Predict Asteroids Potentially Hazardous or Not Using NASA Asteroids Dataset

Group member: Yangxin Fan

Part I: Abstract

Asteroids larger than approximately 35 meters across can pose a threat to a town or city if they hit earth. Numerous scientific papers proved that asteroid impact cause the end-Cretaceous dinosaur extinction [1]. Hence, accurately predicting whether a particular asteroid within the Solar system is hazardous or not to Earth plays an important role in safeguarding the development of human civilization. NASA JPL (Jet Propulsion Laboratory) [5] has collected the most complete dataset of known asteroids and with labels to be either hazardous or not. For analysis, I split the dataset into 75% training and 25% testing using stratified sampling. I use PCA (Principal Component Analysis) to reveal the fact that hazardous asteroids are always clustered together. I build a Random Forest classifier to predict whether an asteroid or not. I achieved test accuracy **99.99%** and recall **0.9922**. I implemented a series of data preprocessing or engineering techniques, for example min-max normalization, SMOTE resampling [6], and features selection using tree ensemble method – Extra Trees. My model perfectly handles the data imbalance issue (only 0.22% asteroids are hazardous within dataset).

Part II: Introduction

An asteroid is a minor planet of the inner Solar System. Historically, these terms have been applied to any astronomical object orbiting the Sun that did not resolve into a disc in a telescope and not observed to have characteristics of an active comet such as a tail. Most of asteroids are located in the main asteroid belt and the others are co-orbiting Jupiter. Some asteroids travel close to the Earth which are potentially hazardous objects if they are estimated to be large enough to cause regional devastation. Several Machine Learning models have been utilized to identify Earth-impact asteroids [2, 3, 4]. My goal is to develop a straightforward Machine Learning model to predict hazardous asteroids with a high recall and overall accuracy. I use the most complete asteroids dataset from JPL. It consists of 958,524 asteroids and 45 features. After data preprocessing (ex: one-hot-encoding) and cleaning (ex: drop columns with over 50% missing values), the dataset consists of 932,335 asteroids and 128 features. The target variable is pha (potential hazardous asteroids or not).

Part III: Methodology

a): I build a Random Forests Classifier to predict whether a particular asteroid is hazardous or not.
b): I implement PCA (Principal Component Analysis) for dimensionality reduction and reveal inner structure of the dataset.
c): I use tree ensemble method – Extra Trees to do feature selection to select most relevant features to boost model performance.
d): I improve the model results through applying min-max normalization to numeric features and SMOTE resampling to handle the data imbalance issue.

Part IV: Data Analysis

a): Supervised method – Random Forests

Due to the limited computational power I have in my laptop, I am only capable to do limited Grid-search based hyper-parameters tunin. The best parameter combination I have is max_depth = 20, max_feature = “sqrt”, min_sample_leaf = 2, min_sample_split = 5, and n_estimator (i.e: number of trees) = 150. This Random Forests model is trained by selected top 20 most important features selected by tree-based method. Below is the confusion matrix on the test dataset, the precision is 0.9838, recall is 0.9922, accuracy is 0.9999, and F1 score 0.988. Note: 0 means non-hazardous and 1 means hazardous. Overall, I achieve a pretty good model performance and solve the potentially low recall/precision issue caused by the data imbalance.

