Hazardous Asteroid Classification

ECE 532 Final Project

Blake Lopez

Table of Contents

Introduction

A binary classification dataset generated from NASA's Near Earth Object Web Service (NeoWs) was chosen [1]. This dataset contains 39 features for 4687 asteroids that are classified as either hazardous or safe. Approxaitly 16% of the data is classified as hazardous and 84% are classified as safe. Many of the features are repeated information in different units, identification numbers, and dates of calculations or discovery so the number of features will be reduced before applying any algorithm. The main task for this dataset is feature selection for finding other hazardous and safe asteroids.

Methods

Elimination of features with repeat information in different units, dates, and identification numbers reduced the number of features from 39 to 20. One more feature was removed by determining the estimated minimum and maximum diameters of the asteroids were correlated; the max estimated diameter was chosen. This makes the number features to be selected from and perform classifications with to be 19. From this reduced dataset, a random set of 10% of the data was separated to be used for final testing and comparisons of models generated from three different algorithms.

All features, which are positive values, were scaled to be between 0 and 1 to better understand their importance. Visualization of the 19 scaled features showed two sets of features can be used to draw a decision boundary, but the decision boundary is not linear. These features are shown in Figure 1. The sharp right angle in the decision boundaries makes it difficult for classification.

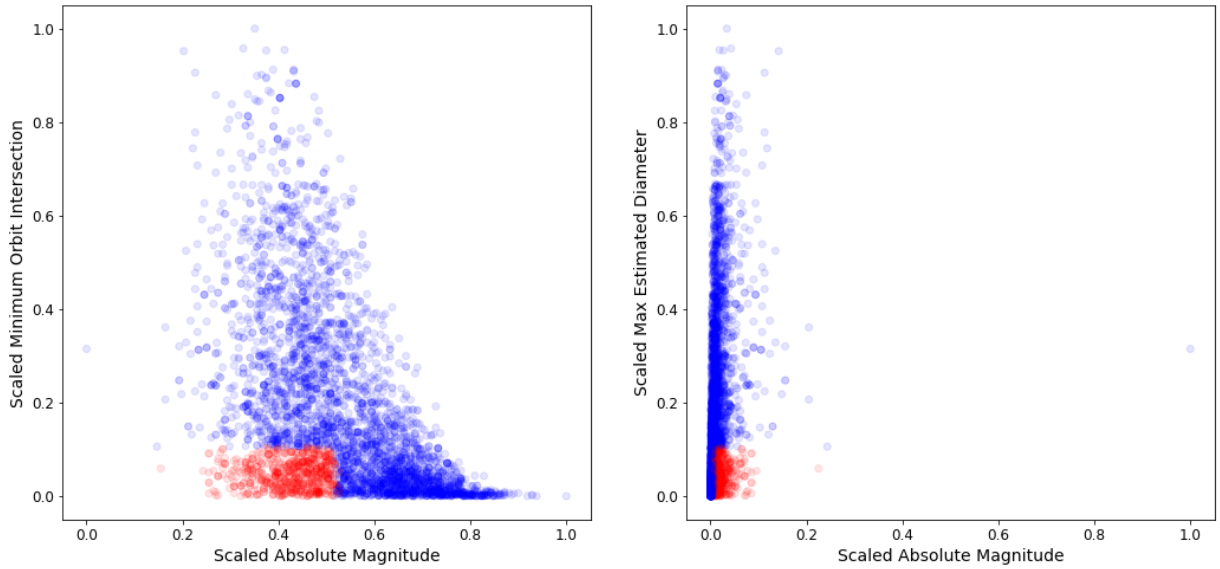

Figure 1 - Scaled features that lead to clear decision boundaries; red is hazardous and blue is safe

The three different algorithms that were chosen are least squared error, least absolute shrinkage and selection operator (LASSO), and a three layer neural network. In addition these algorithms, synthetic minority oversampling technique (SMOTE) will be used to understand and limit the effects of the imbalanced number of classes [2]. SMOTE uses five nearest neighbors to stochasticly generate new data until there is an equal number of datapoints for each class [2].

For least squared error, a method of ranking feature importance by repeatedly deleting the feature that, when removed, has the highest accuracy when compared to the other removed features. This should identify other less correlated features to be able to classify with. A validation step of adding back one feature will be used to check for any features that pair well.

