Hazardous Asteroid Classification with Machine Learning using Physical and Orbital Asteroid Properties

Arjun Ramakrishnan

Abstract

Asteroids, rocky objects orbiting the sun, have been a key focus of scientific study as they can provide insights into planet formation. With a seemingly infinite number of asteroids in space, the possibility of one colliding with our planet and leading to devastating effects constantly looms large. Asteroids that could come in close proximity or collide with earth are classified as potentially hazardous asteroids, PHA (NASA, n.d.). However, it becomes cumbersome for humans to manually analyze large datasets for identifying all the possibly dangerous asteroids. Thus, machine learning techniques are ideal to study trends and make predictions. Machine learning is a method of data analysis based on computer algorithms that model relationships and improve our ability to analyze asteroid threats. It has been applied to automate the asteroid classification process in the past, for instance by Anish Si in 2018 at the Vellore Institute of Technology in India, where his 15-tree Random Forest model performed the best (Si, 2018). The goal of this study was to train multiple machine learning models on physical and orbital asteroid features and identify the model that most accurately classified the asteroids as hazardous or non-hazardous. The key enhancements were that a different subset of features and significantly different list of models were used for classification. The results showed that a 50-tree Random Forest classification model had a 98.45% accuracy on the test set validating that the Random Forest is the most optimal model for asteroid classification.

Keywords

Potentially Hazardous Asteroids, Machine Learning, Random Forest Classification

Introduction

Asteroids are considered the building blocks of our solar system. They are chunks of rock in outer space that are smaller than a planet and orbit the sun. The majority of them are known to be found in the asteroid belt between Jupiter and Mars, however the orbit of an asteroid could alter due to gravity exerted by other objects or collisions, putting the asteroid on the path to impact earth. There have been numerous asteroid collisions with Earth throughout time, for instance in Chelyabinsk, Russia in 2013 when an asteroid twenty meters wide exploded upon entering the Earth's atmosphere (The Planetary Society, n.d.). The shockwave from the explosion was so strong that it injured more than a thousand people and destroyed buildings across six cities. The Chicxulub crater marks the site of an asteroid impact about 65 million years ago in the Gulf of Mexico region where an asteroid approximately ten to fifteen kilometers wide impacted Earth, eradicating 70% of our planet's species (The Planetary Society, n.d.). Asteroid 2022 EB5 struck the north of Iceland as recently as 2022 and was the fifth asteroid to be discovered before impact with Earth (NASA, 2022). Asteroid impacts remain a constant threat and a future impact could easily cause widespread devastating damage to Earth and its inhabitants. Hence, it is crucial to keep researching asteroids and their properties, with a focus on those that can experience close approaches to Earth with a greater potential for impact. These asteroids are currently classified as potentially hazardous asteroids or PHA. The PHA classification is primarily based on the physical properties of the asteroid, such as its size and its orbital properties like the orbital period and tilt of the orbit. Specifically, asteroids with a minimum orbit intersection distance of 0.05 au or less and an absolute magnitude of 22 or less are classified as PHA. Minimum orbit intersection distance is the closest distance the asteroid is to earth in its orbit. Absolute magnitude refers to the brightness of the object as seen from a fixed distance (NASA, n.d.). By improving the classification process of PHA, scientists can identify future asteroid threats quicker, allowing for more efficient evacuations and minimizing possible destruction from the event. The landscape in outer space is constantly changing, and it is important to have an accurate mechanism to study and identify hazardous asteroids and gain a deeper understanding of our solar system.

Given the sheer number of asteroids in outer space and the volume of data for their physical and orbital trajectory properties, it is not feasible for humans to manually process it in a timely manner to make educated decisions about how dangerous orbiting asteroids could be.

Thus, machine learning techniques are ideal to study trends and make predictions. Machine learning has been applied to hazardous asteroid classification in the past, as shown in a paper titled *Hazardous Asteroid Classification through Various Machine Learning Techniques* by Anish Si. In this study, logistic regression, support vector machine, decision tree, K nearest neighbor, random forest, naïve bayes, adaboost and xgboost machine learning algorithms were evaluated on an asteroid dataset with 755 hazardous asteroids and 3932 non-hazardous asteroids. The dataset used in the study contained the following asteroids features: absolute magnitude, minimum orbit intersection, Ascending node longitude, orbit uncertainty, perihelion time, inclination, semi major axis, Anomaly, perihelion arguments, perihelion time, relative velocity, perihelion distance, eccentricity, aphelion distance and Jupiter Tisserand Invariant. The random forest (15-tree) and xgboost models were found to have the highest performance of 100% accuracy, while naïve bayes gave the lowest at 80.70%.

