Springboard

Data Science Career Track

Capstone Project 2

Predicting the Diameter of Asteroids

Emily Bocim

May 2020

Table of Contents

Asteroids
Emily Bocim

Introduction

Asteroid data is tracked for determining the amount of risk an asteroid poses to Earth. It can be difficult to determine the actual size through visual methods, due to the amount of reflected light. Determining the amount of light reflected can require variables that are unknown, such as the compounds that make up the asteroid, which makes other methods of finding diameter more feasible than others. Asteroid diameter is important in risk analysis for determining severity of impact on Earth and space missions.

Client

The use of predicting asteroid diameter has a very targeted audience. This includes organizations with the intent of sending various objects to space and those studying space objects. Being able to know how large an asteroid is helps to predict the severity of impact, whether our planet or another. In regions with smaller asteroids, it may determine how much an impact a machine needs to be able to take.

Database

- 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.
- There are features that can be pulled from the database that were not included in the dataset, that could potentially be used to increase the accuracy of the model. Further exploration into how they relate to diameter would be required before introducing into the modeling process.

Asteroids
Emily Bocim

Data Wrangling

Data Summary:

Before working with the dataset, it consisted of 31 features and 839,714 observations. After the cleaning process, this decreased to 19 features and 137,636 observations.

Data Type Inconsistencies:

- **Diameter (Target Feature) -** There were added non-numeric characters which caused diameter to be recognized as an object type. These were removed and then it was able to be converted to numeric.

Missing Data:

- **Diameter -** Initially, ~84% of the observations were missing a value for the diameter. As this is the target, those observations were removed. Even with removing these observations, there were still 12 features with more than 50% of their values missing. These features were dropped from the dataset as well.
- **Categorical Features -** Features of this type, that remained until model building, had no issues.
- **Numerical Features**
    - **Data Arc -** This was the only other feature to have problems with missing values. As figure 1 shows, it consisted of many outliers, so median was chosen to fill in these values.

