PREDICTING ASTEROID DIAMETER FOR EARTH IMPACT ASSESSMENT

GEORGETOWN UNIVERSITY

TEGVEER GHURA, ANTHONY MOUBARAK, RAMDAYAL REWARIA, RAGHAV SHARMA

Introduction

Asteroids have long held a significant role in the rich history of our solar system, as they have existed since its very formation. Throughout the ages, these celestial bodies have impacted Earth on multiple occasions, resulting in mass extinctions and leaving indelible imprints on our planet. Consequently, esteemed space organizations such as NASA diligently monitor asteroids due to their potential threat to Earth, and comprehending their behavior and trajectory is of paramount importance in developing efficacious mitigation strategies.

Precisely determining the diameter of an asteroid is a critical factor in comprehending its potential impact on Earth and formulating appropriate mitigation strategies. Even a minor disparity in diameter can prove to be the demarcation between a benign atmospheric entry and a calamitous impact with far-reaching consequences.

Data Science Problems

1. Can Machine Learning Models help predict the diameter of an incoming asteroid based on some of its features? How do the models' metrics vary between them?
2. What are the most important asteroid features that help predict the diameter of an asteroid ?

Dataset

This project utilized data provided by NASA's Jet Propulsion Laboratory, which contains over 20 numerical variables that describe more than 130,000 asteroids. However, some features had to be dropped due to a high number of empty values and/or being populated exclusively by one or two values.

Methods

1. **XGBoost** – Uses gradient boosting and decision tree-based techniques to achieve high predictive accuracy and speed on structured datasets.
2. **MLPRegressor** – Uses multi-layer perceptron architecture for predicting continuous numeric values.
3. **Gradient Boosting** – Uses an ensemble of weak prediction models, such as decision trees, to iteratively correct the errors of previous models and achieve high accuracy on supervised learning tasks.
4. **Random Forest Regressor** – An ensemble machine learning algorithm that uses multiple decision trees and bootstrap aggregation to predict continuous numeric values.

Data Exploration

- Correlation of various features of the asteroids are checked against the diameter of the asteroid.
- Features such as Absolute magnitude, Mean Anomaly, Perihelion Distance, Eccentricity are highly correlated with Diameter of the Asteroid which warrants feature selection.
- Aphelion Distance and Perihelion Distance are highly correlated with each other, which induces multicollinearity and hence one of them should be dropped.

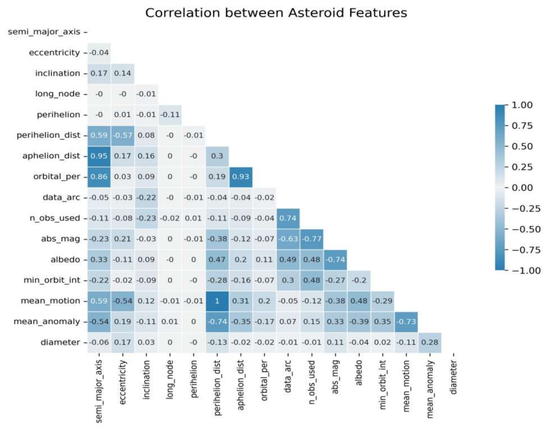

Figure 1 – Heat Map depicting the correlation between asteroid features.

Feature Selection

- After cleaning the dataset, a total of 19 predictors were left. This warranted us to employ dimensionality reduction methods to avoid multicollinearity. These methods include best subset selection, forward selection, backward selection, and LASSO regression.
- Best subset selection, forward selection, and backward selection returned 13 variables. LASSO Regression returned 4 variables, as seen in the coefficient plot using the log of lambda:

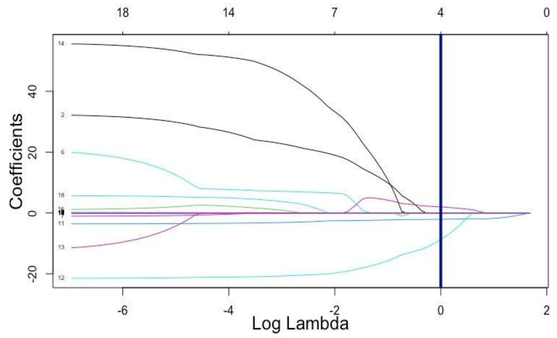

Figure 2 – LASSO Regression simplifies the parameter space without compromising performance. At log lambda = 0, 4 variables remained.

