Blake Lopez
ECE 532 - Project Update 1
11/17/2020

Hazardous Asteroid Classification

The nasa.csv has been loaded and preprocessed. The Load Data.ipynb removes unnecessary columns such as ID numbers, dates of discovery/calculations, and columns that have repeat information in different units/scales. This brought the number of features from 39 to 20. Then, 1 more feature was removed based upon the perfect correlation of the min and max estimated diameter of the asteroids. A nasa_filt.csv was created and has all of the data with the 19 features. From this, the nasa_test.csv was created from 10% of the data for final testing and comparison of the three algorithms. The remaining 90% of the data was placed in nasa_train.csv for training of the three algorithms.

Using nasa_train.csv and the Least Squared.ipynb, the least squared algorithm was used to see which features were the most important. Using all of the features, the expected classification error was 9.5%. A method of ranking the features based on their impact on the expected error was used. This technique concluded there are 7 features that allow for an expected classification error of 9.4% which means the other features had little to no impact on the classification. The 7 features are "Absolute Magnitude", "Est Dia in KM(max)", "Orbit Uncertainty", "Minimum Orbit Intersection", "Semi Major Axis", "Perihelion Distance", and "Aphelion Dist". However, the confusion matrix shows an accuracy of about 50% for the hazardous cases. This can be explained by the difference in size between the two classes; the hazardous class is just 16% of the data.

Based on low prediction of the hazardous class due to its low data numbers, a method called SMOTE (Synthetic Minority Oversampling TEchnique) was used to increase the sample size of the hazardous class. This is accomplished by randomly choosing one of the 5 nearest neighbors of a data point and then randomly selecting a point in between for a new data point; this occurs until the number of data points in each class are equal.

Using nasa_train.csv and the Least Squared SMOTE.ipynb, the least squared algorithm along with SMOTE was used to see which features were the most important. Using all of the features for training on the expanded data set and then evaluating on the original data, the classification error was 13.7%. This is higher than just using least squared, but the accuracy of the hazardous case improved from 59.2% to 98.4%. For this application, it is better to be more accurate in correctly determining if an asteroid is hazardous than it is to be more accurate in determining if an asteroid is non hazardous.

The same method of ranking the features by their impact on the expected error was then used. This technique concluded there were 5 features that allowed for a classification error of 13.3%

which means the other features had little to no impact on the classification. The 5 features are "Absolute Magnitude", "Minimum Orbit Intersection", "Semi Major Axis", "Perihelion Distance", and "Aphelion Dist". The confusion matrix shows an accuracy of 99.7% for the hazardous cases and 84.1% for the nonhazardous class. The use of SMOTE increased the overall error but balanced the accuracy of the models. For the next algorithm, LASSO, SMOTE will be expected to have a similar effect.

Significant progress has been made on the project, but I am slightly behind schedule. However, I plan to be able to catch up soon and I have made a revised timeline that is shifted by a few days from the original timeline. The Github page for following the progress of this project is https://lopezbl.github.io/ECE532_Project/.

Revised Project Timeline		
Date	**Tasks**	**Submission**
~~11/01/2020~~	~~Complete pre-processing of the data~~	~~None~~
~~11/08/2020~~	~~Complete the least squares algorithm for classification~~	~~None~~
~~11/16/2020~~	~~Complete the least squares algorithm for classification with SMOTE~~	~~None~~
~~11/17/2020~~	~~Update the Github page with progress~~	~~Update 1~~
11/20/2020	Complete the LASSO algorithm for classification with SMOTE	None
11/27/2020	Complete the neural network algorithm for classification	None
12/01/2020	Update Github page with progress and initial comparison of the algorithms	Update 2
12/03/2020	Complete final comparison and recommended classification	None
12/08/2020	Complete first draft of the final report	None
12/12/2020	Submit final draft of the final report	Final Report
12/17/2020	Evaluate two other projects	Peer Evaluations