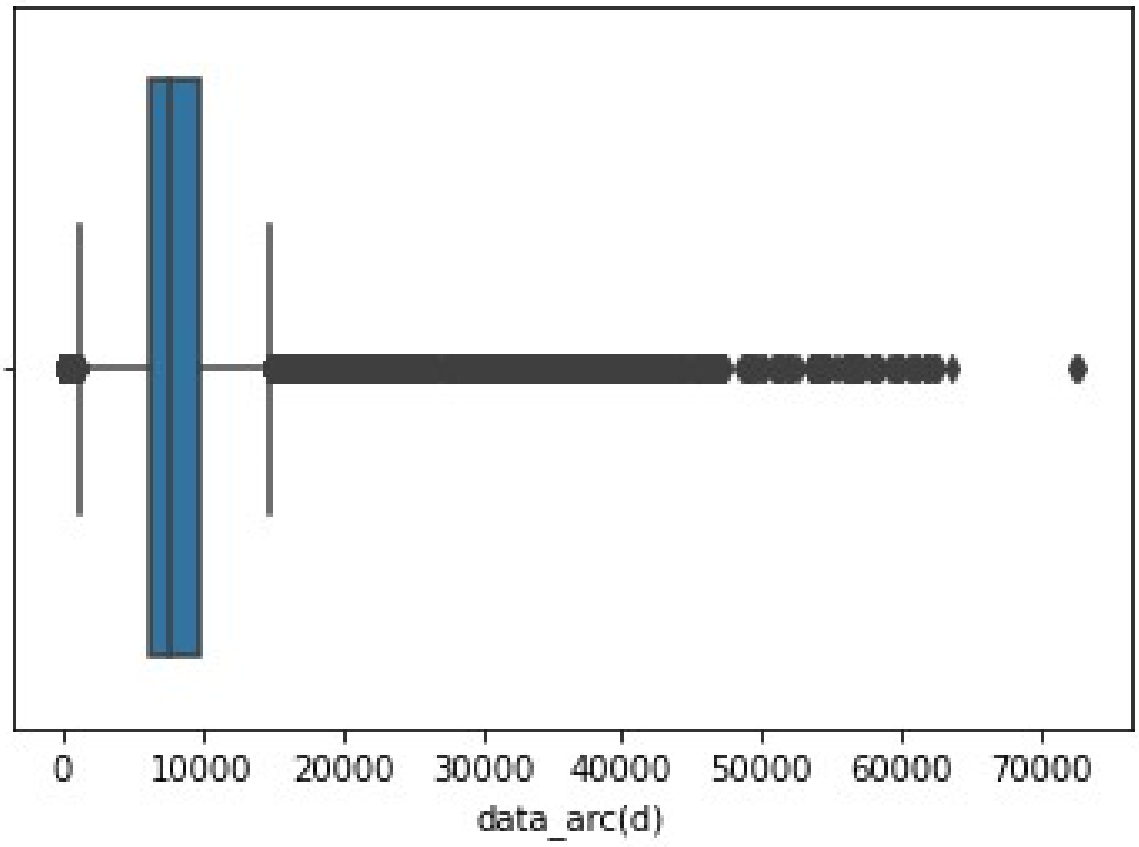

Figure 1: Boxplot of the Data Arc feature.

Data Exploration

Calculated Features:

Some features were directly calculated from other features, so were removed to avoid redundancy or reliance on the target. There were also duplicate features that were removed.

- Near Earth Object Classifier - Based on orbit class.
- Physically Hazardous Classifier - Determined based on diameter.

Orbit Class

- What is it?
    - Where an asteroid orbits, such as if is part of the main asteroid belt, orbits a larger, or is near Earth.
- What defines the different orbit classes? See Figure 2.
- There was an apparent relationship between different orbit classes and diameter that is represented in Figure 3. Orbits are determined by strength of gravity, which is determined by the size of the two objects and how close they are to each other. So, it makes sense that larger asteroids are going to be closer to the larger planets, than to Earth. In future exploration, it is found that perihelion distance is a more precise feature for describing this, but orbit class was useful in initial visualizations.

Asteroids
Emily Bocim

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).

Figure 2: Description of Asteroid Orbit Classes. (University of Maryland, 2020)

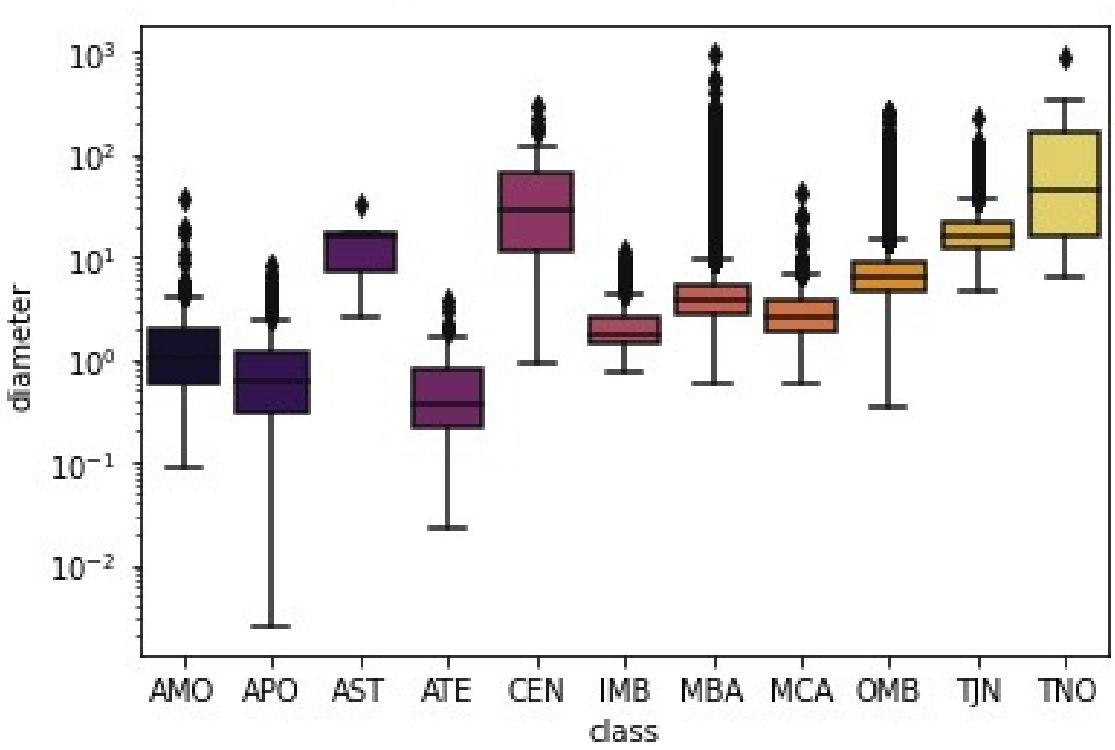

Figure 3: Boxplot of asteroid diameter by orbit class.

