1. Introduction
2. Data Description
3. Exploratory Data Analysis
4. Machine Learning Models
5. Performance of the Models
6. Conclusion
7. Further enhancements
8. References
9. INTRODUCTION

Asteroids are giant, irregularly shaped rocky planetoids in space that orbit our Sun. Asteroids range in different sizes. They may hold the secret about the origins of our solar system and the origin of life. Scientists have attributed the moon to be a result of a collision of an asteroid with Earth. Also, the asteroids are made up of rare metals and rich in minerals, so we might want to mine near-earth asteroids for metals to solve the economic crisis on Earth. Finally, if the asteroid is large and near the Earth, it may prove fatal to the planet. They threaten to collide with Earth, and if one of these giant rocks collides with our planet, it can cause a mass extinction of all lives on Earth. Therefore, knowing the size of the asteroid benefits us.

There are over millions of known asteroids in our solar system. Scientists and researchers have measured the diameters of some asteroids in our orbit. But many have not yet been measured. To answer questions such as whether the asteroid is large enough to be mined for metals or if these bodies are close and large enough to be dangerous to the planet, we need to determine the size of the asteroid.

This project aims to develop a machine learning-based prediction system for asteroid diameters. A clean version of Kaggle's JPL asteroid catalog is used as the dataset to train and test the prediction system. As more and more asteroids are discovered and cataloged, our system can be used to estimate their diameters quickly and inexpensively. Armed with this information, scientists and policymakers can help predict the dangers posed by Earth-intercepting asteroids. It also helps us understand how our solar system formed and how objects in space interact.

For this project, we have used the NASA dataset for asteroids and applied several machine learning algorithms for regression to determine and predict the diameter of an asteroid.

2)METHODOLOGY:

2.1) Data collection-

Data was collected from Kaggle. It contains (839736, 27) rows and columns.

2.2)Dataset description

Due to technical, financial and time constraints, we cannot directly measure the diameter of many of these asteroids. NASA's Jet Propulsion Laboratory maintains a database that catalogs known asteroids and stores their properties such as distance from the Sun and eccentricity. So far, scientists have measured the diameters of about 16% of all known asteroids. Asteroid sizes are measured by diameter and can be predicted using these features.

The features included in the dataset are

1. A -> semimajor axis: The semimajor axis of an asteroid is used to describe the dimension of its orbit around the sun, and its value determines the orbital period of small planets.
2. E -> Eccentricity: The axis marked eccentricity measures how far each orbit is from a circle. The smaller the eccentricity number, the more circular the orbit.
3. G -> magnitude of gradient: Its value depends on how the particles on the asteroid's surface scatter light.
4. I -> x-y tilt to ecliptic plane (degrees) Orbital inclination measures the inclination of an object's orbit around a celestial body. It is expressed as the angle between the reference plane and the orbital plane or the orbiting body's directional axis.
5. Om -> Longitude of the ascending node: This is one of the orbital elements used to indicate the trajectory of an object in space. This is the angle from the reference direction, called the origin of longitude, to the movement of the ascending node, measured in the reference plane Link: https://en.wikipedia.org/wiki/Longitude\_of\_the\_ascending\_node
6. W -> argument of perihelion: In astronomy, the periapsis (ω) argument is a way of talking about the orbit of a planet, asteroid, or comet. Also known as the perihelion argument or the near focal argument. It is the angle (starting from the center of the orbit) between an orbiting body's periapsis and its ascending node Link: https://simple.wikipedia.org/wiki/Argument\_of\_periapsis
7. Q -> perihelion distance

The closest distance between an asteroid and the Sun

1. Ad -> aphelion distance

The furthest distance between an asteroid and the Sun

1. Per\_y -> Orbital period
2. The amount of time an asteroid takes to orbit the Sun
3. H -> Absolute Magnitude parameter: An asteroid's absolute magnitude is the visual magnitude an observer would record if the asteroid were placed 1 Astronomical Unit (au) away and one au from the Sun and at a zero phase angle. Link: <https://cneos.jpl.nasa.gov/glossary/h.html>
4. Diameter -> Diameter of asteroid (Km)
5. rot\_per -> Rotation Period(h): The period of time the asteroid takes to rotate on its axis.
6. Albedo -> geometric albedo: Albedo refers to an object's measure of reflectivity or intrinsic brightness. Link: <https://nssdc.gsfc.nasa.gov/planetary/text/asteroids.txt>
7. GM -> Standard gravitational parameter: It is the product of the gravitational constant G and the mass M of the body. <https://en.wikipedia.org/wiki/Standard_gravitational_parameter>
8. Neo -> Near Earth Object
9. Pha -> Physically Hazardous Asteroid
10. Moid -> Earth Minimum orbit Intersection Distance(au): Minimum orbit intersection distance (MOID) is a measure used in astronomy to assess potential close approaches and collision risks between astronomical objects.