In this research study, a significantly different set of machine learning models and features were trained to automate the hazardous asteroid classification process given physical and orbital asteroid properties obtained from a NASA asteroid database in an attempt to identify the most performant model for asteroid classification.

Methodology

Aim of the Study

The aim of this study was to use machine learning to accurately classify hazardous asteroids. Seven machine learning models were trained on a training subset of the data and tested for performance on a development subset of the data. The machine learning model with the best accuracy on the development set was planned to be tested on a test set, a portion of the data set aside before training that the models had never seen before.

Research Design

In this exploratory research study, the independent variable was the machine learning model trained on the asteroid dataset, and the dependent variable was the model’s ability to accurately classify hazardous asteroids on the development set, which was quantified by its accuracy score..

First, the JPL database was determined as the dataset source, and a portion of the database with the appropriate objects and features was accessed. After downloading the dataset, missing values in the dataset, specifically five missing values in the absolute magnitude feature, were filled with the mean of all the absolute magnitude values. The data was then split into training, development, and test sets, 80% of the data being for training and 10% each being for the testing and development sets. Feature scaling was applied to all the data, meaning the data for each feature was scaled to a normal distribution with a mean of zero and unit standard deviation, accounting for any possible bias towards larger values in the original dataset before scaling. Each model was then implemented, trained on the training data and tested on the development set to get a baseline for its performance. For each algorithm, the performance was evaluated using either an R^2 value for regression models and a confusion matrix and accuracy score otherwise.

The performance of the 4 best algorithms, which was determined by model performance on the development set, were then attempted to be improved through the use of k-fold cross validation and grid search (Nadeem, 2020). Of these algorithms, the highest performing algorithm was evaluated on the unseen test set.

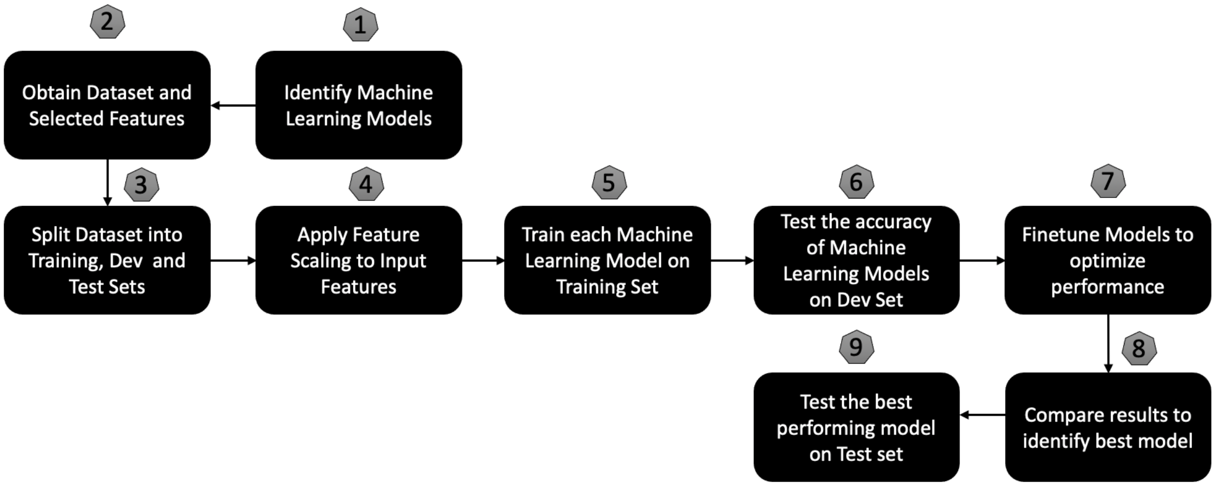
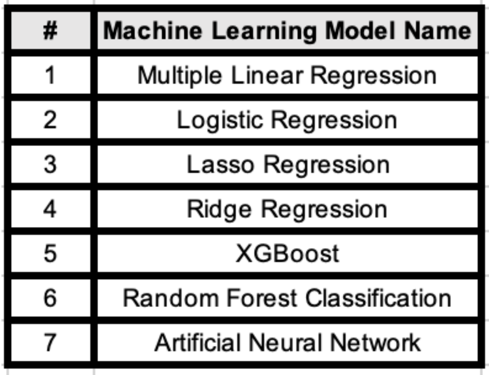

