Comparing Accuracy and Efficiency of Statistical Methods in Estimation of Asteroid Diameter

**Aly Kapasi - Sergio Jimenez - Youngchan Park**

Aly Kapasi:

* **Model Fitting**
* **Final Report**

Sergio Jimenez:

* **Model Fitting**
* **Final Report**

Youngchan Park

* **Exploratory Data Analysis**
* **Final Report**

**Introduction:**

Measuring objects in space is very difficult, more so when the objects are meteors due to their irregular shapes and their distance from Earth. NASA has the Sentry: Earth Impact Monitoring project which keeps a catalog of all asteroids that may have a potential impact with Earth in the next 100 years. This is one example of a project where knowing the diameter of an asteroid is critical; asteroids larger than 2km in diameter have potential to lay devastation to small nations, so being able to measure and predict potential impact of asteroids is a very important task. Determining the diameter of the asteroid would be essential for evacuation efforts because inhabitants would all have to be evacuated to a safe distance from the impact site. For this project, we intend to use analytical models we learned in class to help determine the diameter of the asteroids provided in our dataset.

We will be using a plethora of different models and procedures to predict the asteroid diameter. We will compare the models’ accuracy as well as their run times. We want to find the most accurate model and the most efficient model. We predicted that a light gradient boosted model will be the best for this data, because there is a great range and depth of data, so the models will be constructed with a great range of data and should be able to predict diameter with good accuracy.

Our initial model was a linear regression model, it had a mean absolute error of 2.54 and had a wall clock runtime of 0.123s to fit the data. The model had the diameter of the meteor as the y-variable and the remaining variables as the x variables. The next model we applied was the Gradient Boosted Decision Tree, this model had a mean absolute error of 0.46 and a wall clock runtime of 73.915s to fit the data. This model was a lot more accurate than the linear regression model, but it was far slower as well. The final model used was a light gradient boosted model, this model returned a mean absolute error of 0.40 making it the most accurate model; this model had a runtime of 1.087s. Our hypothesis was proven correct, as the gradient boosted models were very accurate.

**Methodology:**

We are using a dataset from the Jet Propulsion Laboratory of the California Institute of Technology. The dataset contains predictor variables such as eccentricity, the semi-major axis, orbital period and 28 others. All these variables can help in predicting the orbit of the asteroid and by that derive the diameter of the meteor, as the orbit is highly dependent on the geometry of the orbit. The dataset has an initial n = 839714, and a cleaned n = 137636. We conducted some Exploratory Data Analysis on the data, we did this so that we could identify simple trends in the data which would help us understand the data better. After looking through the data and our objective, we decided to use some supervised learning methods such as linear regression, a gradient boosted decision tree, a light gradient boosted model and a multi-layer perceptron.

Linear Regression was one of the primary analyses used on the dataset, this model was used to check whether any number of the variables had a linear relationship with the response variable, we didn’t expect this model to be very accurate. We chose to use both Gradient Boosted Decision Trees and a Light Gradient Boosted Model, we used these models because we thought they would be well suited to handle large quantities of data, we anticipate these models will be well performing.

**Implementation Details:**

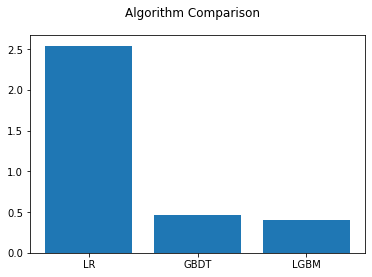
When exploring this dataset, we noticed some features of the dataset which would severely impact the quality of the analyses we would conduct. The dataset was heavily populated with NaNs, as such we decided to omit all rows with NaN entries. We noticed that the majority of the entries for diameter were NaN, after omitting all rows that did not have entries for diameter, we were left with a dataset with n = 137636 which was significantly smaller than the original dataset. We also checked the amount of NaN entries in individual columns, columns with many of those entries were then dropped from the dataset. After filtering the dataset, we needed to format the data types so that we could use them for our analysis; all the datatypes were converted to either integer or float format. When splitting the data, we decided to use an 80:20 split.

