Data 200
Graduate Project Report

Dong Wang 3036361873 dongwangdw@berkeley.edu,
Xiling Long 3036339643 xiling_long@berkeley.edu,
Ruicong Li 3036360300 ruicong.li@berkeley.edu

Abstract

There is an infinite number of asteroids in the universe, and sometimes a few asteroids get too close to Earth, eventually getting into the atmosphere and even hitting the ground. If an asteroid is large enough, it could hurt people or even destroy Earth. Therefore, we would like to come up with a model that could use some information about the asteroid to predict if it is hazardous. Since we decide to treat this as a classification problem, we try to use Logistic Regression, Decision Tree, and Random Forest to solve it. Before fitting our model, we did some necessary data cleaning and feature selection. As a result, our Decision Tree and Random Forest models are excellent at predicting whether an asteroid is hazardous to Earth.

Introduction

Near-Earth Objects (NEOs) are comets and asteroids that have been nudged by the gravitational attraction of nearby planets into orbits that allow them to enter the Earth's neighborhood. To evaluate whether a given asteroid might have an impact on Earth, NASA conducts a lot of computation and prediction of how close that orbit comes to Earth.

The dataset consists of information about Asteroids like their Name, IDs, relative velocities, Estimated diameter inclination, etc, and a label of whether it's hazardous or not from NASA. Current calculation and assessment rely on a traditional way and linear methods, based on the collected data, we applied multiple machine learning models to predict whether an asteroid could be potentially hazardous and look for important features that are responsible for an asteroid's impact.

Description of the data

This dataset contains 4687 instances and 40 attributes. Some of them measure the same feature of the asteroid but by different standards, like Est Dia, Close Approach Date, Epoch Date Close Approach, Relative Velocity, and Miss Dist.

Most of the features are numerical, and only 5 of them are non-numerical attributes. Hazardous is a boolean, Close Approach Date, Orbiting Body, Orbit Determination Date, Equinox area type object. Besides, there are no missing values in the dataset.

Features	Descriptions
Neo Reference ID	Near-Earth Object (NEO) reference ID number for an asteroid (or a comet) which comes close to earth
Name	'Name' of asteroid (same as NEO Reference ID)
Absolute Magnitude	A measure of the asteroid's luminosity (in H) (the brightness of an asteroid if it is 1 astronomical unit away from both the Sun and the observer, and the angle between the Sun, asteroid, and Earth is 0 degrees)
Est Dia in (in KM, M, Miles, and Feet) (min)	Minimum estimated diameter of the asteroid (Note: Since asteroids cannot be directly measured and because they have irregular shapes, their diameters are estimates. These estimates are calculated using its absolute magnitude and geometric albedo.)
Est Dia in (in KM, M, Miles, and Feet) (max)	Maximum estimated diameter of the asteroid
Close Approach Date	Date at which the asteroid approaches close to Earth
Epoch Date Close Approach	Date at which the asteroid approaches close to Earth (in epoch time)
Relative Velocity (in km per sec, km per hr, and miles per hour)	Asteroid's velocity relative to earth
Miss Dist.(in Astronomical, lunar, km, and miles)	Distance by which the asteroid misses Earth
Orbiting Body	currently 'Earth’
Orbit ID	An ID of JPL NEA orbit that JPL Nasa uses in its analysis
Orbit Determination Date	Date at which the asteroid's orbit was determined

Orbit Uncertainity	A measure of the uncertainity ('measurement errors') in the calculated orbit
Minimum Orbit Intersection	The closest distance between Earth and the asteroid in their respective orbits (in astronomical units)
Jupiter Tisserand Invariant	A value used to differentiate between asteroids and Jupiter
Epoch Osculation	The instance of time at which the asteroid's position and velocity vectors (from which its osculating orbit is calculated) is specified
Eccentricity	A value which specifies by how much the asteroid's orbit deviates from a perfect circle
Semi Major Axis	The longest radius of an elliptical orbit; a easure of the asteroid's average distance from the Sun (asteroids orbit the Sun)
Inclination	Measures the tilt of the asteroid's orbit around the Sun
Asc Node Longitude	(copying from NASA) 'Angle in the ecliptic plane between the inertial
Orbital Period	Time taken for asteroid to complete a single orbit around the Sun
Perihelion Distance	Distance of point in asteroid's orbit which is closest to the Sun
Perihelion Arg	(copying from Nasa) 'The angle (in the body's orbit plane) between the ascending node line and perihelion measured in the direction of the body's orbit'
Aphelion Dist	Distance of point in asteroid's orbit which is farthest from the Sun
Perihelion Time	Length of time of asteroid's passage through the perihelion stage
Mean Anomaly	(copying from Nasa) 'The product of an orbiting body's mean motion and time past perihelion passage'
Mean Motion	(copying from Nasa) 'The angular speed required for a body to make one orbit around an ideal ellipse with a specific semi
Equinox	An astronomical standard to measure against (currently 'J2000.0')
Hazardous	Is the asteroid hazardous? (True or False)