Semi-Major Axis

- What is it?
    - The longest radius of an object's elliptical orbit.
- Relationship to other features:
    - There was an almost linear relationship with Aphelion Distance, with Aphelion Distance being around half of the Semi-Major Axis. This is expected as Aphelion distance is the farthest point of the orbit.
    - In relation to Orbital Period, to appeared to be a non-linear relationship. A pattern between the two is not surprising as they both relate to how the asteroid orbits the sun.

Asteroids
Emily Bocim

Mean Motion

- What is it?
    - The angular speed required the asteroid to complete one orbit.
- Relationship to other features:
    - There do appear to be some groupings between Mean Motion and Orbit Class. See Figure 4. However, there is still enough overlap between some classes that one feature cannot be used to fully describe the other. A similar relationship was also seen when comparing Orbit Class with Argument Perihelion, Data Arc, and the Number of Observations.

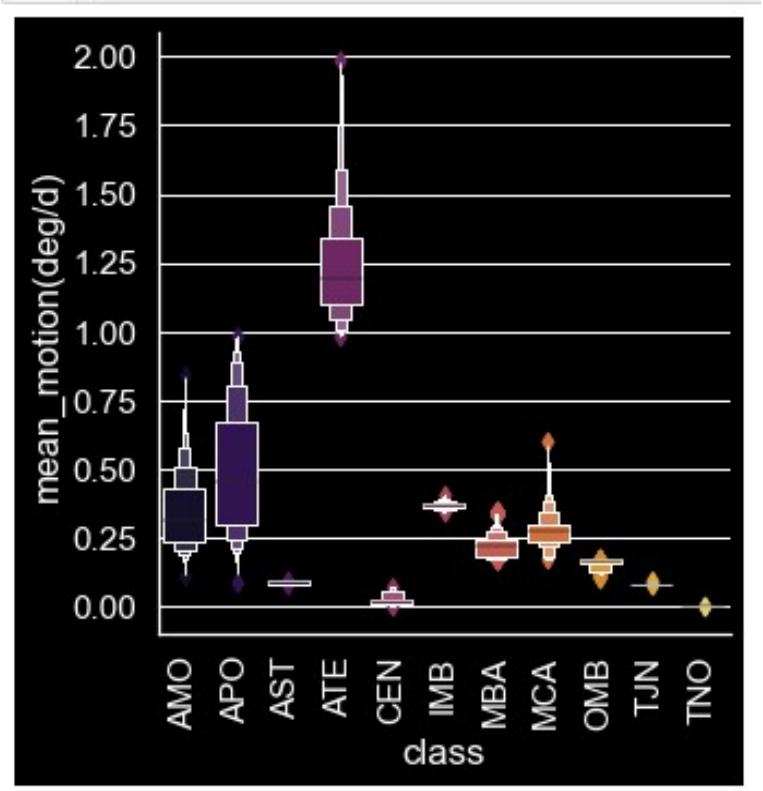

Figure 4: Boxplot of Orbit Class versus Mean Motion.

Perihelion Distance

- The closest distance to the sun of an orbit.
- Figure 5 shows the relationship between Perihelion Distance and Orbit Class. Many of these features define various aspects about an asteroids orbit, where the Orbit Class is more a summary of all such features.

