ISSN 1870-4069

Determination of Hazardous Asteroids Using Machine Learning

Luis Vicarte, Juan Salas, Alfredo Murillo, Hiram Ponce

Universidad Panamericana,
Facultad de Ingeniería,
Mexico

{0199800, 0183674, 0252446, hponce}@up.edu.mx

Abstract. Potentially Hazardous Asteroid, or PHA, are near-Earth asteroids with a minimum orbital intersection distance of 0.05 AU or less with the Earth. This distance is roughly one-twentieth of the Earth-Sun mean distance, and it is thought to be the biggest conceivable orbital perturbation within a 100-year time frame that might result in a collision. This work proposes to build a machine learning model that can be used to provide a preventive forecast that confirms which asteroid are harmful to the Earth and which are not. After an analysis of a public dataset in asteroid and comparing four machine learning models, we found that the XGBoost model performs with 98% of accuracy. We use this model to find the most relevant features in the dataset, and we validated this information with the literature.

Keywords: Asteroids, artificial intelligence, support vector machines, XGBoost, classification

1 Introduction

Potentially Hazardous Asteroid, or PHA, are near-Earth asteroids with a minimum orbital intersection distance of 0.05 AU or less with the Earth. This distance is roughly one-twentieth of the Earth-Sun mean distance, and it is thought to be the biggest conceivable orbital perturbation within a 100-year time frame that might result in a collision [1].

PHAs account for around 20 percent of near-Earth asteroids. Asteroids are thought to have a chance of colliding with Earth, causing damage ranging from minor local destruction to mass extinction [2].

The fall of rock or iron asteroids greater than 50 m in diameter occurs every hundred years on average, causing local disasters and tidal waves. Asteroids greater than a kilometer create global disasters every few hundred thousand years. In the latter instance, the debris from the collision spreads across the Earth's atmosphere, causing acid rain, partial sunlight interruption, and massive flames created by high-temperature particles that fall to the Earth following the collision [3].

We consider these are scientifically isolated phenomena with significant significance for our world. As part of our study, we propose to identify those that are particularly hazardous to the globe and develop a machine learning model that can reliably

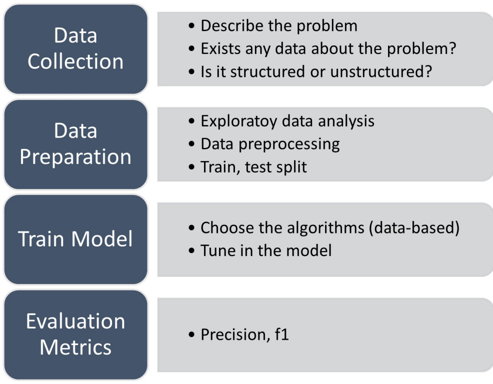

Fig. 1. Machine Learning Workflow.

anticipate them. According to new data from the Jet Propulsion Laboratory, Earth is surrounded by about 28,000 asteroids of all forms and sizes.

However, no evidence exists that any are now in the process of colliding with our planet [4]. This work proposes to build a machine learning model that can be used to provide a preventive forecast that confirms which asteroids are harmful to the Earth and which are not.

The rest of the paper is organized as follows. Section 2 summarizes the related work using machine learning for asteroids identification. Section 3 describes the methodology of the work, including the exploratory analysis over the dataset used, the data prepatation, the machine learning models and the experimentation. Section 4 shows the experimental results and discussion. Finally, Section 5 concludes the paper.

2 Related Work

Machine learning algorithms have recently been utilized to identify members of asteroid families and identify resonant argument pictures of asteroids in three-body resonances, among other uses, in the field of asteroid identification [5]. In comparison, machine learning applications to Solar System bodies, a broader topic that encompasses imaging and spectrophotometry of tiny bodies, have previously been classed as in progress [5].

The application of machine learning leads to the identification of new celestial objects or features, and research groups and procedures are more established. Machine learning is commonly used to study asteroid dynamics, although it is still in its infancy, with smaller groups and fewer articles yielding results [6].

Large observational studies of asteroids, like as those carried out at the Vera C. Rubin Observatory, will yield crucial data sets on their physical and orbital features [7]. ML applications are now being developed for clustering, picture recognition, and anomaly detection, among other things, and are predicted to be quite useful [8].