2
Obtain Dataset and Selected Features
1
Identify Machine Learning Models
3
Split Dataset into Training, Dev and Test Sets
4
Apply Feature Scaling to Input Features
5
Train each Machine Learning Model on Training Set
6
Test the accuracy of Machine Learning Models on Dev Set
7
Finetune Models to optimize performance
8
Compare results to identify best model
9
Test the best performing model on Test set

Figure 1: Flowchart of Procedure for Study

Scales used/tools used/instruments used

Coding for the machine learning algorithms was done using the programming language Python. Python libraries for machine learning allow easy access and transformation of data making the implementation of algorithms efficient. The following machine learning models were implemented in this study: multiple linear regression, logistic regression, lasso regression, ridge regression, XGBoost, random forest classification, and artificial neural network as shown in Table 1. The artificial neural network contained two hidden layers with nine nodes each (Chavan, 2013), a 40% dropout layer, and a sigmoid activation function. After evaluating each algorithm on the development set, k-fold cross validation and grid search were applied to the logistic regression (Jain, 2021), XGBoost (Tilli, 2017), random forest classification (Koehrsen, n.d.), and artificial neural network (Brownlee, 2019; Brownlee, 2022) models.

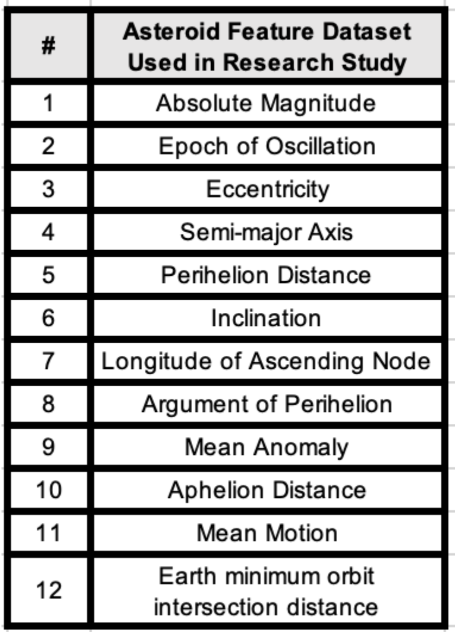


#	Machine Learning Model Name
1	Multiple Linear Regression
2	Logistic Regression
3	Lasso Regression
4	Ridge Regression
5	XGBoost
6	Random Forest Classification
7	Artificial Neural Network

Table 1: Machine learning models implemented and tested in study

Data Collection Procedure

The dataset on which the models were trained and evaluated was a subset of the Small Body Database from the NASA Jet Propulsion Lab at the California Institute of Technology (JPL,n.d.). It contained data about asteroids that were noted as Near Earth Objects (NEOs), or objects whose orbits allow them to pass extremely close to Earth. The dataset contained 29,150 different asteroids, 2257 of which were classified as potentially hazardous, while 26,893 were not classified as such. Both physical and orbital properties of the asteroids were selected as inputs to the models, as these are part of the analytical criteria for determining whether an asteroid is potentially hazardous or not. The specific features are absolute magnitude, epoch of oscillation, eccentricity, semi-major axis, perihelion distance, inclination, longitude of ascending node, argument of perihelion, mean anomaly, aphelion distance, mean motion, and Earth minimum orbit intersection distance as shown in Table 2.