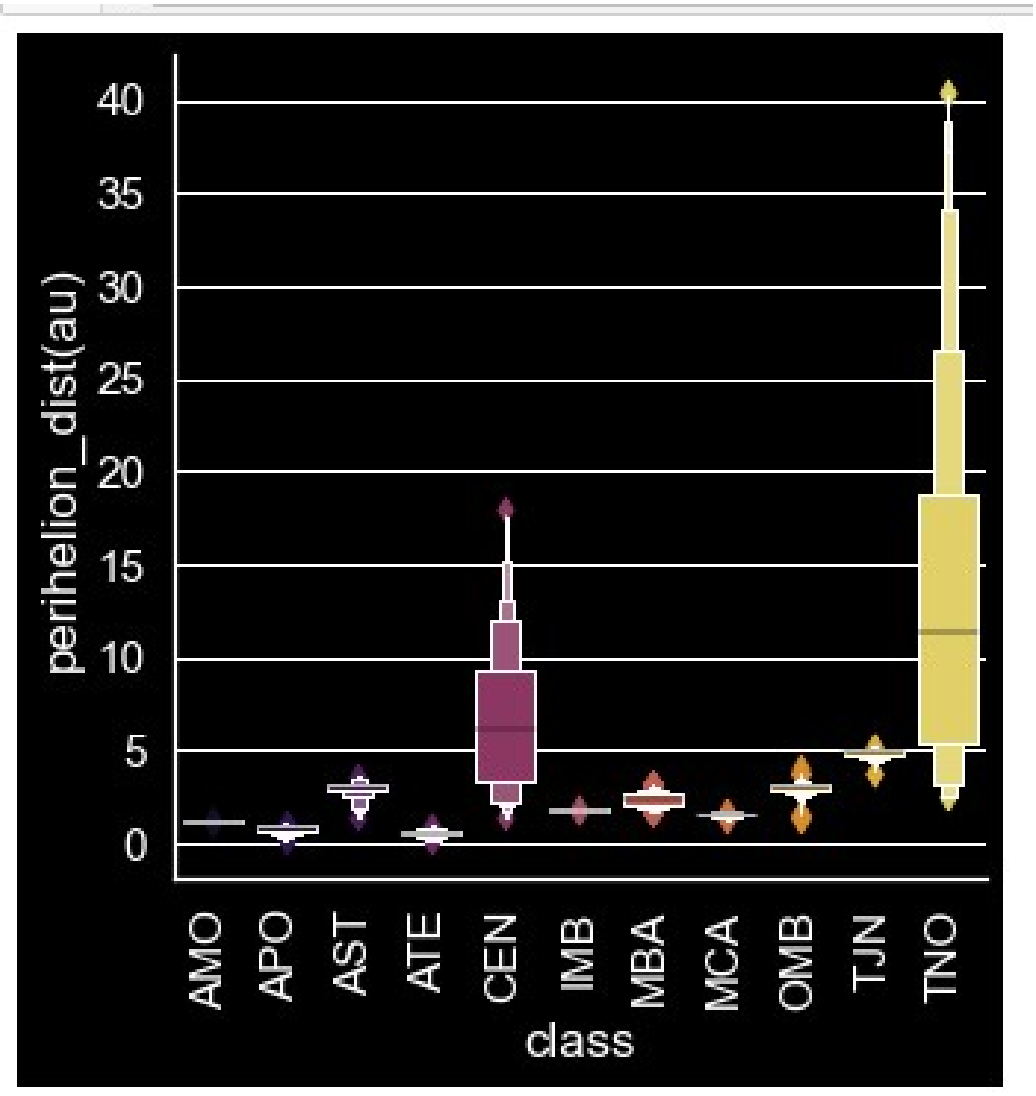

Figure 5: Orbit Class versus Perihelion Distance.

Other features:

- **Eccentricity**
    - Describes how elliptical an orbit is; an eccentricity of 0 describes a circular orbit, and an eccentricity of 1 describes a parabolic orbit.
- **Mean Anomoly**
    - The product of an orbiting body's mean motion and time past perihelion passage.
- **X-Y Inclination**
    - The tilt of the objects orbit.

Data Modeling

Features:

The features that were included at the start of the modeling process were:

- Semi-Major Axis
- Eccentricity
- X-Y Inclination
- Longitude Ascending Node
- Argument Perihelion Distance
- Data Arc
- Number of Observations Used
- Orbit Class
- Mean Motion
- Orbital Period
- Mean Anomaly
- Perihelion Distance
- Diameter

After initial tests it was found that there was a non-linear relationship between Diameter and the features. To adjust for this, the target became the log of Diameter, which yielded better performing models before even adjusting parameters.

It was also found that Orbit Class was not a significant factor in the performance of the models, based on Linear Regression p-values and Random Forest feature importance values. As previously discussed, Orbit Class is defined by other features that more specifically define and asteroid's orbit, so Orbit Class was dropped from further modeling tests.

Data Scaler:

For all models tested, a Standard Scaler was used to transform the data.

Linear Regression with OLS:

Using the default parameters, this model has R^2 and Adjusted R^2 values of 0.716. Figure 6 shows how the actual diameters versus predicted diameters, with the red diagonal being the direct correlation between the two. As can be seen, it performs well in the lower range, but falls further away from the actual value as diameter increases. Any change in combination of features decreased this accuracy even further, so all remaining features were used for this model.