In contrast, our study focuses primarily on a mathematical and computational approach rather than astrophysics or space. Many solutions to these difficulties include imaging and visualization challenges, which our project does not consider at all and is exclusively dependent on machine learning model execution.

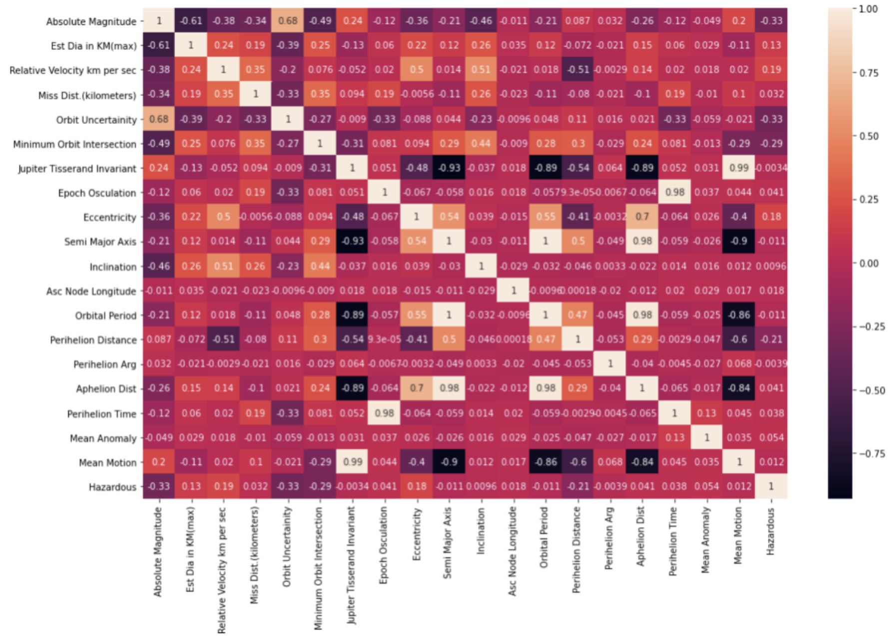

Fig. 2. Correlation Analysis.

3 Methodology

We adopt the classical workflow of machine learning, as depicted in Fig. 1. It comprises four main steps: data collection, data preparation, train model, and evaluation model.

3.1 Data Collection

The data is about Asteroids and is provided by NEOWS (Near-Earth Object Web Service[1]). This collection contains 4,687 rows and 41 columns.

In summary, our variables include the asteroid's name, diameter, proximity to the Earth, inclination, and distance from the Earth's orbit, among others. The variables are mostly integers, floats, and strings.

3.2 Data Preparation

The exploratory data analysis began with a review of null or zero values, which were replaced by their mean to be used in the model; it was also reviewed if the number of nulls was not very significant, eliminating these values to eliminate any possibility of noise and that they did not affect the total number of classes and asteroids; and it was also reviewed if the number of nulls was not very significant, eliminating these values to eliminate any possibility of noise and that they did not affect the total number of classes and aster.

[1] http://neo.jpl.nasa.gov

Table 1. Evaluation of the models.

Model	Accuracy	Precision	Recall	F1-score
LR	0.95	0.89	0.91	0.90
SVM	0.96	0.90	0.90	0.90
XGBoost	1.0	0.99	1.0	0.99
MLP	0.95	0.90	0.91	0.91

Following that, we conducted a correlation analysis of variables, as shown in Fig. 2, discovering that the mean movement is connected to Jupiter's invariant; oscillation is related to time; and the period of orbit is related to distance, among others.

We deleted the data columns for the variables that were not important after selecting the most and least important variables, which were: Unnamed; Neo Reference ID; Est Day in Miles, min, max; Orbiting distance; Relative velocity in kilometers per hour.

The "Hazardous" variable was then label encoded such that the model 0 or 1 informs us if it is harmful to the earth or not. The preceding is critical in ensuring that the model functions properly and that the data is normalized.

