Predicting the Diameter Of Asteroids

Emily Bocim

Springboard Data Science Capstone Project

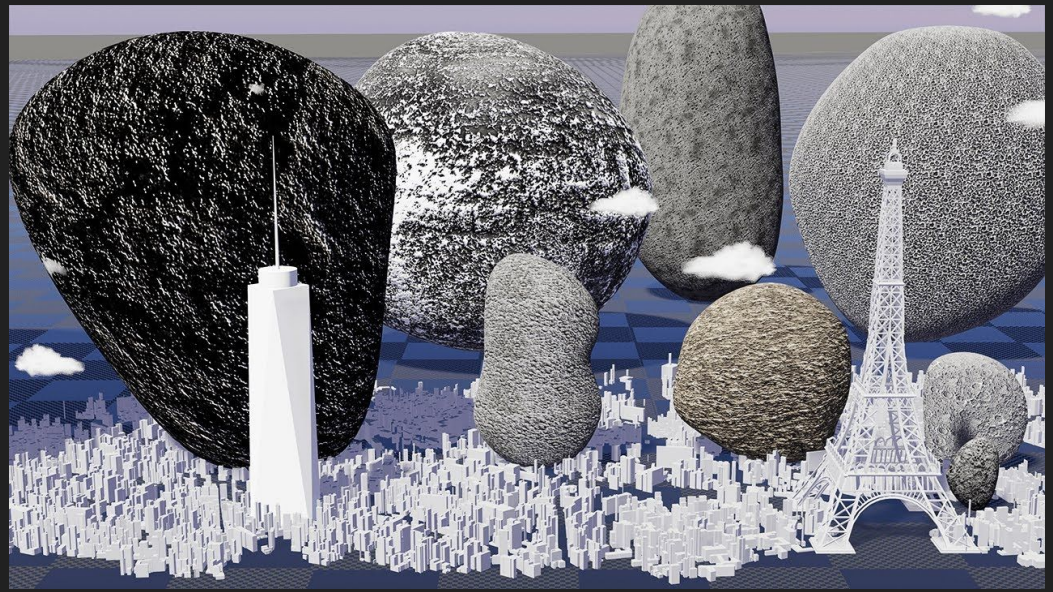
May 2020

Thanks to Springboard Mentor

Branko Kovac

Project Statement

The purpose of this project was to determine a method for predicting the diameter of asteroids.



Who Cares?

- Asteroids range from a height equivalent to the average human to more than a quarter the size of the moon and orbit in all different areas of the solar system.

- NASA and other space organizations are interested in their size for modeling the effect of their impact on Earth and other bodies in the solar system.

- There are also talks about traveling to certain asteroids for mining purposes or (not just for the movies) for altering their trajectory.

Data Information

- The database used for this project was provided by the Jet Propulsion Laboratory of California Institute of Technology database, an organization under NASA. It can be accessed at:
  - https://ssd.jpl.nasa.gov/sbdb_query.cgi

- The version of the data was supplied by Victor Basu on Kaggle at:
  - https://www.kaggle.com/basu369victor/prediction-of-asteroid-diameter
  - This dataset was generated in 2019.

What are some indicators of asteroid size?

There were initially 31 features available in the dataset, but these features can be summarised as follows:

- Where the asteroid orbits.

- What the asteroid’s orbit looks like.

- Measurements of different types of light.

Data Cleaning

- When initially working with the data, it was found that there was too much missing data from any of the features involving light.

- These features were removed from any further exploration.

- Only features related to the number of observations and the orbit of the asteroid were able to be kept for further exploration and eventually the modeling process.

Data Exploration

General visualization of an object's orbit in space.