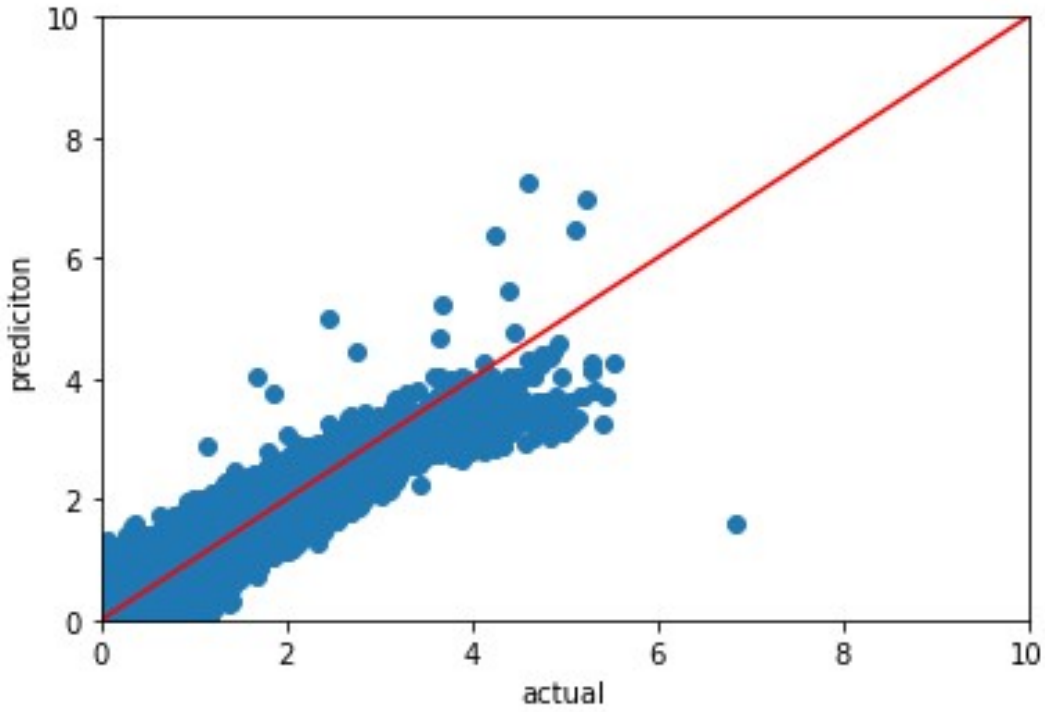

Figure 6: Comparison of actual diameter values to predicted for linear regression using default parameters and all features.

In an attempt to improve performance, ridge regressor and lasso regressor were also tested. For ridge regressor, alphas between 0.01 and 1 were tested, but there was no significant improvement from the linear regression model. The same alphas were then tested for lasso regressor. In this case, any increase in alpha decreased performance.

Random Forest Regressor:

With the default parameters, the random forest regressor performed as follows:

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

This was a significant improvement from the linear regression model. In Figure 7, it can be seen that there are fewer observations that deviated away from the actual values and there is a consistency in performance through the range of values. Other parameters were tested for this model, but they did not improve performance by any significant amount.

All remaining features were used for this model, as well, and Figure 8 shows how they are ranked in terms of importance. As with the linear regression model, and changes to what features were being used resulted in a decrease of performance.