2.3)Data preprocessing/Data cleaning:

Data preprocessing is the first step in a machine learning project. We need to convert our raw data into a format that can be successfully interpreted, understood, and modeled by ML algorithms.

2.3.1) Handling missing values

Most models will not work in a case where datasets have missing values. Furthermore, missing data can cause inaccuracies in the model. Therefore handling missing data is a vital part of an ML pipeline. For our project, first, we deleted the columns with missing values of less than 83.7%. After removing these columns, the columns decreased from 27 to 17.

The remaining columns are full\_name a, e, i, om, w, q, ad, per\_y, data\_arc, condition\_code, n\_obs\_used, H, Diameter, neo, pha, moid. As the next step of handling missing data, we deleted rows that had missing values. After deleting these rows, we got 136799 rows out of 839736.

2.3.2)Feature engineering/Data transformation:

For the feature engineering step, we checked the datatype of our dataset's columns.

Table

Description automatically generated As our goal is to predict the Diameter, it should be a float value, and since it is in a string format and has inconsistent entries, we convert it to float datatype for further processing.

Our dataset also contains categorical values; there are 2 columns called neo and pha which has categorical values. We used one hot encoding using pandas to convert it into numerical values. we used the pandas 'get dummies' method to perform this task.

2.3.3) Detecting and handling outliers

for detecting outliers in our data from our original dataset, we created a separate data frame that contains only numerical values so that we can perform computational analysis. We then computed the dataset's IQR, calculated the lower and upper limits, and removed the outliers from the dataset. The filtered dataset is of (106646, 12) rows and columns.

2.3.4) feature selection

Using the segregated numerical values data frame, we created a heatmap to check the correlation between the numerical variables.

Timeline

Description automatically generated with medium confidence

The heatmap shows that a, q, ad, per\_y,moid, per: are strongly correlated. Similarly, H is the only variable that we found was positively correlated with Diameter. However, rot\_per, e, i, w, and om are less correlated with the Diameter.

We then dropped the columns with the low correlation and got a dataframe of shape (136799, 10).

2.3.5) feature scaling

Now, we transform our input data to fit within a

specific scale. We use the standardization method using a class from sklearn. Standardization is a statistically derived process of taking a data set (or distribution) and transforming it so that it is centered around zero and has a standard deviation of 1.

Many ML models, especially those involving regularization, require a proper standardization of input data.

3)EDA and data visualization

4)BUILDING MACHINE LEARNING MODELS

Metrics used:

1. R2 Score: R-squared is a statistical measure of how closely the data fit the fitted regression line. It is also called the coefficient of determination of multiple regression or multiple decision measure.
2. Mean Squared Error: Mean Squared Error (MSE) indicates how close the regression line is to a set of points. To do this, take the distances from the points to the regression line (these distances are the "error") and square them. The squaring is necessary to remove all negative signs. It also makes a big difference as the weight increases. It's called the mean squared error because you're averaging a set of errors. The lower the MSE, the better the prognosis.
3. measurements. The difference between the measured value and the "true" value. Mean Absolute Error (MAE) is the average of all absolute errors. The formula is:
4. Root mean square error: Root mean square error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far the data points are from the regression line. RMSE is a measure of how these residuals are distributed. In other words, it tells you how much the data is clustered around the line of best fit. Mean squared error is commonly used to validate experimental results in climatology, forecasting, and regression analysis.The project used various models to predict asteroid diameters. To do this, I first split the data set into a training set and a test set and got the target value as the diameter.

Using scikit learn we trained the following models.

Linear regression

Polynomial features

Ridge

Lasso

Random forest regressor

XGB regressor

Neural Networks

RESULTS

DISUCISSION

COCNLUSION

References

<https://medium.com/analytics-vidhya/asteroid-diameter-prediction-using-machine-learning-3e81eb14dd7a>

<https://en.wikipedia.org/wiki/Minimum_orbit_intersection_distance>

<https://github.com/gayatripatil29/Asteroid_Diameter_Prediction>

<https://medium.com/analytics-vidhya/asteroid-diameter-prediction-using-machine-learning-3e81eb14dd7a>

https://www.kaggle.com/datasets/basu369victor/prediction-of-asteroid-diameter