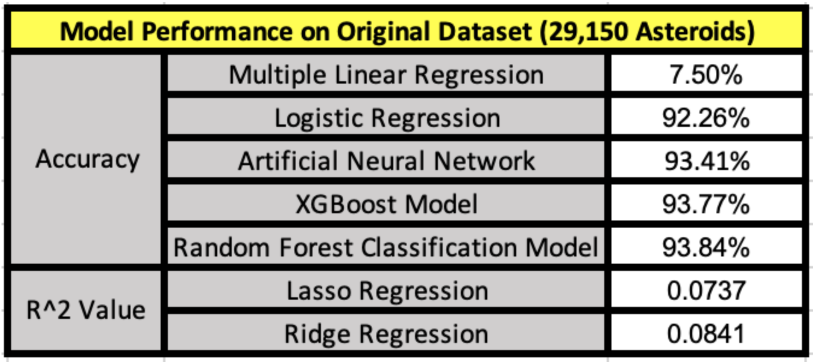


#	Asteroid Feature Dataset Used in Research Study
1	Absolute Magnitude
2	Epoch of Oscillation
3	Eccentricity
4	Semi-major Axis
5	Perihelion Distance
6	Inclination
7	Longitude of Ascending Node
8	Argument of Perihelion
9	Mean Anomaly
10	Aphelion Distance
11	Mean Motion
12	Earth minimum orbit intersection distance

Table 2: Features in dataset used for classification prediction

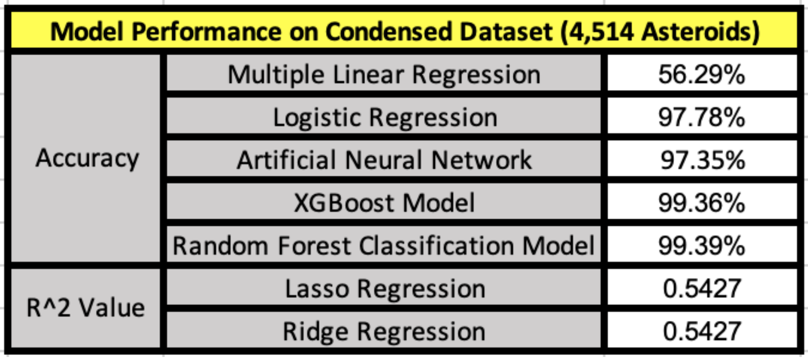
A condensed version of the dataset with the 2257 hazardous asteroids and a random sample of 2257 non-hazardous asteroids was then created and the process of testing the models was repeated. Due to the large proportion of non-hazardous asteroids compared to hazardous, it was found that the models were struggling to correctly classify hazardous asteroids. Thus, this condensed dataset sought to resolve this issue by presenting an equal proportion of both hazardous and non-hazardous asteroids to facilitate better classification.

Results



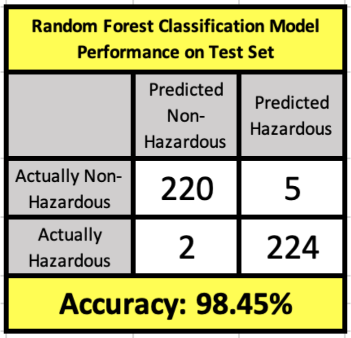
Model Performance on Original Dataset (29,150 Asteroids)		
Accuracy	Multiple Linear Regression	7.50%
	Logistic Regression	92.26%
	Artificial Neural Network	93.41%
	XGBoost Model	93.77%
	Random Forest Classification Model	93.84%
R^2 Value	Lasso Regression	0.0737
	Ridge Regression	0.0841

Table 3: The table shows the performance of each model on the original dataset of 29,150 asteroids



Model Performance on Condensed Dataset (4,514 Asteroids)		
Accuracy	Multiple Linear Regression	56.29%
	Logistic Regression	97.78%
	Artificial Neural Network	97.35%
	XGBoost Model	99.36%
	Random Forest Classification Model	99.39%
R^2 Value	Lasso Regression	0.5427
	Ridge Regression	0.5427

Table 4: The table shows the performance of each model on the condensed dataset with equal numbers of hazardous and non-hazardous asteroids



Random Forest Classification Model Performance on Test Set		
	Predicted Non-Hazardous	Predicted Hazardous
Actually Non-Hazardous	220	5
Actually Hazardous	2	224
Accuracy: 98.45%		

Table 5: The table shows the final evaluation and confusion matrix of the best performing model, random forest classification, on an unseen test set

Discussion

As shown in Table 3, when the models were tested on the original asteroid dataset containing 29,150 asteroid data points, the Random Forest Classification model, containing 50 trees, performed the best with an accuracy of 93.84%. Similarly, as shown in Table 4, on the condensed dataset, the Random Forest had the highest performance again at 99.39%. In both cases, the XGBoost model matched the performance of the random forest model, staying within 0.1% of that of the random forest model. These results match the trends of those detailed in Si's study, who also concluded that the Random Forest and XGBoost models performed the best for this task. When using the original entire dataset, the multiple linear and lasso regression models performed the worst, while in the condensed dataset case, the lasso and ridge regression models performed the worst.