3.3 Machine Learning Models

We split the data into 80% for training and 20% for testing. We consider four machine learning models for our proposal:

– *Logistic regression (LR)*– it is a supervised learning model that allows classification based on the transformation of numerical values into categories (or classes) using the sigmoidal function. The training of a logistic regression uses the error function called logistic loss. It penalizes wrong predictions rather than reward correct predictions [9].
– *Support Vector Machines (SVM)*– Support vector machine (SVM) is a computer algorithm that learns by example to assign labels to objects!. For instance, an SVM can learn to recognize fraudulent credit card activity by examining hundreds or thousands of fraudulent and nonfraudulent credit card activity reports. Alternatively, an SVM can learn to recognize handwritten digits by examining a large collection of scanned images of handwritten zeroes, ones and soporth. SVMs have also been successfully applied to an increasingly wide variety of biological applications [10].
– *XGBoost*–xgboost is short for eXtreme Gradient Boosting package. It is an efficient and scalable implementation of gradient boosting framework. The package includes efficient linear model solver and tree learning algorithm. It supports various objective functions, including regression, classification and ranking [11].
– *Multilayer Perceptron (MLP)*–The perceptron algorithm is due to Rosenblatt in the late 1950s. The perceptron, a simple computing engine which has been dubbed a 'linear machine' is best related to supervised classification [12].

3.4 Experimentation

We use the following metrics to compare the performance of the models: accuracy (1), precision (2), recall (3), and F1-score (4); where, TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively:

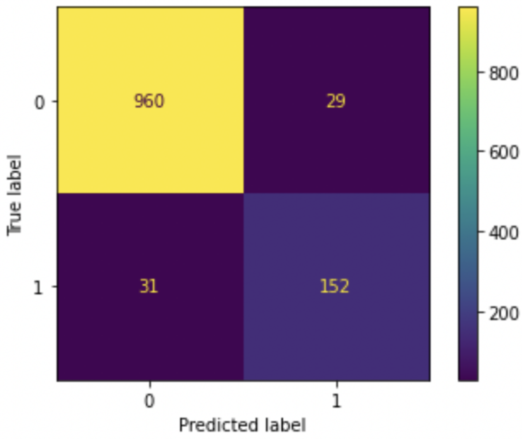

Fig. 3. Confusion matrix of the XGBoost model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{4}$$

4 Experimental Results

We conduct a 5-fold cross validation in each of the models, and Table 1 summarizes the results of the models in terms of accuracy, precision, recall, and F1-score.

As shown in the results, the XGBoost and the multilayered Perceptron models produced the greatest outcomes. In Fig. 3, we show the confusion matrix obtained by the predictions of the best model found so far.

Moreover, we found the most relevant variables in the data set using the XGBoost model, especifically its feature importance option, as shown in Fig. 4. They include the following:

- Absolute magnitude: the magnitude that a star would appear to have if it were located at a standard distance of 10 parsecs.
- Minimum orbit intersection: measure used in astronomy to assess potential close approaches and collision risks between astronomical objects.
- Inclination: measures the tilt of an object's orbit around a celestial body.
- Aphelion distance: the point in the orbit of an object where it is farthest from the Sun.

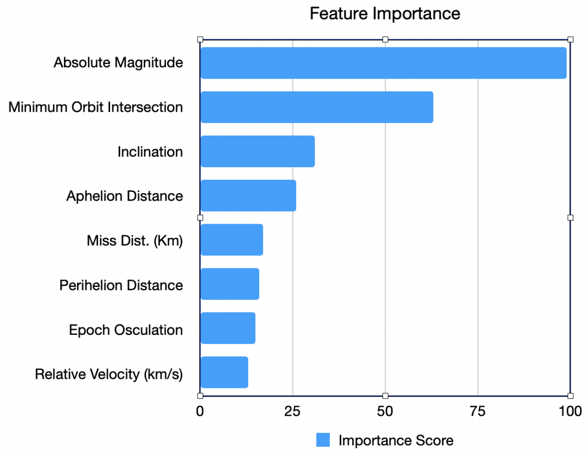

Fig. 4. Analysis of feature importance.

- Miss distance: the maximum distance at which the explosion of an artifact head can be expected to seriously damage its target.
- Perihelion distance: the point in orbit where an object is nearest to the sun.
- Epoch osculation: the instant of time at which the position and velocity vectors are specified.
- Relative velocity km per sec: the velocity of an object B in the rest frame of another object A.

