PREDICTING DIAMETER OF ASTEROIDS USING PRINCIPLES OF MACHINE LEARNING

Abhishek Jain[1#], Vishwas Rajaram[2#]
[1]Department of Physics, National Institute of Technology Calicut, Kozhikode-673601
[2]Department of Mechanical Engineering, National Institute of Technology Calicut, Kozhikode-673601

abhishekjain.nitc@gmail.com, vishurajaram8122@gmail.com
Equal authorship

Abstract: We have used NASA's Jet Propulsion Laboratory's Small Object Database to compare three different regression models using an adjusted R2 score method, absolute mean square error and root mean squared error to find the most accurate method to determine the diameter of a small object and asteroids using physical parameters, orbital elements and discovery circumstances. We check the correlation matrix of all the attributes of the dataset to select the correct attributes for modeling. We have used an adjusted R2 score over the naïve R2 score for this particular database as it has a lot of attributes. After comparing, we observe Random forest model gives the best accuracy.

Keywords: - Asteroids, small body, diameter estimation

1. Introduction

Asteroids are objects of varying sizes formed during impact of planetary objects and can be categorized as near-earth objects (NEO), potentially hazardous objects (PHA), Main belt asteroids (MBA), Inner Mail-Belt Asteroids (IMB) and Hyperbolic Asteroids (HYA) (https://ssd.jpl.nasa.gov/sbdb.cgi). The impact of the asteroid upon collision with earth or other nearby planetary objects will depend on its size and hence, the estimation of the size is very important. A review on size estimates of over 1.6 lakh asteroids is available (Mainzer et al 2015) which also gives details on its emissivity properties. The largest sample size for asteroid was made available through the WISE survey which involved usage of thermal flux modeling (Wright et al 2010). An open source for thermal flux modeling (https://github.com/moeyensj/atm) has also been published (Moyens et al 2020). Different methods of estimating diameter of an asteroid have come up in recent times, which included that based on thermal modeling (Masiero et al. 2018), optical data on the basis of correlation between SDSS colour and optical albedo (Izevic 2021) and multilayer perceptron regressor model (Basu 2019). The availability of open data in NASA's Jet Propulsion Laboratory's Small Object and Asteroid Database, which contains more than 800,000 instances of asteroid data (https://ssd.jpl.nasa.gov/sbdb.cgi), allows for trials of different algorithms for increasing the accuracy of prediction of asteroid diameter. In this paper we aim to predict the diameter of asteroid using the principles of Machine Learning (ML) and applied Artificial Intelligence (AI) and find the best fit for our data.

2. Model implementations

A. Why do we need to predict diameters of asteroids:

Predicting asteroid size will be important to define any future collisions and its impact and hence, one needs to evolve newer methods to predict the exact diameter of the asteroids and other orbital parameters. The most widely used method is using the absolute magnitude and albedo values to predict the diameter. The diameter (D) is approximated by:

$$D(H, p_v) = \frac{1329}{\sqrt{p_v}}.10^{-0.2H}[km] \quad (1)$$

where p_v is the geometric albedo and H is the absolute magnitude (Harris and Lagerros 2002).

B. Structure and information in the Dataset

The NASA's Jet Propulsion Laboratory's Small Object and Asteroid Database have the following values:

1. "a": The semi-major axis of trajectory
2. "e": The eccentricity
3. "i": The inclination w.r.t x-y ecliptic plane
4. "y": The time period of an asteroid
5. "H": The magnitude of an asteroid at zero phase angle and at unit heliocentric and geocentric distances.
6. "MOID": Minimum asteroid intersection distance
7. "Albedo": Geometric albedo value, the ratio of body's brightness at zero phase angle to the brightness of a perfectly diffusing disc at the same position and with the same apparent size.
8. "Diameter": The diameter of the asteroid.

Although the database has more than 8,70,000 entries, only about 1,30,000 of these have the diameter of the asteroid specified, which have been used in our analysis. In the absence of certain other parameters in this data set, the missing values have been replaced by the median of that parameter based on the bigger data set. Using this database, we aim to create a robust model to predict the diameters with maximum accuracy.

C. Data Visualization

Data visualization helps us better understand the relation between the parameters and gives us relevant insight (Fig. 1).

