Predicting Hazardous Asteroids Using Machine Learning

Komal Gohel
Stevens Institute of Technology
Hoboken, United States
rgohel1@stevens.edu

Jason Bauer
Stevens Institute of Technology
Hoboken, United States
jbauer5@stevens.edu

Nese Morali
Stevens Institute of Technology
Hoboken, United States
nmorali@stevens.edu

***Abstract*—Abstract—Dozens of asteroids pass the Earth every single day. Our project focuses on classifying asteroids as potentially hazardous or not using data obtained from the Jet Propulsion Laboratory of California Institute of Technology(JPL). Since potentially hazardous asteroids (PHA) can only be classified as yes or no, we decided that using logistic regression as a predictive model made sense. Before we built our model, we cleaned up the data by removing any unnecessary columns. We used two different feature selection algorithms when setting up our logistic model, RFE and Random Forest. Our logistic regression achieved an accuracy of 100 percent, which we will further explore in other sections. An accurate model could identity a PHA quickly so that NASA and other organizations have plenty of time to prepare if need be.**

I. Introduction

For all of Earth's history, asteroids have made an impact on the planet as well as its inhabitants. When people think of hazardous asteroids they usually imagine a giant asteroid that will crash into Earth and cause destruction like in the movies. In a real-world dataset, most of the hazardous asteroids we look at are smaller asteroids. These small asteroids can still cause millions of dollars' worth of damage to buildings, space stations, or satellites. Luckily for us, NASA has been working on techniques to deflect dangerous asteroids from hitting Earth. A key example of this is NASA's recent DART mission. This mission was designed to test asteroid deflection by changing the asteroid's direction through kinetic impact. The deflection technique is extremely important, but NASA must first detect an asteroid before they can deflect it. This is the idea we will be looking at in this paper. Right now, many of these calculations are done by hand. There currently is no standard for how to solve this problem via machine learning. This project will aim to reduce the workload for calculations. We believe this is achievable with an accurate machine learning model.

II. Related Work

One model created by Radbeen and Denneau 2021 used a neural net to detect hazardous asteroids based on the ATLAS dataset. The big difference between this project and ours is that the ATLAS database is comprised entirely of images. Our JPL dataset is entirely numerical. Another model created by Trotter McLemore 2022 created an XGBoost model based on the NeoWs dataset. This dataset uses XGBoost where we will be using Random Forest.

III. Our Solution

The most important part of solving this problem is selecting the proper variables. Real life asteroid data collection can be much simpler if we know exactly which variables predict hazardous asteroids. We will run multiple machine learning

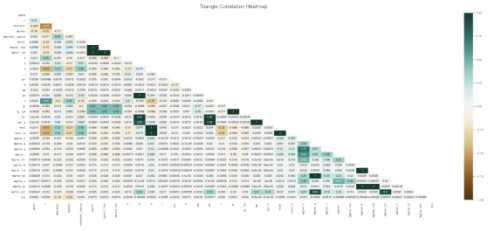

Fig. 1. Correlation Heat Map

algorithms and compare them to have a better idea of the ideal model for predicting hazardous asteroids.

A. Description of Dataset

The data used for this project is from the Jet Propulsion Laboratory of California Institute of Technology(JPL). This lab is funded by NASA and was created to share data about astronomical data to the public. This dataset contains information about 958,524 different asteroids, where around 98 percent are not considered hazardous. Many variables in our dataset are directly related, for example moid and moidld are both earth minimum orbit intersection distance. The difference is that moidld is in lunar units and moid is in au units. We also removed columns that were not relevant for analysis. For example, fullname and prefix were removed because the name of the asteroids is not relevant and the prefix column was 100 percent NULL. As part of our preprocessing, we removed any rows with NaN values. We also converted the neo and pha columns to numeric by replacing 'Y' and 'N' with 1 and 0 respectively. The next step in our data cleaning was looking for perfectly correlated variables. Perfectly correlated variables will decreased the accuracy of a logistic regression so we wanted to remove any that could hurt the accuracy of the model. We plotted a correlation matrix with a heatmap to see the relationships between all variables. After we cleaned the dataset, we saved 28 of the original 45 columns.