LASSO uses a regularization parameter so the optimal value will have to be chosen for each model generated. A method of ranking will also be used by repeatedly deleting the feature with the lowest magnitude weight. The same validation step will be used to check for paired features.

The three layer neural network uses backpropagation to determine the weights for the input layer and hidden layer. The number of hidden nodes, learning rate, and repetitions will have to be studied for this approach. There will be no feature ranking for this algorithm and attempting to fit the decision boundary in Figure 1 will be the main focus.

Results

Least Squared Error

The three features identified with a clear boundary line was determined to have an error of 16.2% with 10-fold cross validation. However, the error for the hazardous class was 98.4% and the other was just 0.3%. This imbalance in the errors can be attributed to the difference in the amount of data for each class. When SMOTE was used the overall error was 13.4% and the error for the hazardous class was 23.1% and 11.6% for the other. This demonstrates the need for balanced datasets.

Focusing on the results for the oversampled data, the ranking method generated Figure 2. While the lowest estimated error occurs using seven of the ranked features, there is marginal difference using just five of the features. Using these five features, validation was performed by adding each of the removed features to the model as shown in Figure 3. This demonstrates there is marginal improvement by adding any of the removed features.

The five features are ‘Absolute Magnitude’, ‘Minimum Orbit Intersection’, ‘Semi Major Axis’, ‘Perihelion Distance’, and ‘Aphelion Dist’. Testing on the data before oversampling, the error was 13.3% with just 0.3% for the hazardous class and 15.8% for the other. For this application

these are promising results because if an asteroid is classified as safe then it is likely to actually be safe.

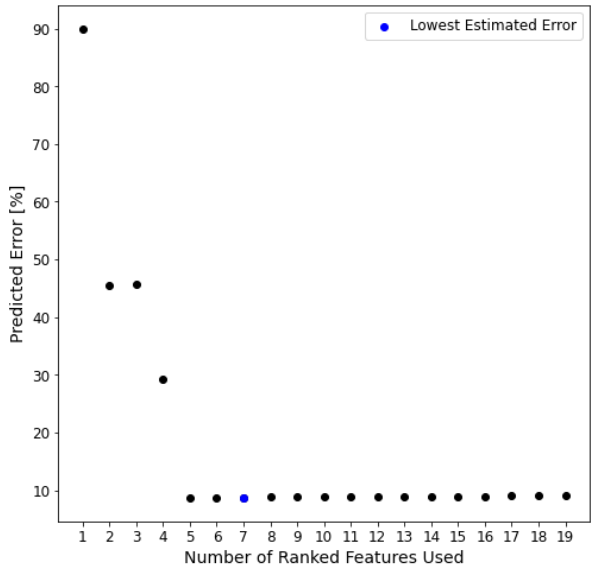

Figure 2 - Ranking results for the oversampled data using least squared error with 10-fold cross validation

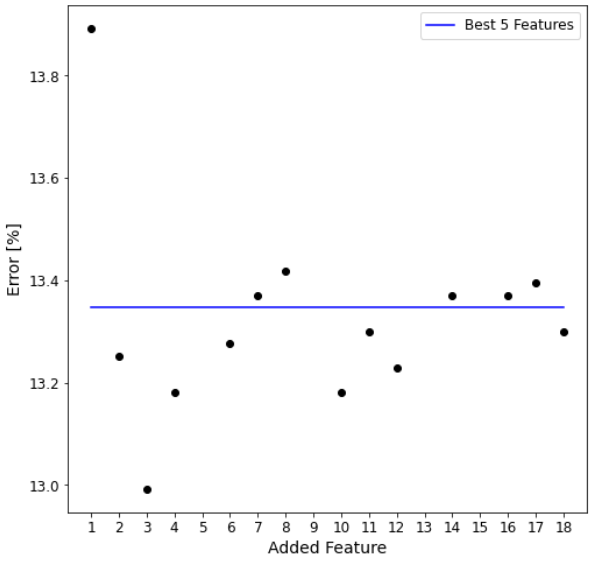

Figure 3 - Validation of the ranking results of the least squared error

LASSO

LASSO generates sparse results for the weight vectors and is often used for feature selection. LASSO generated similar results for the three feature model as least squared error but resulted in a different ranking. The first ranking result generated Figure 4 and then seven best features were chosen and validation was attempted. During validation in Figure 5, it was discovered some of the features did not have an impact on the expected error. As a result, one feature was removed and validation occurred again. In this case, it was found removal of another feature did not impact the error greatly but did increase the error of the hazardous class greatly. As a result, this feature was not removed.

