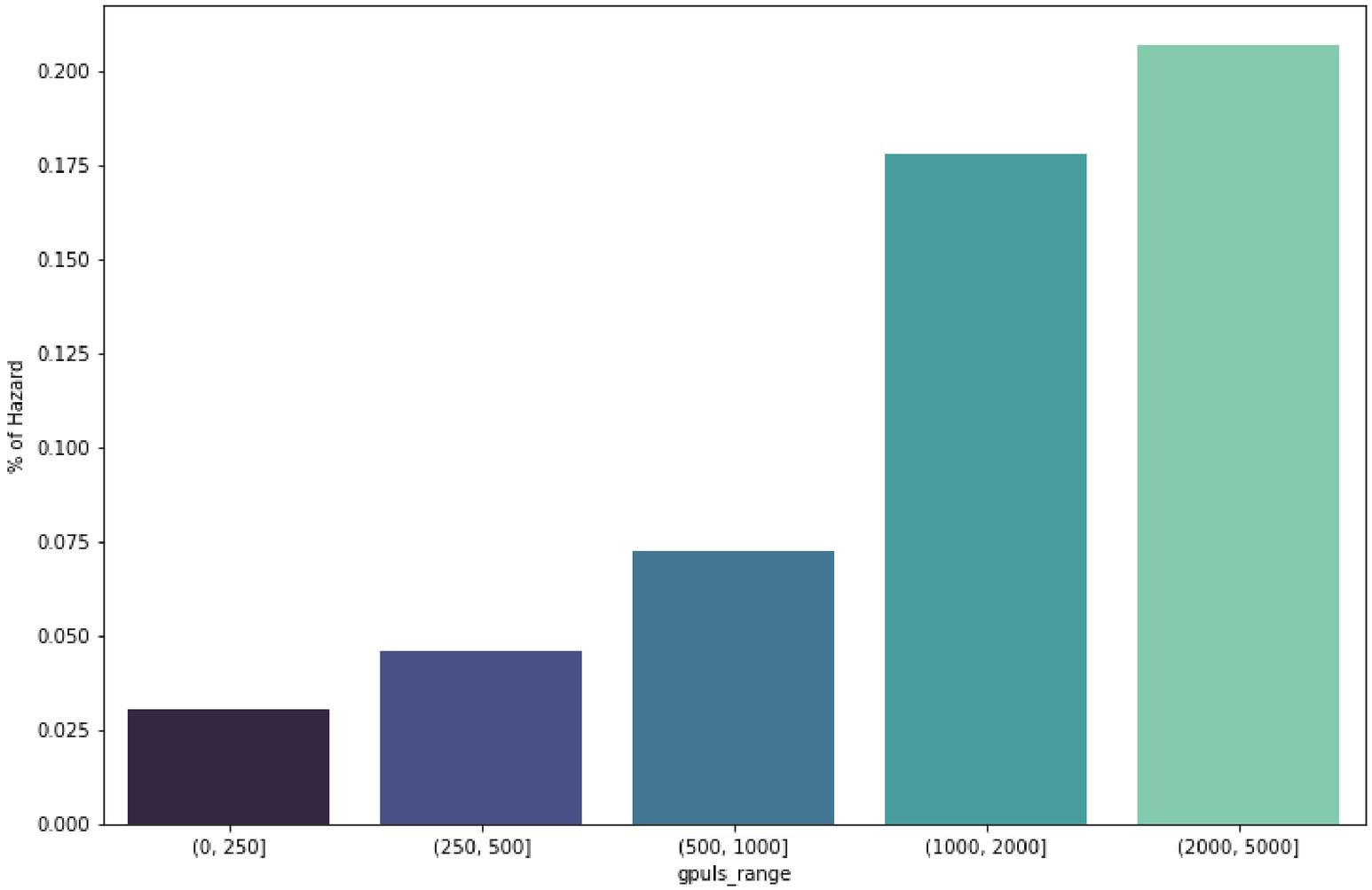
0 : preparation shift
1 : mining shift

Sesimic hazards are more likely to happen in mining shift than preparation shift



Based on Energy

The greater the energy, the chances of seismic hazard occured will be higher too.



Based on Pulse

The more seismic pulse that are recorded, the chances of seismic hazard happened will be higher too

Data Preprocessing

Splitting Train and Test

Split dataset to train and test with 75:25 proportion

Feature Selection

Select feature for the model based on correlation value and personal judgement

Handling Outliers

Cleaning the data from extreme value

Feature Scaling

Equalize the range of numerical column to help model train better

Handling Imbalance

Equalize the proportion of target labels to help model learn the minority label better

The Machine Learning Models

- Logistic Regression
- Suppor Vector Machine
- Random Forest
- Gradient Boosting

Model Evaluation Metric

Model	Train Accuracy	Test Accuaracy	Recall 0 Test	Recall 1 Test
Logistic Regression	0.73	0.79	0.80	0.63
Support Vector Machine	0.75	0.81	0.83	0.56
Naive Bayes	1.00	0.91	0.97	0.07
Gradient Boosting	0.87	0.83	0.86	0.47

Model Improvement

After we choose the models that have better performance,. Then we make improvements by tuning the model's hyperparameters using GridSearchCV function

Model	Accuracy	Recall 0	Recall 1
Logistic Regression	0.79 > 0.78	0.80 > 0.80	0.63 > 0.63
Support Vector Machine	0.81 > 0.82	0.83 > 0.83	0.56 > 0.63

Overall Summary

- The performance of Support Vector Machine model is still far from expectations, especially in accurately predicting hazardous state (Recall value : 0.63)
- Mistakes in predicting hazardous state to be non-hazardous state will lead to fatal consequences, there could be casualties from site workers. Company will also suffer from this incident, at worst company's business license could be revoked by regulator.
- For non-hazardous state predictions, the model's performance is quite good (Accuracy value : 0.82).

Next Steps

Try another machine learning models in addition to four models that already created.

Perform hyperparameter tuning for other models.

Looking for additional related data to support this research.

For more details on this project, please visit:
https://github.com/H8-Assignments-Bay/p1---ftds-001-hck--ml2-dimitriasta