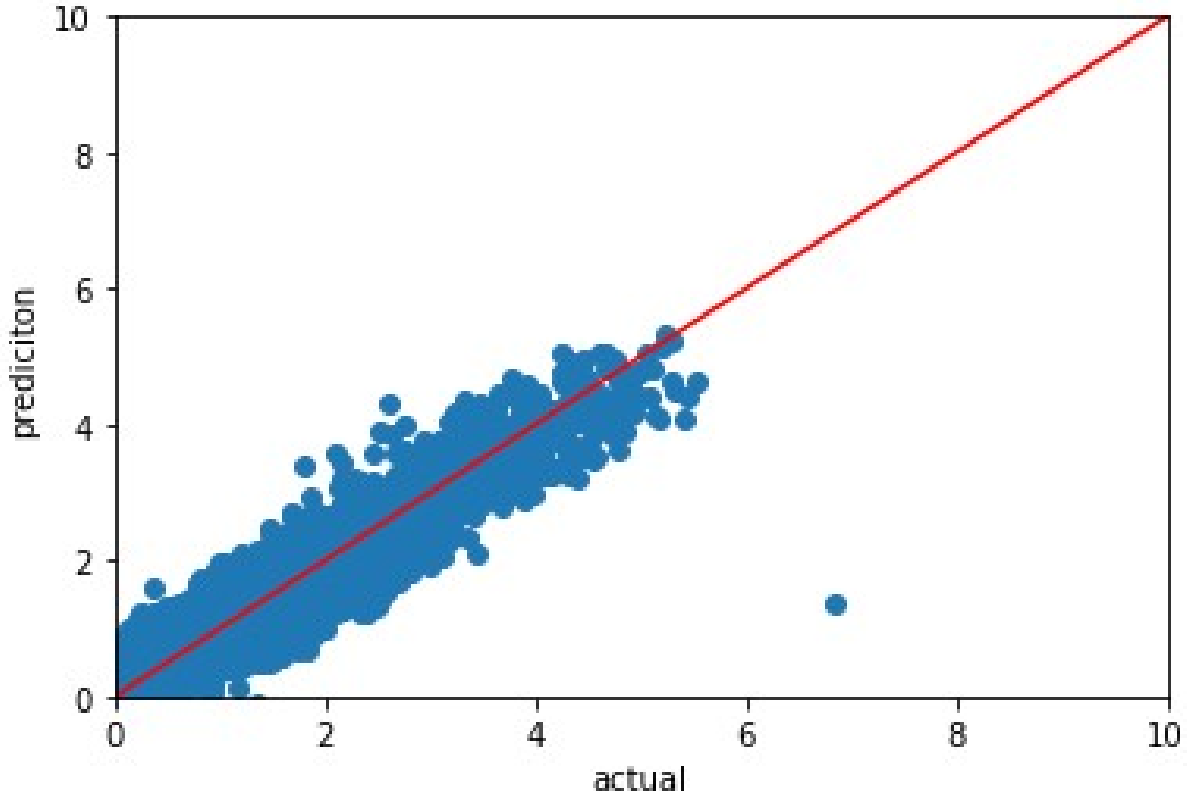

Figure 7: Comparison of actual diameter values to predicted for random forest regression using default parameters and all features.