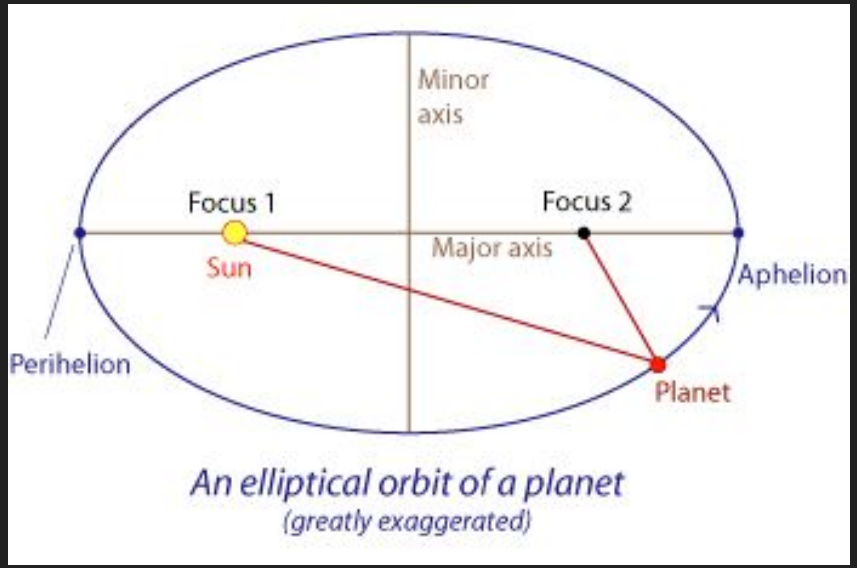
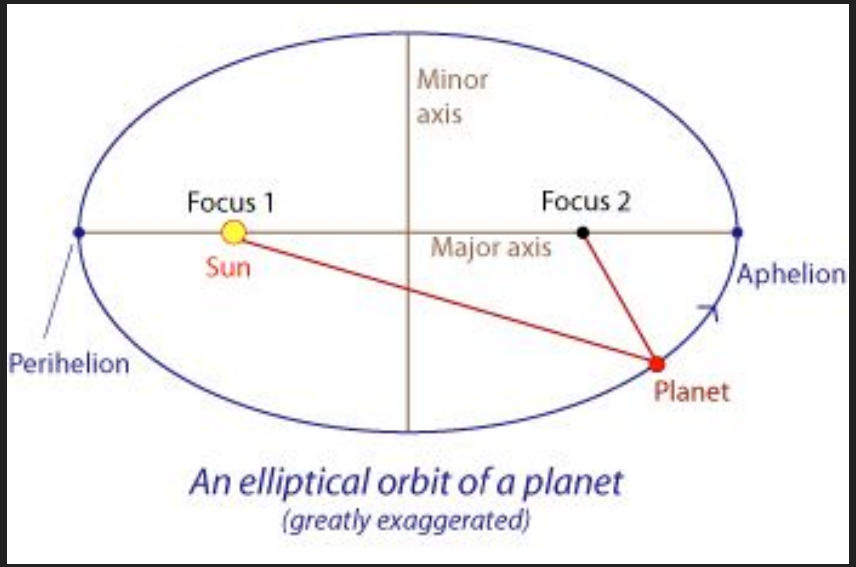


Image Reference: http://astronomer-wpengine.netdna-ssl.com/wp-content/uploads/2013/06/kepler1.gif

Data Exploration: Semi-Major Axis

- The long radius of an object's elliptical orbit.
- Obvious relationships with aphelion and perihelion distances, with each being near half the distance of the major axis depending on other orbital features.
- Relates to orbital period as well since orbit size is a factor, as well how fast it is orbiting.



Data Exploration: Orbit Class

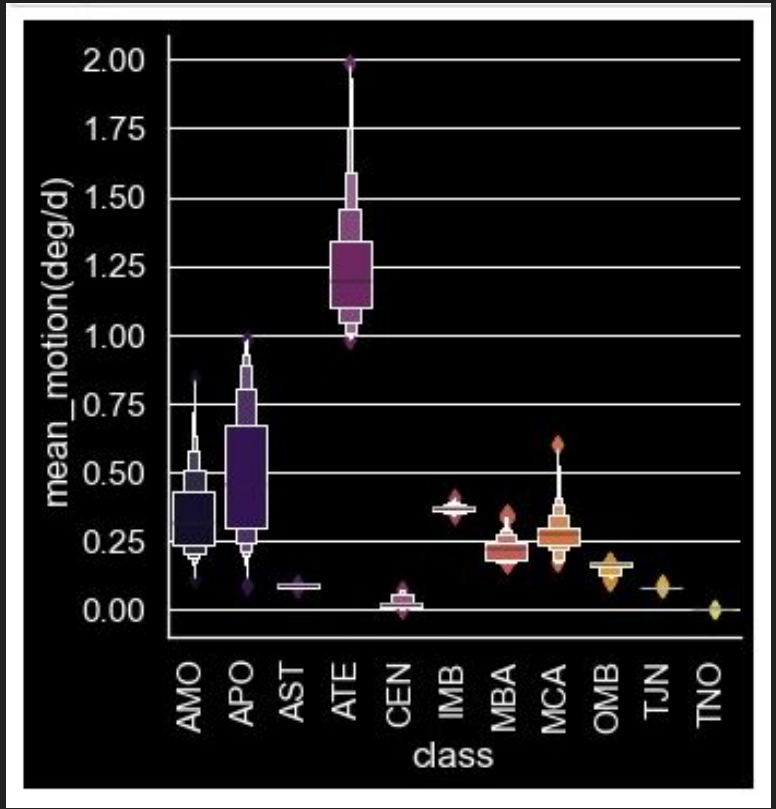
This is where an asteroid orbits and is used to generalize to generalize an asteroids distance from the sun when exploring other features.