From an astronomical standpoint, two primary characteristics are examined for an asteroid to be deemed dangerous, among which the following stand out: absolute magnitude and minimum orbit intersection, among many other considerations [13]. We may conclude from the above that our primary variables make sense in the real world.

5 Conclusions

In this work, we proposed to build a machine learning model that can be used to provide a preventive forecast that confirms which asteroids are harmful to the Earth and which are not. To do so, we identified the most important variables, ran tests with various classification models, and finally chose the best ones that produced high efficiency in this problem. As a result, we have results ranging from 94% to 100%.

For future work, we consider it is critical to continue gathering data on asteroids, feeding the model, and testing alternative algorithms in order to develop a more accurate method of detecting harmful asteroids for the planet in future research.

References

1. NASA Jet Propulsion Laboratory: NASA System Predicts Impact of Small Asteroid. Asteroids and Comets, https://www.jpl.nasa.gov/news/nasa-system-predicts-impact-of-small-asteroid/ (2022)

2. Bland, P.A., Artemieva, N.A.: The Rate of Small Impacts on Earth. Meteoritics and Planetary Science, vol. 41, no. 4, pp. 607–631 (2006). DOI: 10.1111/j.1945-5100.2006.tb00485.x.
3. NASA Jet Propulsion Laboratory: NASA's Next-Generation Asteroid Impact Monitoring System Goes Online. Asteroids and Comets, https://www.jpl.nasa.gov/news/nasas-next-generation-asteroid-impact-monitoring-system-goes-online (2021)
4. Carruba, V., Aljbaae, S., Domingos, R.C., Lucchini, A., Furlaneto, P.: Machine Learning Classification of New Asteroid Families Members. Monthly Notices of the Royal Astronomical Society, vol. 496, no. 1, pp. 540–549 (2020). DOI: 10.1093/mnras/staa1463.
5. Carruba, V., Aljbaae, S., Domingos, R.C., Barletta, W.: Artificial Neural Network Classification of Asteroids in the M1: 2 Mean-Motion Resonance with Mars. Monthly Notices of the Royal Astronomical Society, vol. 504, no. 1, pp. 692–700 (2021). DOI: 10.1093/mnras/stab914.
6. Förster, F., Cabrera-Vives, G., Castillo-Navarrete, E., Estévez, P.A., Sánchez-Sáez, P., Arredondo, J., Bauer, F.E., Carrasco-Davis, R., Catelan, M., Elorrieta, F., Eyheramendy, S., Huijse, P., Pignata, G., Reyes, E., Reyes, I., Rodríguez-Mancini, D., Ruz-Mieres, D., Valenzuela, C., Álvarez-Maldonado, I., Astorga, N., et al.: The Automatic Learning for the Rapid Classification of Events (ALeRCE) Alert Broker. The Astronomical Journal, vol. 161, no. 5 (2021). DOI: 10.3847/1538-3881/abe9bc.
7. Goldstein, M., Uchida, S.: A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. PloS One, vol. 11, no. 4 (2016). DOI: 10.1371/journal.pone.0152173.
8. Wright, R.E.: Logistic Regression. Reading and Understanding Multivariate Statistics, pp. 217–244 (1995)
9. Noble, W.S.: What is a Support Vector Machine? Nature Biotechnology, vol. 24, no. 12, pp. 1565–1567 (2006). DOI: 10.1038/nbt1206-1565.
10. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K.: Xgboost: eXtreme Gradient Boosting. R package version 0.4-2, vol. 4, no. 1, pp. 1–4 (2015)
11. Murtagh, F.: Multilayer Perceptrons for Classification and Regression. Neurocomputing, vol. 2, no. 5-6, pp. 183–197 (1991) DOI: 10.1016/0925-2312(91)90023-5.
12. Shustov, B.M., Shugarov, A.S., Naroenkov, S.A., Prokhorov, M.E.: Astronomical Aspects of Cosmic Threats: New Problems and Approaches to Asteroid—Comet Hazard Following the Chelyabinsk Event of February 15, 2013. Astronomy Reports, vol. 59, no. 10, pp. 983–996 (2015). DOI: 10.1134/S1063772915100066.