# Summary report

Biliang Wang
bwang135@uottawa.ca

September 24, 2019

I utilized Scikit-Learn to analyze the given data, and the task is a binary classification.

The problem is about seismic bumps prediction, which requires, ideally, every bump should not be missed. On the other hand, we can afford some non-bump samples to be classified as bumps. The most suitable model would be the one with highest recall, in other words, the number of false negative should be as small as possible.

The majority of time was spent on data preparation, which mainly consisted of converting the format of columns of data and building the pipelines. Training models and making predictions were relatively straightforward, provided data sets were well-prepared. All the models were dumped and preserved using joblib.dump.

The key performance of models are as follows:

Table 1: Recall, precision and AUC score of four models

| *Models* | *Decision tree* | *Rule-based* | *Naive Bayes* | *K-nearest neighbors* |
|---|---|---|---|---|
| *recall* | 0.129412 | 0.067797 | 0.900000 | 0.052941 |
| *precision* | 0.077739 | 0.070588 | 0.088388 | 0.243243 |
| *AUC score* | 0.528915 | 0.500249 | 0.683230 | 0.618608 |

According to the performance of the four models, clearly Naive Bayes is the best one among four models, with the highest recall of 0.900000. Thus I would use Naive Bayes algorithm to address this problem.

Finally, due to the small size and imbalance of the dataset(only 170 samples represent positive class), the model might have underfitting concern. We need significantly more data to train a better model.