Asteroid Orbit Classes

Abbreviation	Title	Description
AMO	Amor	Near-Earth asteroid orbits similar to that of 1221 Amor (a > 1.0 AU; 1.017 AU < q < 1.3 AU).
APO	Apollo	Near-Earth asteroid orbits which cross the Earth's orbit similar to that of 1862 Apollo (a > 1.0 AU; q < 1.017 AU).
AST	Asteroid	Asteroid orbit not matching any defined orbit class.
ATE	Aten	Near-Earth asteroid orbits similar to that of 2062 Aten (a < 1.0 AU; Q > 0.983 AU).
CEN	Centaur	Objects with orbits between Jupiter and Neptune (5.5 AU < a < 30.1 AU).
HYA	Hyperbolic Asteroid	Asteroids on hyperbolic orbits (e > 1.0).
IEO	Interior Earth Object	An asteroid orbit contained entirely within the orbit of the Earth (Q < 0.983 AU).
IMB	Inner Main-belt Asteroid	Asteroids with orbital elements constrained by (a < 2.0 AU; q > 1.666 AU).
MBA	Main-belt Asteroid	Asteroids with orbital elements constrained by (2.0 AU < a < 3.2 AU; q > 1.666 AU).
MCA	Mars-crossing Asteroid	Asteroids that cross the orbit of Mars constrained by (1.3 AU < q < 1.666 AU; a < 3.2 AU).
OMB	Outer Main-belt Asteroid	Asteroids with orbital elements constrained by (3.2 AU < a < 4.6 AU).
PAA	Parabolic Asteroid	Asteroids on parabolic orbits (e = 1.0).
TJN	Jupiter Trojan	Asteroids traped in Jupiter's L4/L5 Lagrange points (4.6 AU < a < 5.5 AU; e < 0.3).
TNO	TransNeptunian Object	Objects with orbits outside Neptune (a > 30.1 AU).

Data Exploration: Mean Motion

- The angular speed required for the asteroid to complete one orbit.
- The figure shows the relationship between mean motion and orbit class.
- As is relates to the type of orbit, the relationship is visually similar between the other features .



Data Exploration: Other Features Defined

Perihelion Distance: The closest distance to the sun of an orbit.

Aphelion Distance: The farthest distance from the sun of an orbit.

Eccentricity: How elliptical an orbit is.

Mean Anomoly: The product of mean motion and past perihelion passage.

X-Y Inclination: The tilt of the object’s orbit.

Regression Modeling

After initial models, a non-linear relationship was observed. By using the log of diameter as the target, model performance was increased.

Models used and their best performance:

Model	R-squared for Test Set
Linear Regression	0.716
Ridge Regression	0.714
Lasso Regression	0.658
Random Forest Regression	0.849
Gradient Boosting Regression	0.832

Best Model

The best performing model was the Random Forest Regressor using:

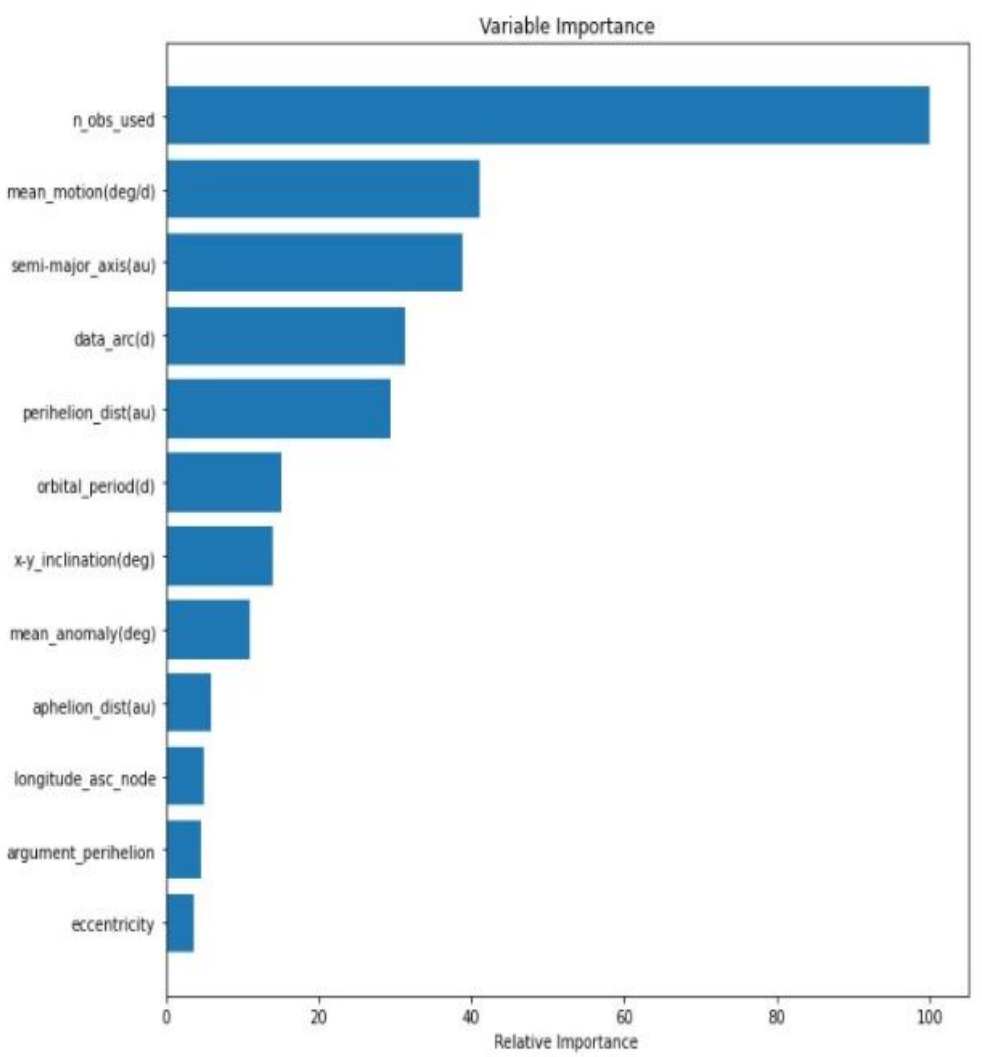
n_estimators = 100

Performance:

- Train Score: 0.97
- Test Score: 0.83
- MSE: 0.07
- RMSE: 0.26

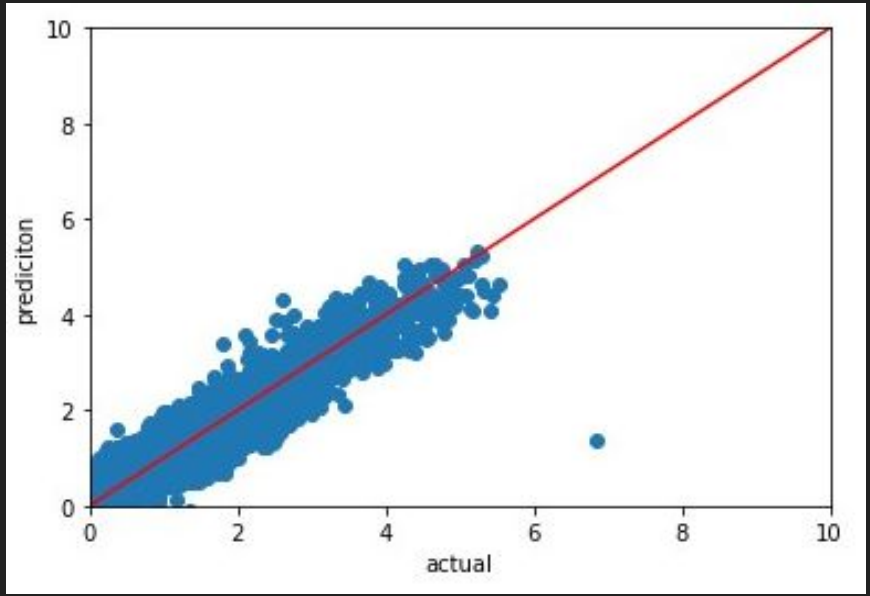
The figure shows the features ranked by

importance for the model.



Random Forest Regressor:

Predicted diameters plotted against actual diameters for the test set.



Assumptions and Limitations

The data as a whole:

- It is assumed that the data is the most accurate available.
- Accuracy of the data is limited by technology and precision of calculations.

Grouping of the Data:

- The majority of the data is from observations in the main asteroid belt.
- Diameters outside beyond the asteroid belt are less know, so the model is prone to overfitting in those regions until more data is available.

Future Work

Infrared Light:

- Infrared light is used by NASA for more accurate diameter measurements. Once enough data is available, this features associated to infrared light could improve model performance.

Asteroid Orbit Class:

- Initial exploration showed potential for creating models that performed better based on asteroid location. This was only done by three sections.
- Further exploration could include breaking it down to each orbit class with the use of different types of models.

Conclusions

- Random forest regressor was the best performing model using all remaining features, an n_estimators value of 100, and a target as the log of diameter.

- As technology and data improves, there will be even more potential for increasing the accuracy of future models.

- Further exploration can be done into using different models based on the location of the asteroids.

Thank you!

Emily Bocim

Email:

emily.bocim@gmail.com

Linkedin:

https://www.linkedin.com/in/emily-bocim/

Project Repository:

https://github.com/emilybocim/springboard/tree/main/Capstone%20Two