We chose to implement a variety of models to see which would best be able to fit the data and help predict asteroid diameter; the various methods were linear regression, gradient boosted decision trees, light gradient boosted model and a multi-layered perceptron. We will then select the most accurate and most efficient model. These models were trained with the diameter values being the response variable and the other selected variables being the predictor variables, this process was completed before the splitting of the dataset.

The first model was the linear regression model, it was fitted with the training data and used the x test data to predict the diameter, finally the mean absolute error was determined. The second model is the gradient boosted decision tree that was fitted by the same process as the previous model. The final model was the light gradient boosted model, which was run multiple times to determine the optimal tuning parameters. When tuning the model, we first determined the optimum number of trees was n = 130. After determining the number of trees we needed to find the appropriate depth and number of leaves for the trees, we determined that the depth = 11 and n\_leaves = 211. Learning rate was the next parameter to be calculated, the optimal rate was 0.05. Lastly the boosting type was checked, of which ‘gdbt’ was the best suited. Using these parameters we fit a new light gbm model, which is more accurate.

**Results and Interpretation:**

First, we look at the distribution of the diameter. As expected, the distribution is skewed to the left (Figure 1). There are many asteroids that have many smaller diameters than large diameters. From the correlation matrix (Figure 2), there is multicollinearity between many independent variables. There are also multicollinearity in the independent variables which are highly correlated to the response variable (diameter). From the k-means clustering to all independent variables (Figure 3), it shows that there are 2 different groups with one small group and one big group. From the k-means clustering to independent variables highly correlated to the response variable (Figure 4), it clearly shows and gives the clue that the independent variables have 2 clusters with unique features between big size diameter asteroids and small size diameter asteroids (Figure 5).

High multicollinearity between independent variables with 2 clusters gives good reason have minimum MAE from light gradient boosted model which inspects the most informative datasets than less informative datasets, and three categorical variables have many factors, these give a reason to use gradient boosted models which penalizes irrelevant variables with decision tree methods. 

The first model we fit the data with was a linear regression model. We fit the data to this model, because we noticed in the heat maps that the variables had most variables were weakly linearly correlated with asteroid diameter. Our regression model had a runtime of 0.133s and a mean absolute error of 2.538. This model doesn’t help predict asteroid diameter too well, but is very fast. Linear regression produced a basic model that was capable of making simple predictions about the data.

The next model we used was the gradient boosted decision tree; this model had a mean absolute error of 0.456 and a runtime of 73.915s. This model is far more accurate than the linear regression model, but it is also far slower. The slowdown in runtime was expected because the gradient boosted decision tree models are much more complex and thorough than the linear regression model. The tradeoff between speed and accuracy is quite extreme for this model.

The last model we fit was a light gradient boosted model, this was expected to be our best model and it was, the model had a runtime of 1.087s with a mean absolute error of 0.40. This is the optimal model, because it’s the most accurate and the most efficient model. Our hypothesis was correct that inspecting only informative columns is helpful than inspecting all columns. This model also proves this dataset has limited important independent variables with non complicated relationships.

From the scratch of the datasets, preprocessing the dataset, testing the normality of the response variables, conducting multivariate analysis and unsupervised learning helps to set up the fitted statistical model with the optimal hyperparameters. The latest machine learning technique light gbm gives the least MAE with much less time than the previous version of the gradient boosting model.