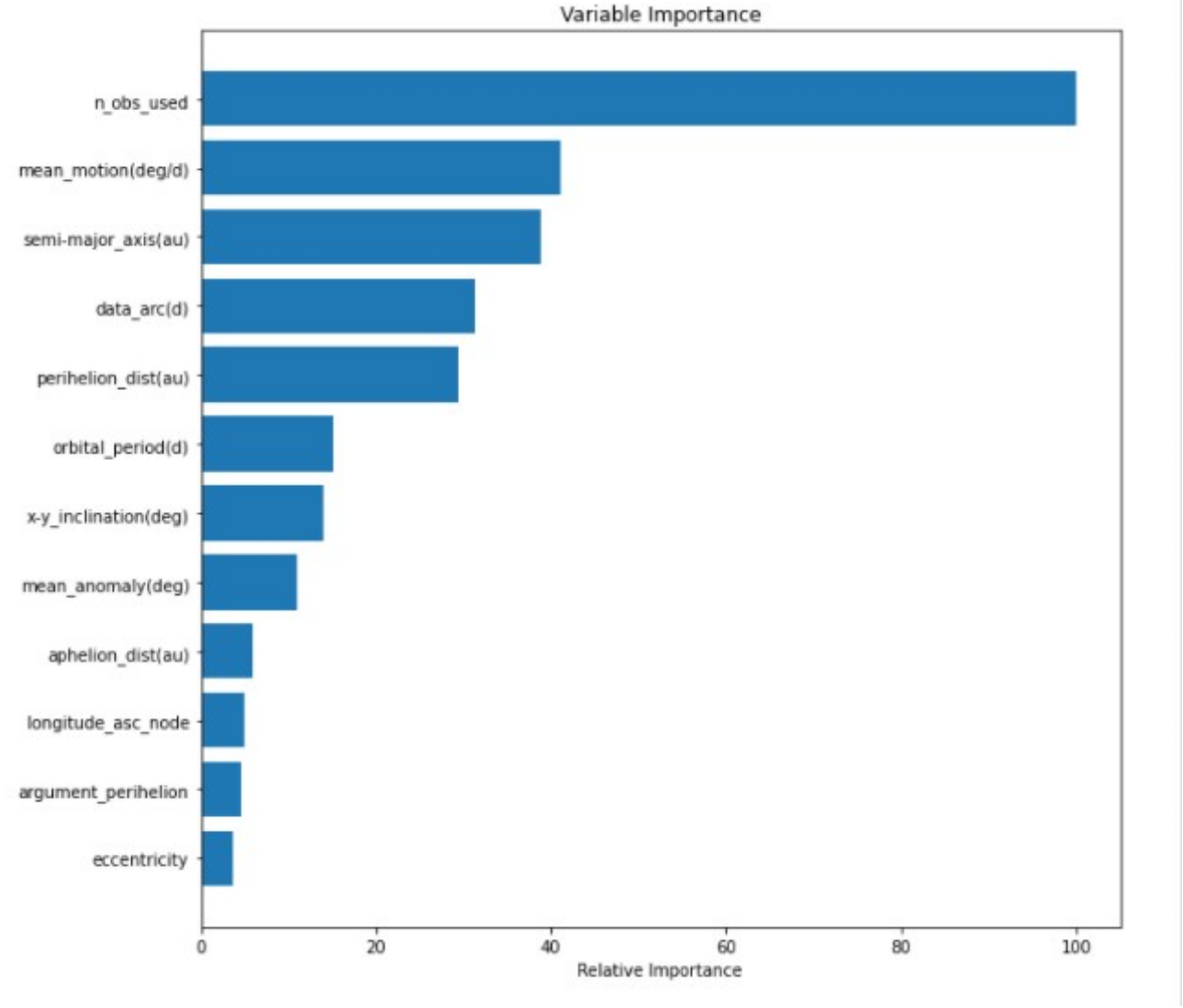

Figure 8: Ranking of the relative importance of the remaining features in the random forest regression model using default parameters.

Gradient Boosting Regressor:

Ranges of parameters were tested for this model, but its best performance was equivalent to the performance of the random forest regressor.

Assumptions and Limitations

Overall Data:

It is assumed that the dataset contains the most accurate values available for all features, including diameter. Even if this is true, there is an inherent amount of inaccuracy to any value that involves a type of distance. This is a result of these types of features being calculated, as physically measuring these values are not possible. Without further research and a better understanding of the topic, it is not known how these features are calculated and to what degree of accuracy is maintained for the values used in these calculations.

Asteroid Class:

The majority of the dataset contained asteroids within the main asteroid belt. This is the largest known collection of asteroids and, because of their location, they are easier to observe than other asteroids. In contrast, there were very few asteroids included from the trans-Neptune objects (TNO) that were used in the model. When looking back at the initial dataset, before any cleaning, many TNO's were included. Most of these observations were unusable in the modeling process because their diameters are still being determined, along with many of the other features.

Technology and Knowledge:

Ability to determine the different features of an asteroid is limited by what technology is currently available. As technology advances, the ability to determine these features will improve. In addition, with further research, a better understanding can be achieved of how these features relate to each other.

Recommendations for Future Work

Infrared Light:

Visible light is not a feature that can be used for determining the size of an asteroid, because different compositions can appear the same size. However, NASA and similar organizations can use infrared light as a more accurate measurement. An example of the difference between visible light and infrared light can be seen in Figure 9. There were not enough values available to use for this project. As more information becomes available, this could be a feature that improves the accuracy of future models.

Asteroid Class:

An initial exploration of creating linear regression models based on distance from the sun was performed, but only to the extent of inside the main asteroid belt, the main asteroid belt, and then outside this distance. This had significant improvement for asteroids outside of the main asteroid belt, but no improvement for the other two sections. Further exploration with breaking the data

even further down into asteroid classes and using other types of models could yield more accurate predictions. As previously described, there are limited observations for asteroids beyond Jupiter, so overfitting for this region would be a concern.