Table 1 Feature table

Description of methods

1. Logistic regression

The Logistic Regression is a model where an input data point is received, and a probability will be produced. Logistic Regression is commonly used in classification problems. The model has a set of parameters, corresponding to the intercept and each feature, and it uses a sigmoid

function to convert the product of coefficients and inputs to a probability. The probability represents how likely is one input data point to be classified as a certain category. Additionally, the model uses cross-entropy loss to converge to a global minimum. Here is a picture of the Logistic Regression model compared to the Linear Regression model.

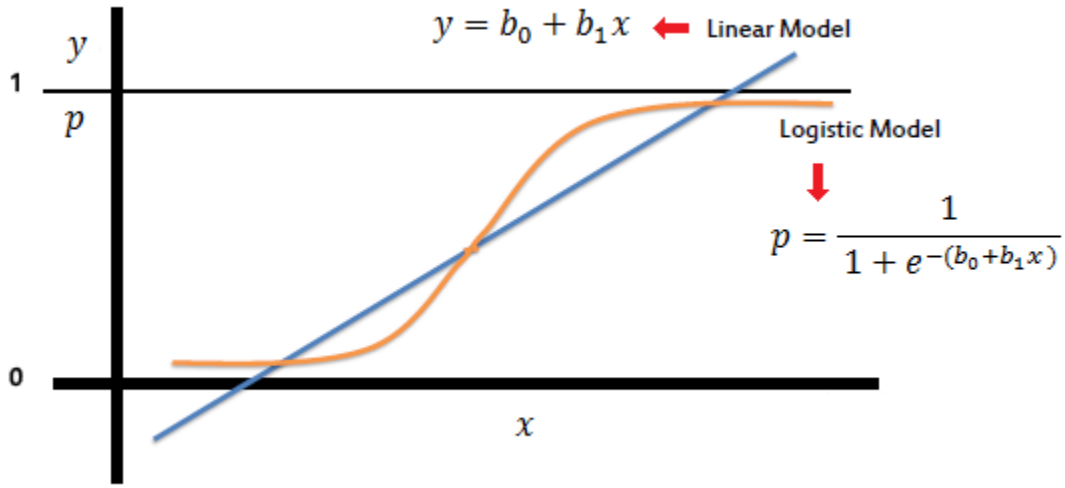

Figure 1 Logistic Regression Model vs Linear Regression Model

source:https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.saedsayad.com%2Flogistic_regression.htm&psig=AOvVaw2D7t3AoO65BsqLvtSdYYgh&ust=1650582214341000&source=images&cd=vfe&ved=0CAwQjRxqFwoTCJCo-43go_cCFQAAAAAdAAAAABAD

2. Decision Tree

Decision Tree Classification is an alternative way to classify data. The reason it is called Tree Classification is that it builds classification and regression models in a form of the tree structure. The algorithm behind the decision tree classification is splitting the dataset into smaller and smaller subsets. At the same time, a sub-decision tree is incrementally developed and finally becomes a tree with decision nodes and leaf nodes.

Figure 2 Visualization of a Decision Tree Model Making a Prediction

source:https://www.saedsayad.com/decision_tree.htm#:~:text=Decision%20tree%20builds%20classification%20or,decision%20nodes%20and%20leaf%20nodes.

3. Random Forest

The Random Forest is an ensemble learning algorithm that combined many single decision trees together and uses the vote to decide the final results. Each individual tree in the random forest generates a classification outcome and the class with the most votes becomes the model's prediction. When training the single trees, the model will select a random subset of features and use a random subset of data, which makes each tree less correlated with others. The low correlation benefits the model, as uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions.