The six features chosen by LASSO are ‘Absolute Magnitude’, ‘Minimum Orbit Intersection’,’Jupiter Tisserand Invariant’, ‘Orbital Period’, ‘Perihelion Distance’, and ‘Aphelion Dist’. Testing on the data before oversampling, the error was 13.2% with just 0.6% for the hazardous class and 15.7% for the other. These results are very similar to the least squared error results but does use a few different features. This questions the results of the rankin method used in both of these algorithms.

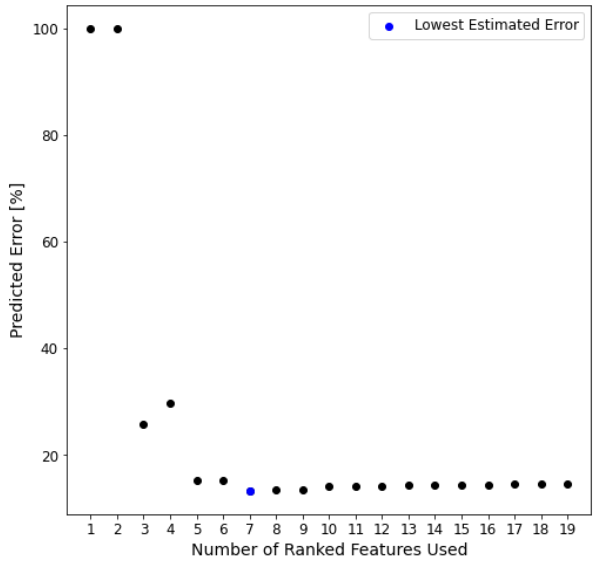

Figure 4 - Ranking results for the oversampled data using LASSO

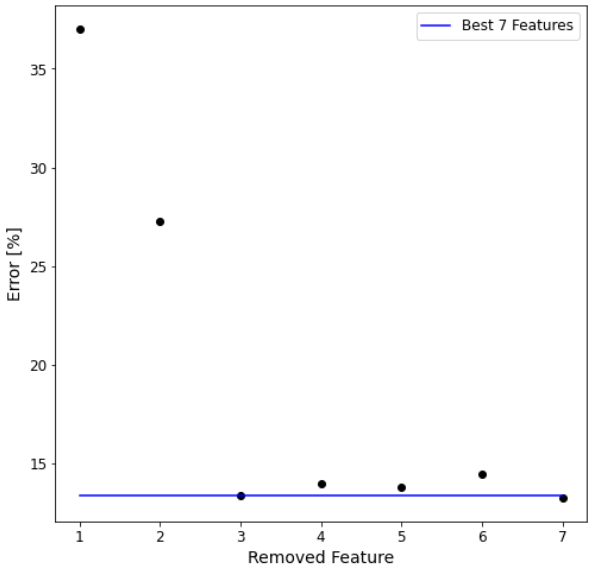

Figure 5 - Initial validation of the ranking results of LASSO

Neural Network

A neural network can create a complicated boundary decision from a limited number of features. This makes a neural network a better method for creating the right angle decision boundary seen in Figure 1. The three features identified were first trained along with a two feature model at the same parameters. This demonstrated the error was comparable between the two and one of the features could be removed. Therefore, two features used to develop the neural network from are 'Absolute Magnitude' and 'Minimum Orbit Intersection'.

To understand the impact of the imbalance in the data on the neural network, the original data and the oversampled data were used to train a neural network using Pytorch under the same operating conditions [3]. This determined the overall error of the two neural networks were the same, but the error for the hazardous class was 1.8% without oversampling and just 0.4% with. Although this difference is not as large as it was for least squared, it is still an important aspect of the dataset that needs to be addressed.