The algorithm with the highest performance on the development set, the 50 tree Random Forest Classification model, was then tested on the test set data set aside from the rest of the dataset initially. As seen in Table 5, the model maintained its performance well, with a 98.45% accuracy on this unseen test set. Specifically, the model only incorrectly classified two hazardous asteroids as non-hazardous and five non-hazardous asteroids as hazardous. The high performance of the model on the development and test sets indicated that minimal overfitting was present in the model.

Conclusion

In this study, multiple machine learning models were used for hazardous asteroid classification based on twelve asteroid features encompassing different physical and orbital asteroid properties. The 50-tree Random Forest classification model performed the best at identifying hazardous asteroids, which was reflected in its consistent high performance on the development and test set data. The results of this study could be applied to further develop and improve the field of asteroid detection and classification, leading to more accurate classifications and faster response times improving the security of our planet as a whole.

Limitations

In this study, only k-fold cross validation was implemented to improve model performance, as the grid search did not show any significant improvement for any of the models.

If grid search were able to be implemented, the performance of this algorithm could have improved by optimizing algorithm parameters such as the number of trees, nodes per tree, and tree depth. To further improve this study, a larger dataset with the same proportion of hazardous and non-hazardous asteroids could be used. Also, feature-condensing algorithms like Principal Component Analysis could have been applied to all the features in the JPL database to obtain the asteroid features most directly related to its classification as hazardous.

Acknowledgements

I would like to thank Jon P. Stingel for his mentorship and guidance on this project.

References

Brownlee, J. (2019, August 6). *How to improve deep learning performance*. Machine Learning Mastery. Retrieved August 3, 2022, from https://machinelearningmastery.com/improve-deep-learning-performance/

Brownlee, J. (2022, August 4). *How to grid search hyperparameters for deep learning models in python with keras*. Machine Learning Mastery. Retrieved August 5, 2022, from https://machinelearningmastery.com/grid-search-hyperparameters-deep-learning-models-python-keras/

Chavan, P. (2013, January 24). *How to decide the number of hidden layers and nodes in a hidden layer?* Research Gate. Retrieved July 28, 2022, from https://www.researchgate.net/post/How-to-decide-the-number-of-hidden-layers-and-nodes-in-a-hidden-layer

Jain, K. (2021, March 14). *How to improve logistic regression?* Medium. Retrieved June 30, 2022, from https://medium.com/analytics-vidhya/how-to-improve-logistic-regression-b956e72f4492

Koehrsen, W. (2018, January 10). *Hyperparameter tuning the random forest in python*. Medium. Retrieved July 27, 2022, from https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Nadeem, Maryam. (2020, November 26). *Hyperparameter tuning using GRIDSEARCHCV and Kerasclassifier*. GeeksforGeeks. Retrieved August 5, 2022, from https://www.geeksforgeeks.org/hyperparameter-tuning-using-gridsearchcv-and-kerasclassifier/

NASA. (2022, March 15). *NASA system predicts impact of small asteroid*. NASA. Retrieved July 7, 2022, from https://www.jpl.nasa.gov/news/nasa-system-predicts-impact-of-small-asteroid

NASA. (n.d.). *Neo basics*. NASA. Retrieved June 30, 2022, from https://cneos.jpl.nasa.gov/about/neo_groups.html

NASA. (n.d.). *Small-body database query*. NASA. Retrieved June 23, 2022, from https://ssd.jpl.nasa.gov/tools/sbdb_query.html

Notable asteroid impacts in Earth's history. The Planetary Society. (n.d.). Retrieved July 7, 2022, from https://www.planetary.org/notable-asteroid-impacts-in-earths-history

Si, A. (2020, March). Hazardous Asteroid Classification through Various Machine Learning Techniques. Tamil Nadu; International Research Journal of Engineering and Technology .

Tilli, Dan. (2017, October 13). *Hyperparameter grid search with XGBoost*. Kaggle. Retrieved July 27, 2022, from https://www.kaggle.com/code/tilii7/hyperparameter-grid-search-with-xgboost/notebook