**Fall 2018 - DATA MINING FOR BUSINESS (BYGB-7967-003)**

**Seismic Bumps Dataset: Evaluation and Conclusions**

**Project Report**

**Team 4:**

**Lakshmi Sravya Katta**

**Ila Srivastava**

**Josh Simpson**

**Classification Statement:**

Main aim of all seismic hazard assessment methods is to predict (with given precision relating to time and date) increased seismic activity which can cause a rockburst. 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').

**Dataset:**

Source of our dataset is <https://archive.ics.uci.edu/ml/datasets/seismic-bumps> from UCI datamining Archive. The number of records in dataset are 2584.The data describe the problem of high energy (higher than 10^4J) seismic bumps forecasting in a coal mine. Data come from two of longwalls located in a Polish coal mine.

**Techniques Applied:**

We applied restructuring of columns, balancing the node, dealt with outliers and extremes, built neural network, decision tree, logistic regression and ensemble classifier. Also, we have calculated precision and recall for our final classifier and decision tree model.

**Analysis:**

* We placed the data audit node to understand our data and the quality of data. Luckily we had no missing values so no imputation was required but we do had outliers and extremes which we discarded before creating our model.
* Also, some of the fields like seismic, seismoacoustic, shift and ghazard had few categorical variables, which need to be transformed into dummy variables. So we applied restructuring node to handle them.
* One of our classes had too fewer records in comparison with other i.e., the ratio was 2000:170 approx for Class 0 vs Class 1. Therefore, we have used balancing node to increment the records of class having less records.

After all the adjustments we have built our three models:

*Decision Tree:* As per the model, gpuls is the most important predictor. The accuracy with training data is 96.8% whereas with test data it is 97.29% accurate. The model is fairly good.

*ANN:* The accuracy of the final model is 78.7%. It is the percentage of records for which the predicted value matches the observed value for our training data. The most important predictor is nbumps as per the model. The neural network does best at predicting those in the *hazardous* category (81.6%).

*Logistic regression:* The summary tab shows the inputs as well as target that were used by model. In this case these inputs and target was chosen for Forward stepwise. It was observed that there were no missing values and all inputs were 100% complete. Therefore, no imputation was performed. The overall accuracy is 70.37% and our model did not perform that well.

Among all the models **Decision tree** performs the best of all. But we will use ensemble to see if the overall accuracy can still be increased or not.

**Comparison of all the models built:**

|  |  |
| --- | --- |
| **Model** | **Overall Accuracy (%)** |
| *Decision Tree* | 96.8% |
| *ANN* | 78.7% |
| *Logistic Regression* | 70.37% |
| *Ensemble* | 82.21% |

* By comparing the decision tree, neural network and logistic regression models’ overall accuracy along with ensemble classifier, we can conclude that decision tree model is better at predicting the occurrence of high energy seismic bumps.

**Conclusion:**

* Dataset was clean but still many improvisations were made to get good percentage of accuracy.
* Decision tree is the best amongst all the classifiers built.
* Precision and recall for class 0 (‘Non-Hazardous state) was better for decision tree in comparison to ensemble classifier.
* Ensemble didn’t increase the overall percentage and therefore we have decided to stick to Decision tree as final model.
* The coal mine safety experts can use the model to predict whether the seismic bump in the next shift will be hazardous or not.
* In order to prevent further incidents they can take safety and preventive measures beforehand.