Potentially Hazardous Asteroids

By: Rachel Goodridge

Abstract

Near-Earth Objects (NEOs), such as asteroids, occasionally have the potential to collide with Earth and cause enormous damage. NASA has been documenting these NEOs as well as many other types of space-going debris and developing diversion tactics given the event of an impact threat. From NASA's Small-Body Database, information about asteroids was used to predict which are potentially hazardous. Following oversampling to address class imbalance and optimization of many parameters, the best model based on F1 validation scores was XGBoost Classifier. The score on the holdout data was 0.784 and the most important features were absolute magnitude, closest distance to the sun, and orbit uncertainty.

Design

Near-Earth Objects (NEOs) such as asteroids, most of which orbit the sun, may be affected by the gravitational pull of other planets enough to change the path of their orbit. Any NEOs larger than 460 feet would cause enormous damage if they collided with Earth. NASA is under congressional orders to document these, as most space debris is undocumented. Ideally, by collecting various information about each asteroid and its orbit, we can determine which should be classified as a Potentially Hazardous Asteroid (PHA). Then, warning systems and diversion tactics can be developed and deployed in response to potential threats.

Data

NASA has created a Small-Body Database (SBDB) that is freely available for the public to download, and from an SBDB query, I have retrieved all asteroids that are considered NEOs. There are 14 features and over 27,000 rows in this particular dataset. Each row is an individual asteroid and features in the data include the mean/min/max distance to the sun, inclination of the orbital plane, uncertainty of the orbit, absolute magnitude (a measure of luminosity), PHA flag, and more.

Algorithms

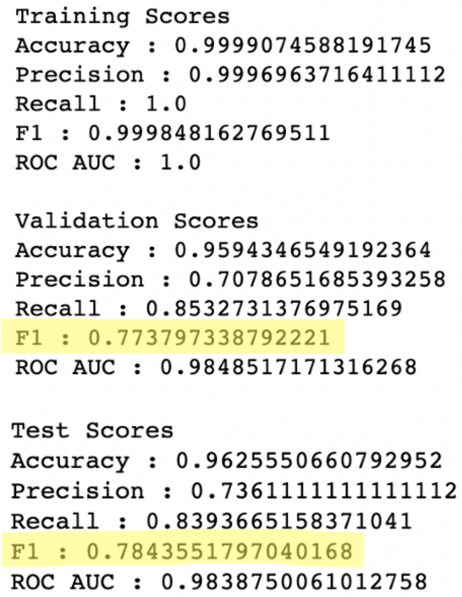
Due to the class imbalance in this dataset (many more nonhazardous asteroids), I oversampled the rarer class using SMOTE. The models I used to fit the data include K-Nearest Neighbors, Logistic Regression, Decision Tree, Bagging Classifier, Random Forest, Extra Trees, AdaBoost Classifier, Gradient Boosting, XGBoost, Voting Classifier, Stacking Classifier, and Gaussian Naive Bayes. Then I optimized several different parameters of each model to find the best fit. I also made both hard and soft predictions, and found the best soft cutoff threshold.

Tools

- numpy and pandas for data exploration
- matplotlib and seaborn for graphing
- sklearn, imblearn, and xgboost for modeling

Communication

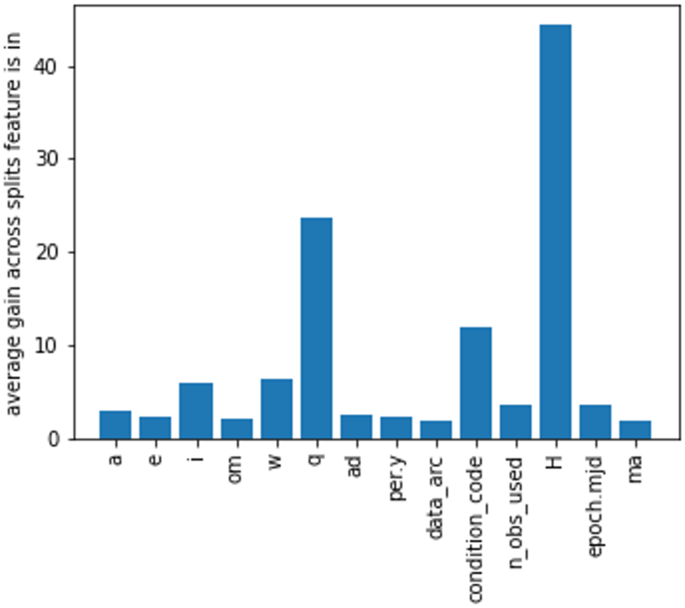
The model with the best performance, based on F1 validation scores, was XGBoost Classifier. The recall score was a bit higher than the precision score, which could be beneficial in this case. It might be better to classify a few harmless asteroids as hazardous than to miss some of the actually hazardous ones. Feature importance was also determined by finding the average score gained by each feature across the splits in which the feature appeared. From this, we can see that absolute magnitude (H) is the most important, followed by closest distance to the sun (q) and orbit uncertainty (condition_code). The final F1 score on the holdout data was 0.784.



```
Training Scores
Accuracy : 0.9999074588191745
Precision : 0.9996963716411112
Recall : 1.0
F1 : 0.999848162769511
ROC AUC : 1.0

Validation Scores
Accuracy : 0.9594346549192364
Precision : 0.7078651685393258
Recall : 0.8532731376975169
F1 : 0.773797338792221
ROC AUC : 0.9848517171316268

Test Scores
Accuracy : 0.9625550660792952
Precision : 0.7361111111111112
Recall : 0.8393665158371041
F1 : 0.7843551797040168
ROC AUC : 0.9838750061012758
```


average gain across splits feature is in
a
e
i
om
w
q
ad
per.y
data_arc
condition_code
n_obs_used
H
epoch.mjd
ma