NY-P3-background

Data Choice

[NASA Hazardous Asteroid Classification dataset from Kaggle (https://www.kaggle.com/shrutimehta/nasa-asteroids-classification)](https://www.kaggle.com/shrutimehta/nasa-asteroids-classification)

Identify features

- The first two columns contain identical identifier values and are not too important or beneficial for a model.
- The next feature is the absolute magnitude which looks looks at the brightness of an celestial object, according to the definition of absolute magnitude, as it would be seen at a distance of 10 parsecs (equal to 1.9174E+14 miles).
- The next set of features are related to the diameter of asteroids. Estimates are made in kilometers (km), meters (m), miles (mi), and feet (ft), with data for the maximum and minimum of each distance.
- There are two columns addressing the date asteroids will approach earth, by date and periods (epoch).
- Features also include the speed of the asteroid, the distance from the earth the asteroid will pass, measured in astronomical, lunar, km, and mi units.
- There are a number of columns dealing with the orbit pattern including the orbital period, perihelion distance, aphelion distance, eccentricity and the like.

Identify target

The target variable is `hazardous` column, showing whether the asteroid is hazardous or not, based on size, speed, and orbit.

Describe processing

The `hazardous` column contains `True` / `False` values so this column would need to encoded in order to properly predict if an asteroid is hazardous or not. I would use pandas `get_dummies()` to convert values to 0's and 1's respectively.

Classification/Regression

This would be a classification project. The goal is to determine whether an asteroid is hazardous or not based on certain features of said asteroid.

Metric

The metric I would use is Mean Squared Error as it takes the average of the errors. Getting this value to the smallest value shows an accurate model.

Estimate of Bayes Optimum

Using Bayes Optimum, I would predict the MSE sould be low and the Bayes Optimum score will be high. According to other models using this data, they scored very well. There are a number of features to choose from and a number of True/False values to predict with.