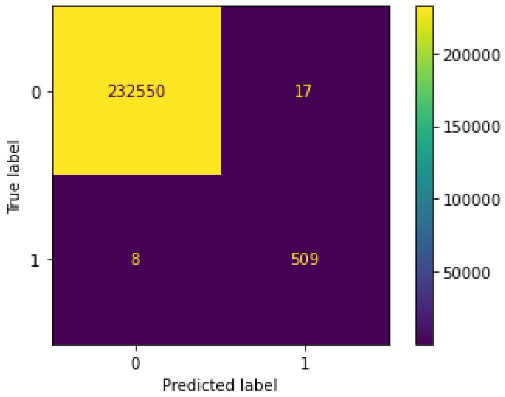

Figure 1. Confusion matrix of Random Forest Model on test dataset

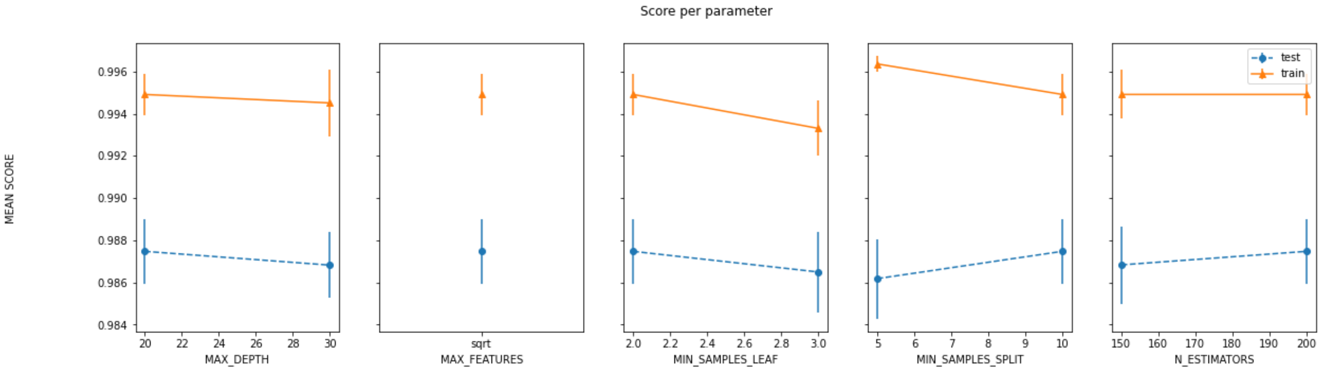

Figure 2. Preliminary Grid-search based hyperparameters tuning result

b): Unsupervised method – Principal Component Analysis

For unsupervised learning, I implement PCA for dimensionality reduction. As you can see from PC1 vs. PC2 plot, only 15.39% variance is explained by first two principal components and there are three obvious clusters: left, middle, and right ones. Interestingly, we see red dots (hazardous asteroids) only appear in the right cluster while blue dots (non-hazardous) show in the left and middle cluster and blend with hazardous clusters in the right cluster. Hence, PCA tells us that hazardous asteroids highly likely appear within the red bounding box. Domain researchers may further analyze the attributes of asteroids within this box.

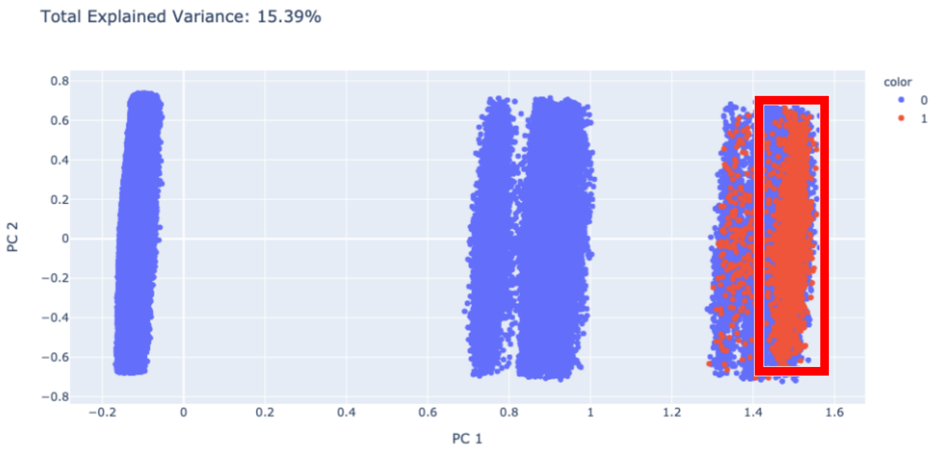

Figure 3. PC1 vs. PC2

In order to achieve over 80% variance explained, we have to include first 35 principal components, this means PCA does not work well in terms of dimensionality reduction since we have 128 features in original dataset. However, PCA does provide us good insight about where to find hazardous asteroids in the PC1. vs. PC2 plot.

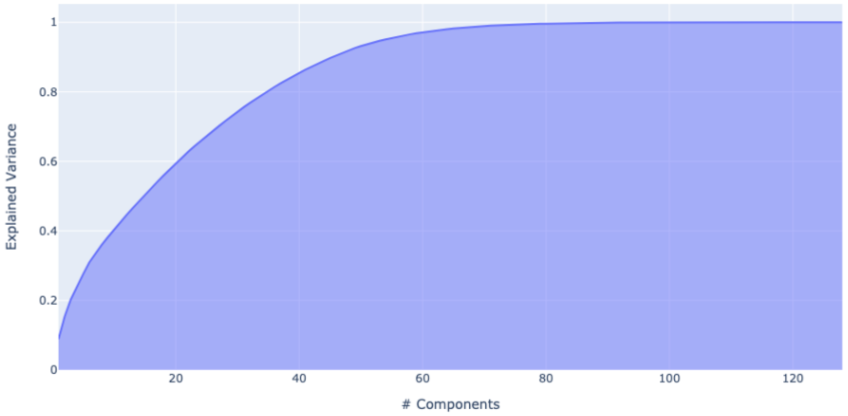

Figure 4. Number of Principal Components vs. Explained variance

c): Feature selection method

Since my supervised method is Random Forests, I use a similar tree ensemble method – Extra Trees for feature selection. The top 20 most important features are shown in below graph. These features are selected for building Random Forests classifier. Noticeably, the model with top 20 features performs even slightly better than the full model with all features. The F1 score is 0.988 (top 20 features) vs. 0.9875 (full model).