The neural network was then trained and tuned for the oversampled data using the two feature model. First the impact of the number of hidden nodes while scaling the number of repetitions and keeping the learning rate constant was analyzed. This resulted in Figure 6 and found around 10 hidden nodes were appropriate. The impact of the learning rate for 10 hidden nodes and a constant number of repetitions was determined. This resulted in Figure 7 and demonstrates a learning rate of 0.2 is appropriate. Final training was then performed using 10 hidden nodes, a learning rate of 0.2, and letting the backpropagation repetition be large. Testing on the data

before oversampling resulted in an overall error of 0.9% with an error of 0.9% for the hazardous class and 0.8% for the other.

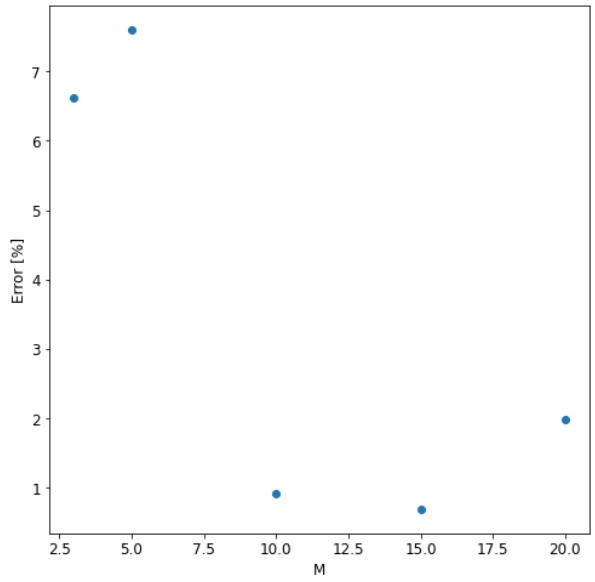

Figure 4 - Impact of the number of hidden layers

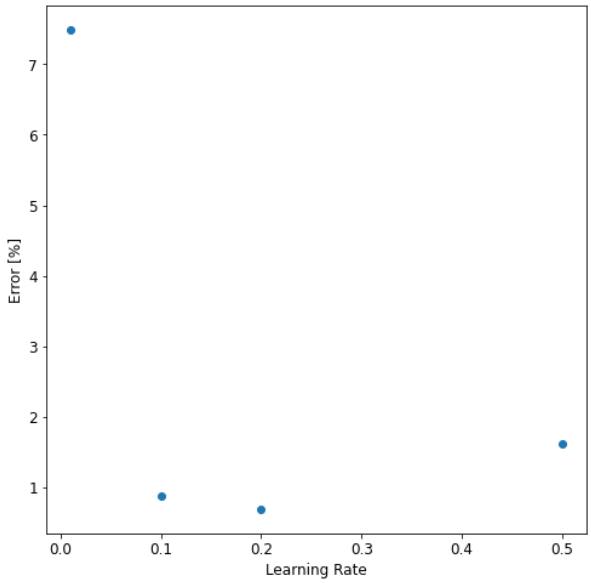

Figure 5 - Impact of the learning rate

Final Testing

Final testing of the different models generated from three different algorithms was performed on the 10% of data separated from original data. The least squared error ranking results did not generalize well and classified each data point as safe. The least squared error on the three feature model resulted in an overall error of 13% with an error of 27% for the hazardous class and 11% for the other. LASSO ranking resulted in an overall error of 19.6% with 0% error for the hazardous class and 23.3% for the other. This result is not as bad as it seems because it means if an asteroid is classified as safe, then one can confidently assume it is.

The final neural network model resulted in an overall error of 0.2% with 0% error for the hazardous class and 0.3% for the other. This result is as close to ideal as one can expect with 0% error for the main class of interest and very low error for the other. As a result, the final recommendation is to use the two features, ‘Absolute Magnitude’ and ‘Minimum Orbit Intersection’, and the neural network model for future classification.

Further details on the methods and results can be found at https://lopezbl.github.io/ECE532_Project/.

References

[1] Mehta, S. (2018, March 01). NASA: Asteroids Classification. Retrieved from https://www.kaggle.com/shrutimehta/nasa-asteroids-classification

[2] Lemaitre, G., Nogueira, F., Oliveira, D., & Aridas, C. (2016). Imblearn.over_sampling.SMOTE¶. Retrieved from https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html

[3] Defining a Neural Network in PyTorch. (2017). Retrieved from https://pytorch.org/tutorials/recipes/recipes/defining_a_neural_network.html