B. Machine Learning Algorithms

We decided to use a logistic regression model for this problem. Logistic regression is powerful when the target variable has binary outcomes so we thought this model could be effective. Before we could fit our regression model, we first needed to select the features of the model. Our correlation matrix offered little help in selecting features for our model. The next thing we tried was using Random Forest Classification to select the features. Once we had our features, we fitted the logistic model.

C. Implementation Details

The Random Forest Classification selected 'neo', 'H', 'diameter', 'q', and 'moid' as the five most important features for determining 'pha'(Potentially hazardous asteroids). Once we fit these features into the logistic regression, we first assessed the model accuracy. We achieved an accuracy of 99.88 percent. We knew this level of accuracy was highly unlikely and wanted to validate our results with a confusion matrix and ROC curve. Both graphics confirmed our outrageously high accuracy value. We believe this unrealistic accuracy value stems from the ratio of non-hazardous asteroids to hazardous asteroids being extremely unbalanced. We will look into solutions for this issue.

IV. Comparison

1) 10 fold Stratified cross validation technique which train data and fitted train data to four different model and give accurate data. However, two of them algorithm are gives more accurate data which is not good

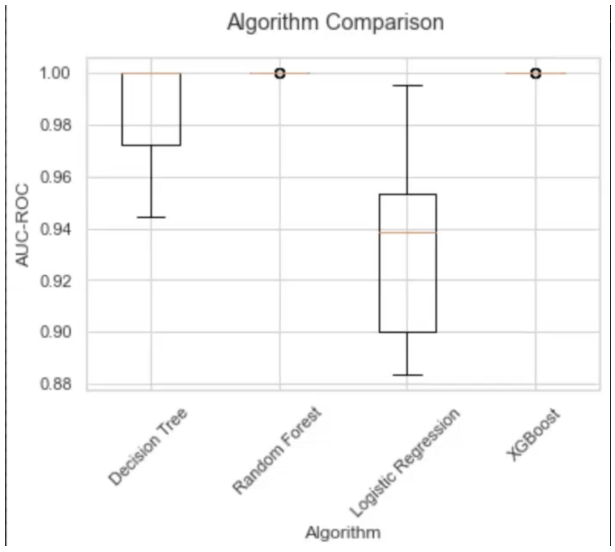

Fig. 2. Algorithm Comparison

2) We don't see any existing solution for out dataset. But in implementation of four algorithm Basic decision trees give pretty well results.Logistic regression is not very accurate. On other hand, Random forest is almost accurate. XGBoost is perfact which gives us red flag.

V. Future Directions

For make it better we can use SMOTE technique for balancing out our binary outcome

VI. Conclusion

XBoost and Random forest algorithm gives almost 100 accuracy most likely because overfitting in data also in dataset it's hard to tell which is good data column that we should not deleted. So there have chances that we deleted columns which suppose to be not deleted. Currently, basic dicision tree gives really good result but in future SMOTE technique will gives more accurate results to this problem dataset

References

1) Rabeendran, A. C., and Denneau, L. (2021). A TwoStage Deep Learning Detection Classifier for the ATLAS Asteroid Survey. arXiv.
2) McLemore (2022). Classifying Hazardous and NonHazardous Asteroids Using Machine Learning.
3) Hossain, M.S., Zabed, M.A. (2023). Machine Learning Approaches for Classification and Diameter Prediction of Asteroids. In: Ahmad, M., Uddin, M.S., Jang, Y.M. (eds) Proceedings of International Conference on Information and Communication Technology for Development. Studies in Autonomic, Data-driven and Industrial Computing. Springer, Singapore.
4)https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
5)https://nhsjs.com/2022/classifying-hazardous-and-non-hazardous-asteroids-using-machine-learning/
6)https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390
https://www.nvidia.com/en-us/glossary/data-science/xgboost/