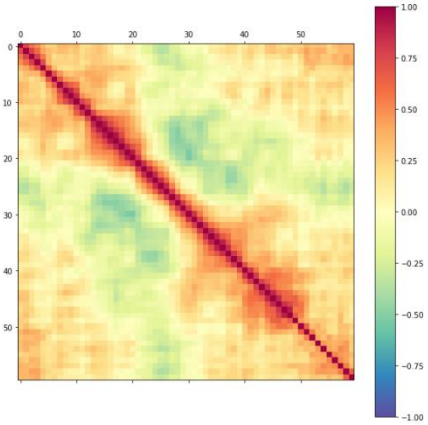
Obstacle Detection in Seismic Data using ML Classification

Background

Seismic data is the key to understanding the subsurface. Whether this be for purposes close to the surface such as uses in building infrastructure or deeper applications such as searching for hydrocarbon deposits, seismic is used and analyzed frequently. The process includes creating sound waves, usually through a hydraulic hammer banging against the ground, which propagates through the subsurface and is reflected by what is within the earth. By calculating the amplitude (strength) and frequency of the reflections, a clear image of the subsurface can be attained. In this study, a seismic dataset is analyzed for the purpose of creating a model that can predict the presence of a metal pipe within the section. Creating such models could help engineers continue to add infrastructure to our cities while refraining from causing costly damage to the existing pipelines.

Data

The data analyzed in the study contains 60 features and 1 label, the features representing the 60 angles in which the sonic waves are recorded once they reach the sensor. The attributes of each feature consist of a number between 0 and 1 which indicates the strength of the signal. The label, which indicates the presence of a rock or metal, is changed to a binary representation within the study to 1 representing metal and 0 representing rock. The data is difficult to interpret just by looking at it, but it is relatively clean so little data processing was necessary before fitting the models. After scaling the data using a standard scalar, the factors that were most correlated or influential were determined using the method SelectKBest from sklearn. From this analysis, angle sensors from 10, 11, 12, 45, and 49 degrees were the most influential in predicting the output of the classification.



Models

Eight models were trained in the classification of the metal pipe versus the rock. The data was split into two sets: a train set (80% of the total data) and a test set (20%). This data was fit to the following models: AdaBoost, Logistic Regression, Gaussian Naïve Bayes, K-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine, and Multi-layer Perceptron. For simplicity of the study, default parameters were used in initializing and fitting the models as there was no direct evidence that using other parameters would prove more beneficial.

Results

Results show that the Multi-layer Perceptron (MLP) was the most accurate in predicting the metal object versus the rock correctly. However, the Support Vector Machine classifier (SVC) and the Random Forest Classifier (RF) often were very close to the same accuracy. To analyze the accuracy of the models, confusion matrices were printed to understand where the models were going wrong. The MLP model predicted 14 out of 14 of the rocks correctly within the test dataset and 25 out of 28 of the metal pipes correctly. The SVC comes in a close second place predicting 13 out of 14 of the rocks correctly and 25 out of 28 of the metal pipes correctly, and lastly the RF model predicted 12 out of 14 rocks correctly and 25 out of 28 metal pipes correctly. As can be seen, the accuracy from all three models is quite good and could be used as a reliable model. These models, however, could be improved to fix the 3 out of 28 metal pipes that were not detected. Because the error of predicting rock with the actual result being pipe is much worse than the opposite, operators need to be careful when using these models. Future studies could also include using grid search or hyper parameter optimization to better tune the models to a validation set and then test them on the test dataset.