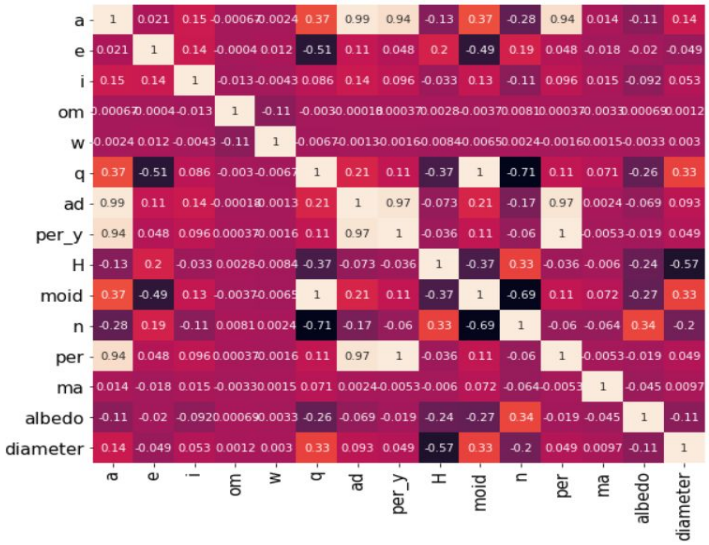

Figure 1: Correlation Matrix for the attributes of database

The correlation between the diameter and absolute magnitude using the database showed an exponential relation [$f(x) = e^{-cx}$] (Fig. 2) correlating well with the earlier defined equation (1).

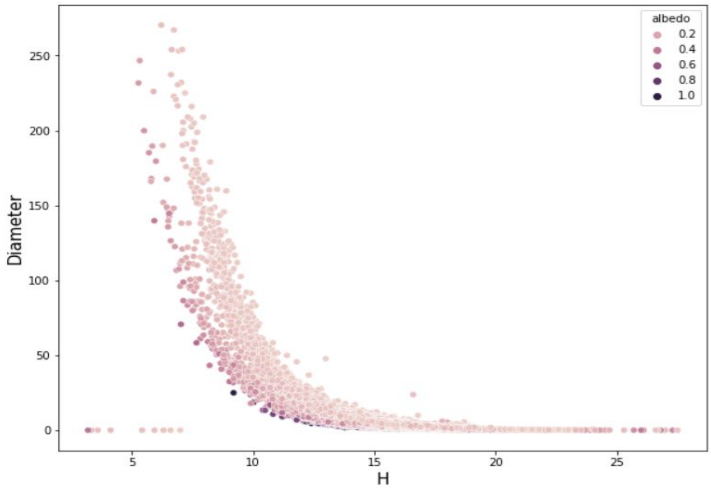

Figure 2: Diameter vs Absolute Magnitude(H):

D. Comparison Metrics:

Different machine learning models have been used to check their performance with the given data and they are compared using an Adjusted R2 score.

An important thing that we need to select is the parameter of comparisons. For this paper, instead of using naïve R2, we use a slightly reprimanded version of the R2 score which we calculated as:

$$R_{adjusted} = 1 - \frac{(1 - R^2) \times (n^{1/3} - 1)}{n^{1/3} - k - 1} \quad (2)$$

Here n is the number of data points and k is the number of independent variables. We used a power of n to scale it to the range of k, so that k isn't redundant. A normal R2 score is calculated as the ratio of Model Sum of squares and Total squares. It can also be the correlation coefficient. A good R2 score shows us how well our prediction fits the data. However, R2 score is bound to be high if the data consists of many independent variables, and hence is not a good metric for comparison for this data. Hence the above mentioned adjusted R2 score is used, which gives us a much better idea about the performance of our Model

E. Algorithms

We have used 3 algorithms to predict the diameter. We tried predicting the albedo and absolute magnitude values and then use their relationship with the diameter of the asteroid. However, this wasn't very useful, due to absence of correlation among the parameters of the asteroid and Albedo and absolute Magnitude (Fig. 1). Hence, we predicted diameter directly, using all other variables as independent using the code generated by us (Github).

E.1: Random Forest Regressor:

Random Forest regressor is an ensemble learning algorithm which uses multiple decision trees. It involves training each decision tree on a different data sample, where sampling is done with replacement. We found that a Random Forest Regressor with depth 10 works best, after which we get a similar R2 score, but with extra computing cost. The first 120,000 data points were used as a method for fitting the model, and the rest were used for testing. The metrics for the algorithm was as follows.:

Mean Absolute Error: 0.0005412433945

Mean Squared Error: 0.000050277417

Root Mean Squared Error: 0.00709065

Adjusted R2 Score: 0.9999835166

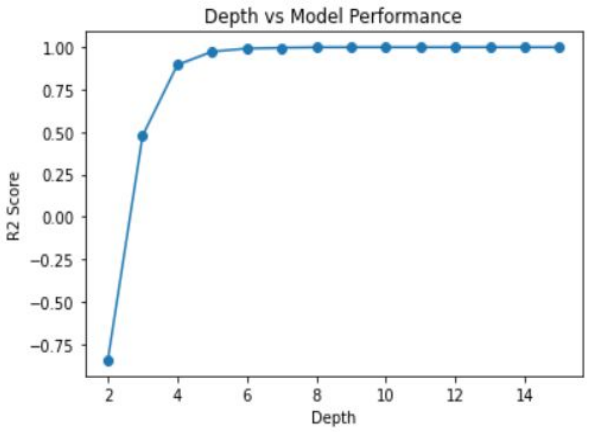

Figure 4: Performance of the model with respect to depth of forest

E.2: XGboost Regressor:

XGB or Extreme gradient boosting is another tree based algorithm that uses multiple additive regression trees, where new models are added to correct the errors made by existing models. It is called so as it uses gradient descent to minimize losses when adding new models. As mentioned earlier, we use the first 120,000 data points for training and the rest for testing. We use a learning rate of 0.05, for which we get the following results:

Mean Absolute Error: 1.0459632872544673

Mean Squared Error: 3.489737096391846

Root Mean Squared Error: 12.178265001933392

Adjusted R2 Score: 0.822

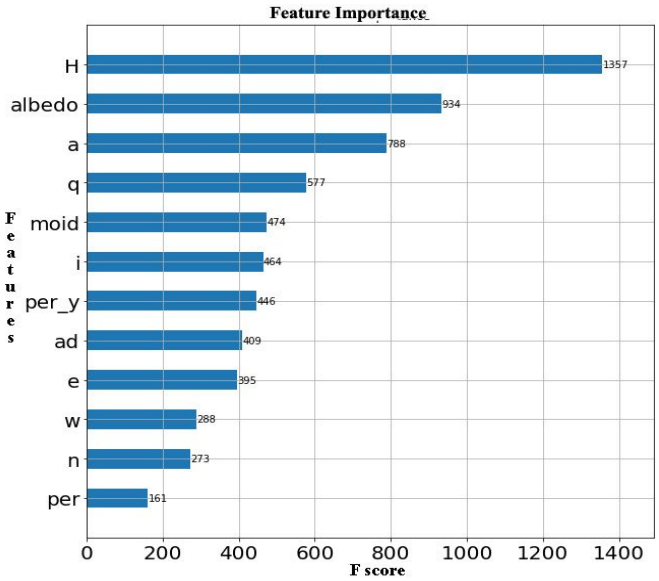

Figure 5: Feature Dependencies

After observing the feature importance table(Fig, 5) we got from XGBoost we can clearly see that Geometric Albedo and Diameter are closely related to the diameter as we had seen earlier in the equation (1).

E.3: Multi-Layer Perceptron Model:

A Multi-Layer Perceptron is not a regressor in true definition, it is a Neural Network. It does not directly use independent variables to predict the dependent variable, but rather uses a network to create dependencies among the variables. Various functions can be used as optimizers, and we found that SGD or Stochastic Gradient Descent is working best, for which the metrics are:

Mean Absolute Error: 0.335367910

Mean Squared Error: 0.5003966881

Root Mean Squared Error: 0.70738722644

Adjusted R2 Score: 0.8524999

F. Conclusion:

Even with various hyperparameter tweaking's, no other algorithm comes even close to a Random Forest Regressor, which gives us an almost perfect fit with very little error. This research shows that Machine Learning can be used for such projects, where non-differentiable non- linearities exist.

G. References

Masiero JR, Redwing E, Mainzer AK, Bauer JM, Cutri RM, Grav T, Kramer E, Nugent CR, Sonnett S, Wright EL. Small and Nearby NEOs Observed by NEOWISE During the First Three Years of Survey: Physical Properties **2018 The Astronomical Journal 156 62.** **https://doi.org/10.3847/1538-3881/aacce4**

Moeyens J, Myhrvold N, Ivezic Z ATM: An open-source tool for asteroid thermal modeling and its application to NEOWISE data **2020 Icarus 341: 113575** https://doi.org/10.1016/j.icarus.2019.113575

Ivezić V, Ivezić Z Predicting the accuracy of asteroid size estimation with data from the Rubin Observatory Legacy Survey of Space and Time **2021 Icarus 357: 114262** https://doi.org/10.1016/j.icarus.2020.114262

Harris, A. W., & Lagerros, J. S. V. 2002, in Asteroids III, ed. W. F. Bottke, Jr.et al. (Tucson, AZ: Univ. Arizona Press), 205

Wright, E.L., Eisenhardt, P.R.M., Mainzer, A.K et al. 2010. The wide-field infrared survey explorer (WISE): Mission description and initial on-orbit performance. Astron. J. 140, 1868–1881. http://dx.doi. org/10.1088/0004-6256/140/6/1868, arXiv:1008.0031.

Mainzer, A., Usui, F., Trilling, D.E., 2015. Space-based thermal infrared studies of asteroids. In: Michel, P., DeMeo, F.E., Bottke, W.F. (Eds.), Asteroids IV. University of Arizona Press, Tucson, AZ, pp. 89–106. http://dx.doi.org/10.2458/azu_uapress_9780816532131-ch005.

Basu V 2019 Prediction of Asteroid Diameter with the Help of Multi-Layer Perceptron Regressor. International Journal of Advances in Electronics and Computer Science. 6: 36-40.

Adjusted R-Squared Formula:
https://www.educba.com/adjusted-r-squared-formula/