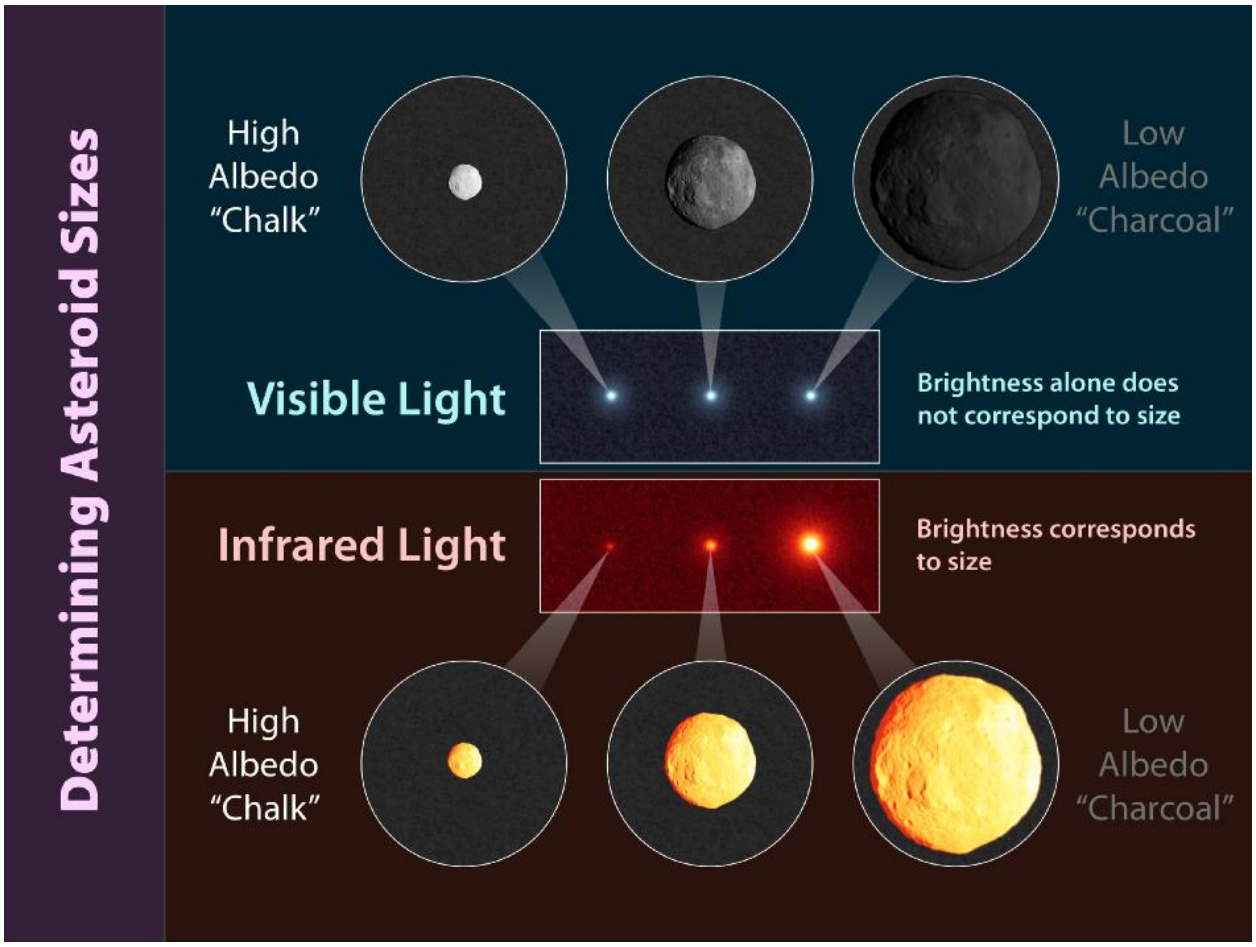

Figure 9: A visual representation of how infrared light is more accurate in determining the size of asteroids of different compositions. (NASA, 2011)

Conclusions

- For this project, regression models were tested for the ability to predict the diameter of asteroids based on a dataset that was created from the JPL Small Body Database. The features that were determined to be the best predictors, from what was available, included features on how observable the asteroid was and various aspects of their orbits.

- It was assumed that the data used was the most accurate available and it is known that the accuracy is limited by the accuracy of technology and calculations.

- The relationship between these features and diameter was shown to be non-linear, so the log of diameter was used as the target feature for fitting the models.

- A random forest regressor model performed best for predicting the diameter.

- In the future, the model could be improved from using data on infrared light and by creating models based on where the asteroid orbits.

References

Jet Propulsion Labratory. “JPL Small-Body Database Search Engine”, 2021, https://ssd.jpl.nasa.gov/sbdb_query.cgi.

NASA. “How to Tell the Size of an Asteroid”, 2011, https://www.nasa.gov/mission_pages/WISE/multimedia/gallery/neowise/pia14733.html.

University of Maryland. “Object Classifications”, 2020, https://pdssbn.astro.umd.edu/data_other/objclass.shtml.

Basu, Victor. “Open Asteroid Dataset”, 2020, https://www.kaggle.com/basu369victor/prediction-of-asteroid-diameter.