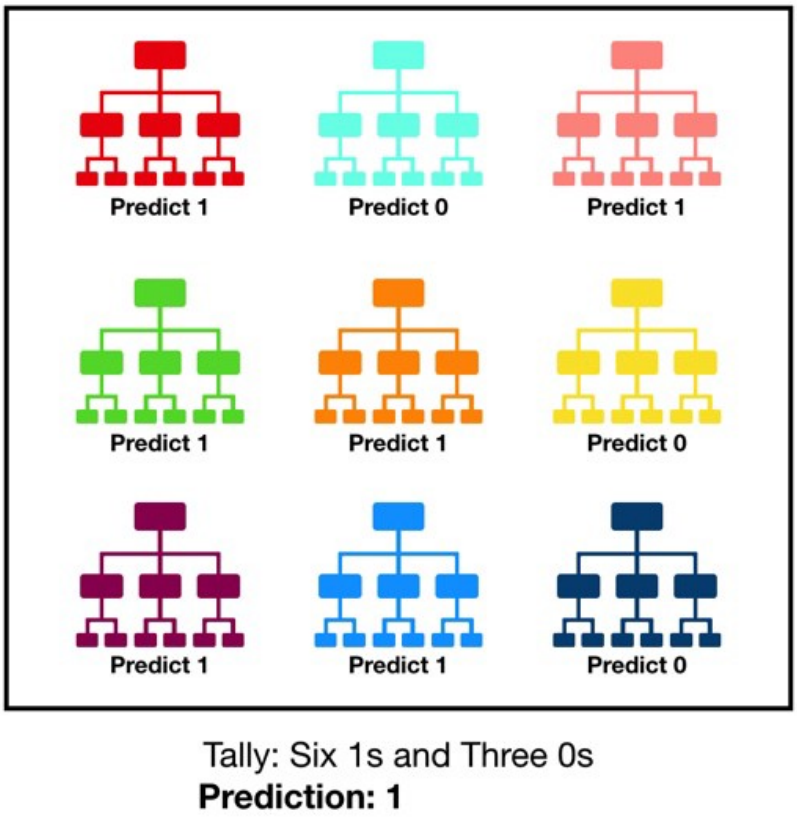

Figure 3 Visualization of a Random Forest Model Making a Prediction

source:https://towardsdatascience.com/understanding-random-forest-58381e0602d2#:~:text=The%20random%20forest%20is%20a,that%20of%20any%20individual%20tree.

Summary of results

	Logistic Regression	Decision Tree	Random Forest
Accuracy(Train)	0.950	1.000	1.000
Accuracy(Test)	0.946	0.996	0.998
Precision(Train)	0.868	0.991	0.991
Precision(Test)	0.841	0.991	0.991

Recall(Train)	0.846	0.986	0.995
Recall(Test)	0.837	0.986	0.995

Table 2 Model Metrics

Above is the metrics of our models. As we can see, Random Forest is the best model among these three, attaining a precision of 99.1% and recall of 99.5% on the test set.

The most important features in predicting the status of an asteroid are Minimum Orbit Intersection, Est Dia in (min), and Absolute Magnitude. We found it interesting that Absolute Magnitude is quite important as it measures the luminosity of an asteroid. According to CNEOS' explanation, brightness is a reflection of the distance between the asteroid and earth, which makes it plausible as the closer the distance, the riskier it will be.