Results

RMSE (Root Mean Squared Error) is the metric used to compare the four models, as it measures the difference between the predicted and actual values. It provides an intuitive understanding of the error magnitude, is sensitive to outliers, and can be easily interpreted in the context of the problem domain.

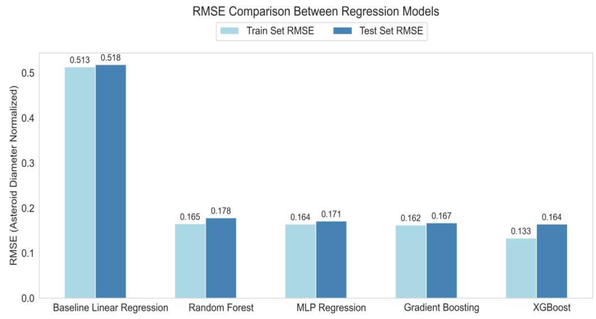

Figure 3 – RMSE comparison of Machine Learning models

All four models were trained on the four selected features and compared against a baseline Linear Regression model trained on the same variables. All models achieved better RMSE scores than the baseline LR model, with XGBoost achieving the lowest Train RMSE of 0.133 and the lowest Test RMSE of 0.164.

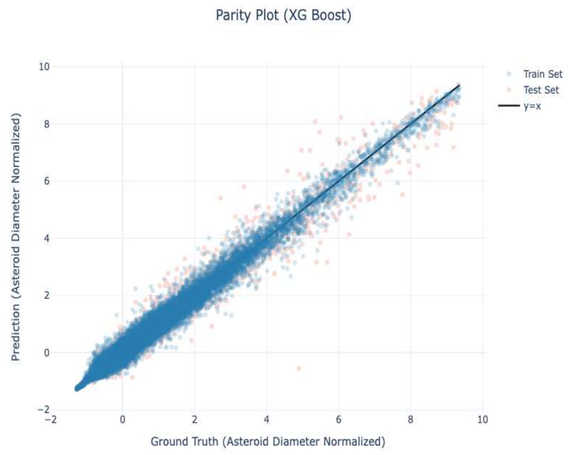

Figure 4 – Parity Plot for Random Forest Regression Model

The parity plot presented above in Figure 4 illustrates the best performing model, XGBoost, which provides further evidence of its accuracy. Moreover, the plot suggests that the model's performance is better for smaller Y values, as the deviation from the line increases for larger Y values.

Conclusion

- The variables with the most predictive power with respect to an Asteroid's diameter are number of days spanned by the data arc, minimum orbit intersection, albedo, and absolute magnitude of the asteroid.
- Asteroids have previously entered Earth's atmosphere, and if we can accurately predict the diameter of an incoming asteroid, it could help us take necessary safety measures. Additionally, institutions such as NASA and ISRO could develop mitigation strategies to handle incoming near-Earth objects.
- Ultimately, this project serves as a strong proof of concept for the application of machine learning in the field of space exploration.

Limitations

The dataset had several limitations, including a high prevalence of missing values. Another significant limitation was the absence of a year column. Specifically, the CNEOS organization's Close Approach Database provided the Close-Approach Date of the asteroids, including the ones in the future, up to year 2100. NASA's Jet Propulsion Laboratory database, analyzed in our paper, did not include these dates. Therefore, knowing the year of each close approach would have allowed for a comparison of asteroid characteristics over time.

References

1. https://www.kaggle.com/datasets/basu369victor/prediction-of-asteroid-diameter
2. https://ssd.jpl.nasa.gov/tools/sbdb_query.html
3. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
4. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
5. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html
6. https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

Acknowledgements

We would like to thank Dr. Hickman, Dr. Nakul, Dr. Purna, and Georgetown University Graduate School of Arts & Sciences. We would also like to acknowledge the Data Science & Analytics (GSAS) program at Georgetown University.