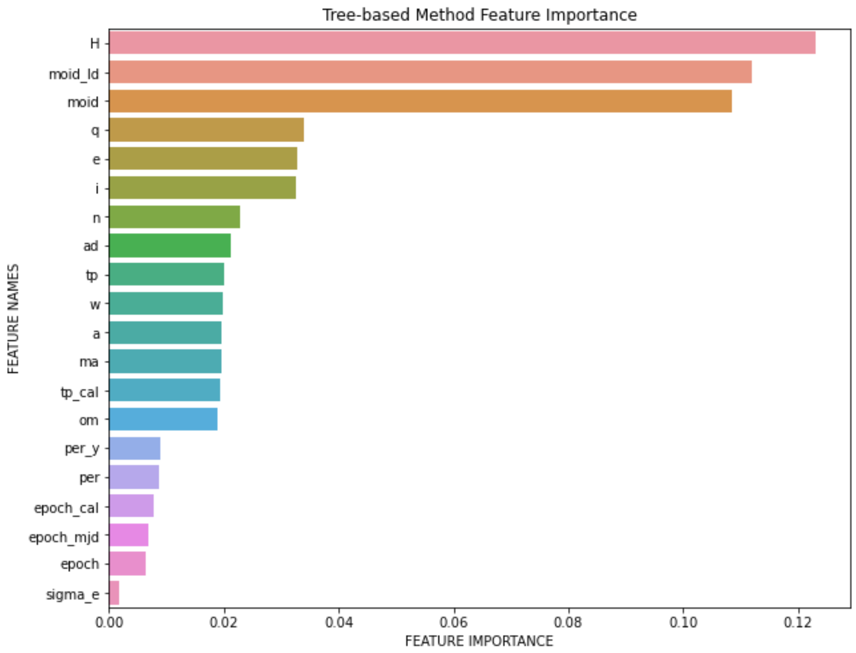

Figure 5. Top 20 features importance ranking

d): methods used to improve model performance
There are two methods used to improve model performance: min-max normalization and SMOTE resampling. First, I apply min-max normalization to restrict all numeric variables to be in range [0, 1]. This solves the problem of incompatible issue among Physics features with different units. Second, I apply SMOTE to deal with the huge data imbalance issue since only 0.22% of asteroids are actually hazardous. SMOTE oversamples the positive (hazardous) asteroids and balance the non-hazardous over hazardous from 450:1 to 1:1. After applying SMOTE, the recall improves from 0.9874 to 0.9922. Improvement from recall is important since we want to identify as many hazardous asteroids as possible.

Part V: Discussions and Critics

Random Forest Model achieves a relatively satisfactory result. In the future, I will tune the model more comprehensively to achieve even better results if I have sufficient GPU resources. Besides, due to the lack of domain knowledge, I cannot explain these discovered important features using Astrophysical knowledge. In the future, after spending more time studying these astrophysical measurements, I will provide more domain-specific interpretations and reveal more facts about the hazardous asteroids themselves.

Part VI: Conclusions

Random Forest model achieves the precision is 0.9838, recall is 0.9922, accuracy is 0.9999, and F1 score 0.988. It predicts the rare hazardous asteroids with a high recall and precision. It can be integrated in building the global asteroids early-warning system. Unsupervised PCA provides domain experts some insights about where to look for potential hazardous from the PC1 vs. PC2 plot. SMOTE resample method is highly effective in handling the imbalance issue of the positive/hazardous case.

Part VII: References

1. https://www.pnas.org/content/117/29/17084.short
2. https://www.aanda.org/articles/aa/full_html/2020/02/aa35983-19/aa35983-19.html
3. https://www.sciencedirect.com/science/article/pii/S0094576517313747?casa_token=DYIpO2SgaWkAAAAA:Jwv8Is26Wc3TpksjUPeAIW3OzFdRb9YzUIuQYbhV5ucWXIUH92Idn9SFajh_rMCh3jQhLco7Ew
4. https://upload.wikimedia.org/wikipedia/commons/1/17/CLASSIFICATION_OF_BOLIDES_AND_METEORS_IN_DOPPLER_RADAR_WEATHER_DATA_USING_UNSUPERVISED_MACHINE_LEARNING_%28IA_classificationof1094564069%29.pdf
5. https://ssd.jpl.nasa.gov/sbdb_query.cgi
6. https://www.jair.org/index.php/jair/article/view/10302