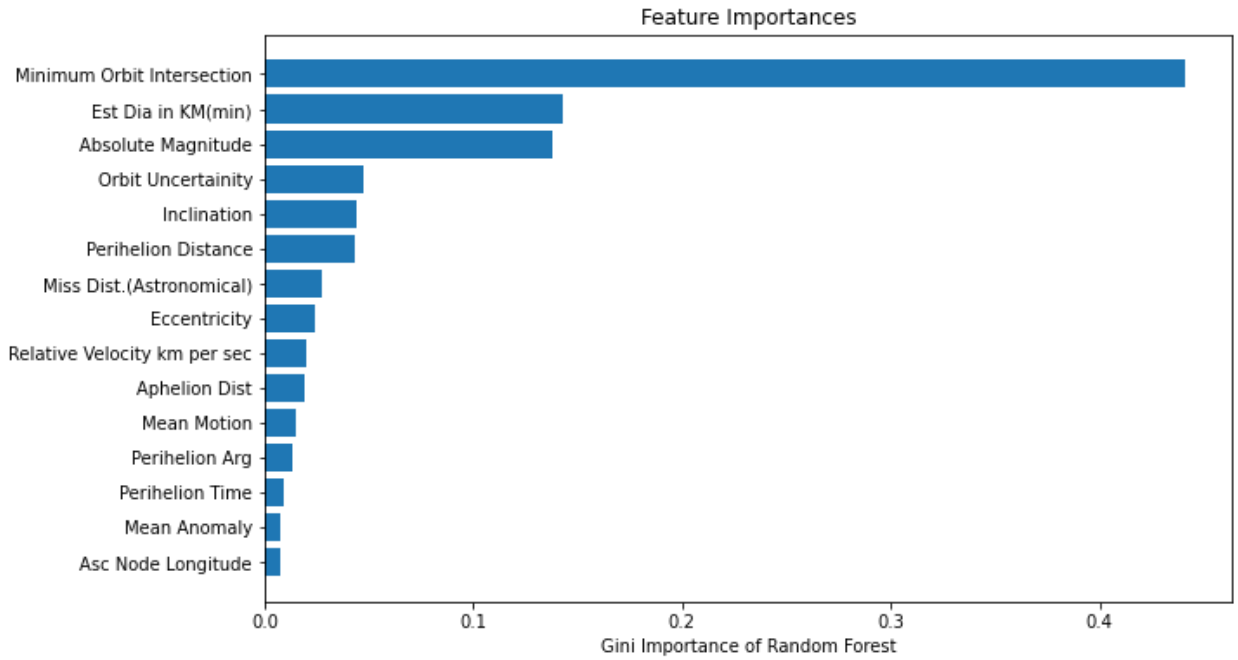

Figure 4 Feature Importances of Random Forest Model

Mean Motion and Aphelion Dist have a highly negative relationship. Mean Motion and Jupiter Tisserand Invariant has a highly positive relationship. Aphelion Dist and Eccentricity are also positively related.

