Slide 1

According to the NASA website definition, Potentially Hazardous Asteroids (PHAs) are a subset of near-Earth objects (NEOs) that have the potential to make threatening close approaches to Earth. They are generally defined based on two primary criteria: their size and their proximity to Earth's orbit.

PHAs have an absolute magnitude (H) of 22.0 or brighter. This roughly corresponds to a diameter of about 140 meters (460 feet) or larger. Asteroids of this size are considered large enough to cause significant regional damage in the event of an impact.

Proximity: PHAs have orbits that bring them within 0.05 astronomical units (AU) of Earth's orbit. One AU is the average distance between the Earth and the Sun, approximately 149.6 million kilometers (93 million miles). Therefore, 0.05 AU is about 7.5 million kilometers (4.65 million miles).

The link provided will take you to the Small Body database query on the NASA website where you can customize your query for near-Earth objects, the kinds, states, etc. There are a total of 79 variables in this dataset. From our team's query, we ended up having 35,367 data points.

Our prime directive here is to challenge the NASA definition for a Potentially Hazardous Asteroid and discover, amongst all the variables, if the near-Earth objects' absolute magnitude and orbit relative to the Earth are truly the key indicators of PHA classification

---

Slide 2

We have a correlation matrix amongst the datas to initially analyze and understand the relationships between the different features.

We have "sigma_a" the sigma major axis 1 sigma uncertainty values strongly correlating with "sigma_per", which is 1 sigma uncertainty (deg) argument of perihelion. A lot of our correlations reside in the sigma variances, but we will not focus on that. With respect to PHA strong correlation, we see approximately 60% correlation strength with the condition code variable. The condition code variable is the orbit condition code (MPC 'U' parameter) is a measure utilized in astrodynamics and orbital mechanics that quantifies the uncertainty in the orbital parameters of a celestial object. U is the parameter known as the orbit uncertainty parameter that ranges from 0 to 9, where 0 designates a well-determined orbit and 9 designates strong uncertainty and unreliability in an object's orbit. The MPC U parameter is a valuable tool for astronomers and mission planners, providing a quick assessment of the certainty in an object's orbital elements.

I believe we are off to a good start, as having strong confidence in the accuracy of NEO locations indicates a high likelihood of accurate PHA classification.

Slide 3

Different machine learning models may provide various insights to the specific characteristics of the dataset. If we can compare the performance of multiple models, we can determine which model provides the highest accuracy and reliability in PHA classification.

Decision Trees, Random Forests, and Gradient Boosting techniques are effective in handling non-linear relationships and interactions between features. They are capable of handling large datasets that possess complex structures.

The K-Nearest Neighbors (KNN) technique aids in understanding local relationships in data and is a simpler and more intuitive model.

Support Vector Machines techniques are highly utilitarian for high dimensional spaces. They are capable of providing a clear margin of separation for classification.

Linear Regression provides a simplistic baseline model to compare against more complicated models. High effective in identifying strong linear relationships to the target variable.

Long Short Term Memory models have the advantage for utilizing sequential datasets effectively that tend to possess time-dependent features. LSTMs can be used to predict future positions or trajectories of asteroids.

Slide 4

Ensuring that our target variable was "PHA", we conducted analysis utilizing the Random Forest Classification technique. I suspected this technique would be a success, as the Random Forest modeling technique is effective in handling large data structures and non-linear relationships. We received a high R^2 value of 0.987, F1 value of 0.99, and high values for recall and precision. We were also able to capture the top ten most important features in PHA classification. The top contributing feature for PHA classification is the "moid" variable which is the Earth Minimum Orbit Intersection Distance (both in AU and LD). MOID in AU is the smallest distance between the orbit of the asteroid and the orbit of Earth, measured in astronomical units. One AU is the average distance from the Earth to the Sun, approximately 149.6 million kilometers (93 million miles). If the MOID is less than 0.05 AU, the asteroid is considered a Potentially Hazardous Asteroid (PHA). This distance indicates that the asteroid's orbit brings it close enough to Earth to warrant concern, as gravitational perturbations or other factors could alter its path, potentially leading to an impact. MOID in LD is the minimum distance between the asteroid's orbit and Earth's orbit, expressed in lunar distances. One lunar distance is the average distance from the Earth to the Moon, approximately 384,400 kilometers (238,855 miles). Expressing MOID in LD provides an intuitive sense of the asteroid's proximity to Earth in terms of familiar celestial distances. A MOID of less than 19.5 LD (0.05 AU) signifies that the asteroid could come close enough to pose a potential threat.

Slide 5

We also have the absolute magnitude (H) being critical in determining whether an asteroid is potentially hazardous. These two parameters help assess the likelihood of a close approach to Earth and the potential impact risk based on the asteroid's size. Monitoring and analyzing these factors allow scientists to prioritize observational efforts and develop potential mitigation strategies to protect Earth from possible asteroid impacts. So far, this appears 100% in line with NASA's definition of a PHA.

Slide 6

We see that the Decision Tree Classification and Gradient Boosting methods provide strong R^2 scores, as well. One attribute that they do not possess is that they do not capture as many top features for PHA classification as the Random Forest does.

Slide 7

The Long Short Term Memory modeling technique provided a relatively good R^2 value, but not as high as the R^2 values from the previous three models. For our Long-Short-Term Memory, we have a confusion matrix displayed.
True label 0, predicted as 0: 6566 data points;
True label 1, predicted as 0: 49 data points;
True label 0, predicted as 1: 13 data points;
True label 1, predicted as 1: 446 data points.

Slides 8 & 9

I have provided all of the relevant scoring values for the 7 modeling techniques utilized on this data query. The Random Forest Classification and the Gradient Boosting Machines methods performed superior to the others. The Linear Regression Score gave me anxiety.

Slide 10

Referring back to our NASA definition of a PHA, where a PHA determination is based on two primary criteria: their size and their proximity to Earth's orbit. We fact checked this a came up with parallel results, where our top features for PHA determination are the absolute magnitude parameter and the Earth Minimum Orbit Intersection Distance (astronomical units) (moid). Ergo, the data agrees that the MOID, or the minimum distance between the osculating orbits of two objects, and the object's magnitude (or size) are the two primary criteria for PHA classification.