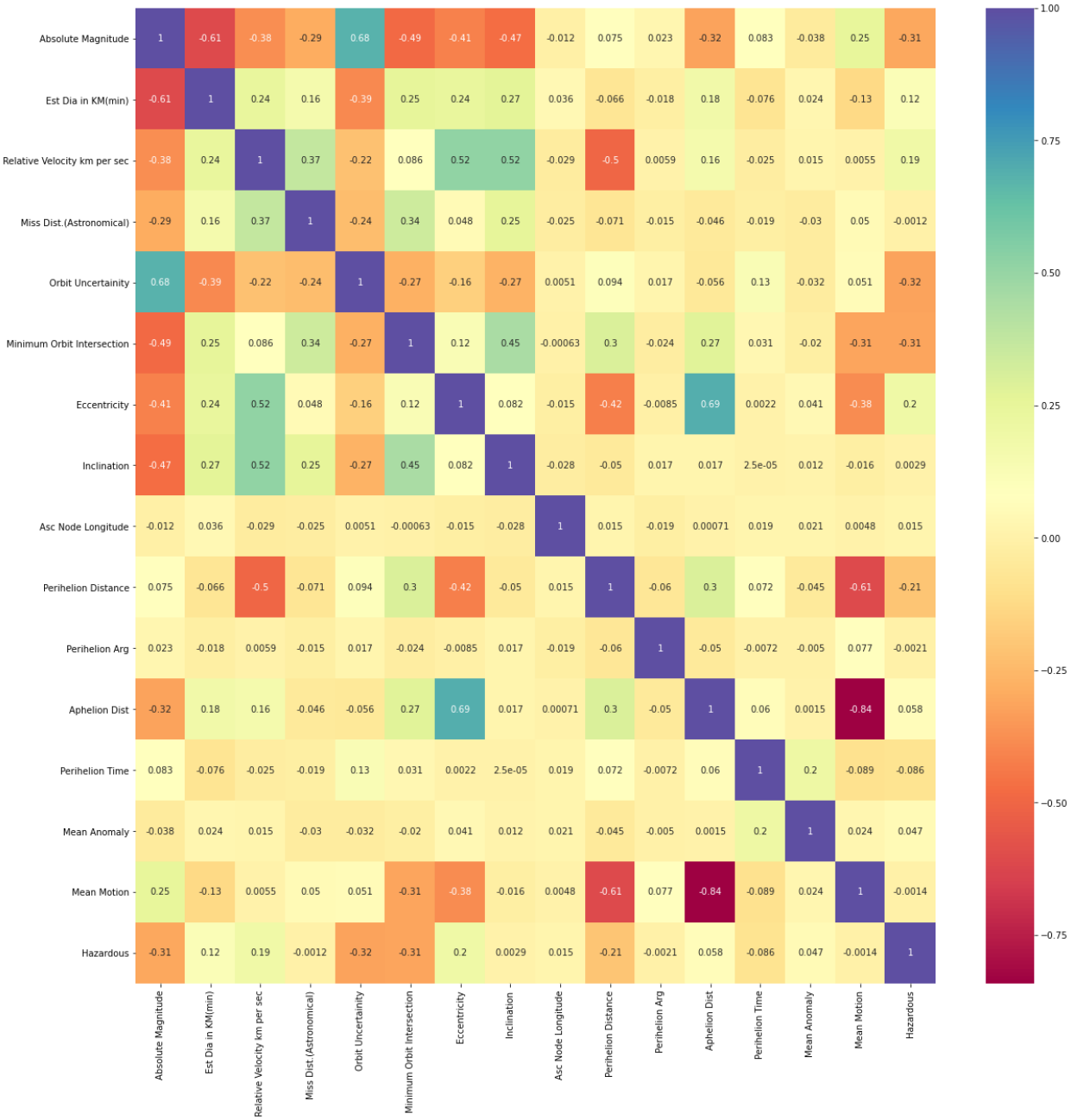

Figure 5 collinearity heatmap after processing data

There are some features like Est Dia in KM(min), Minimum Orbit Intersection and Aphelion Dist has right-skewed distribution. Features like Absolute Magnitude, Perhelion Distance and Eccentricity are pretty close to normal distribution. Besides, Perhelion Time has a left-skewed distribution.

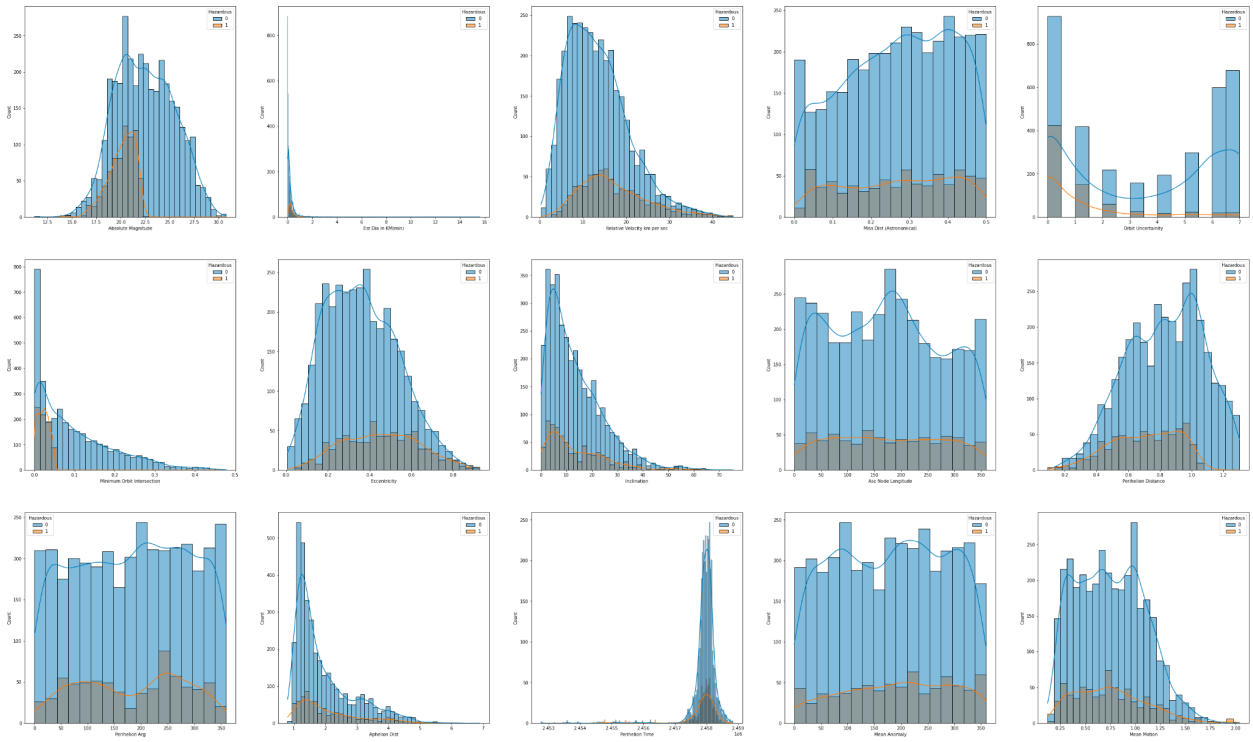

Figure 6 Histogram of each feature

Discussion

1. Dataset bias

The dataset we are using only contains 4600+ rows, which are sampled in advance. Due to the confidentiality of the data of the asteroid, we are unable to access the unsampled and unprocessed data. If there are inappropriate methods applied while collecting original data, it

can generate bias in the dataset. As a result, we might build these models based on an imperfect dataset.

2. Dataset processing

If we take a look at the Histogram distribution of each feature, we can find that there are several figures that have long tails and outliers, which impacts the distribution of the data. Ideally, we need to remove the outliers and center the data to get a better dataset, so that the prediction results will be more accurate.