**Supplementary:**

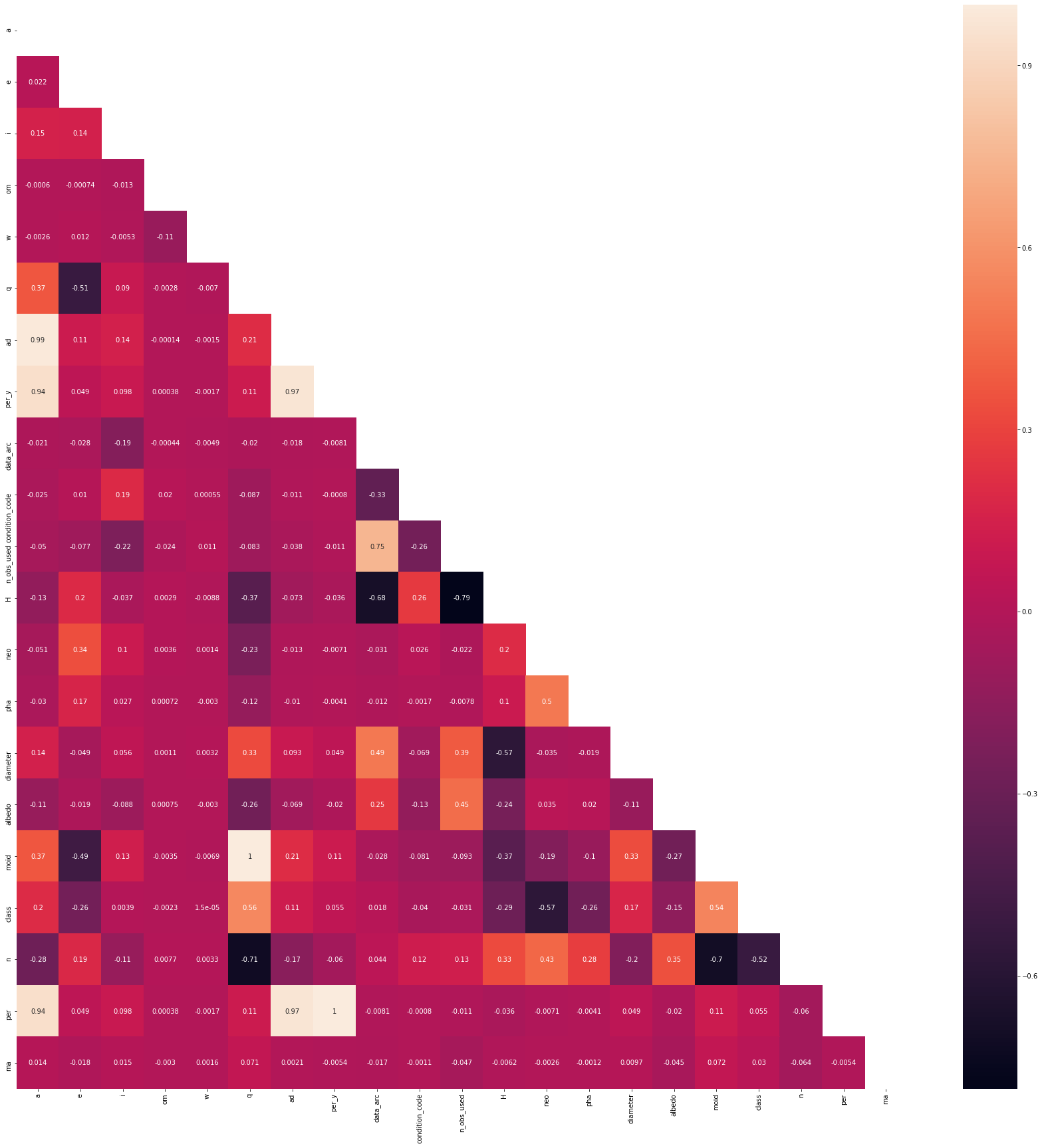


Figure 1. Correlation Matrix(y=Diameter)

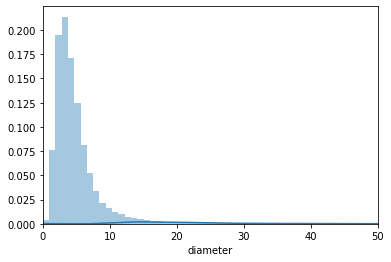
Figure 2. The distribution of Diameter

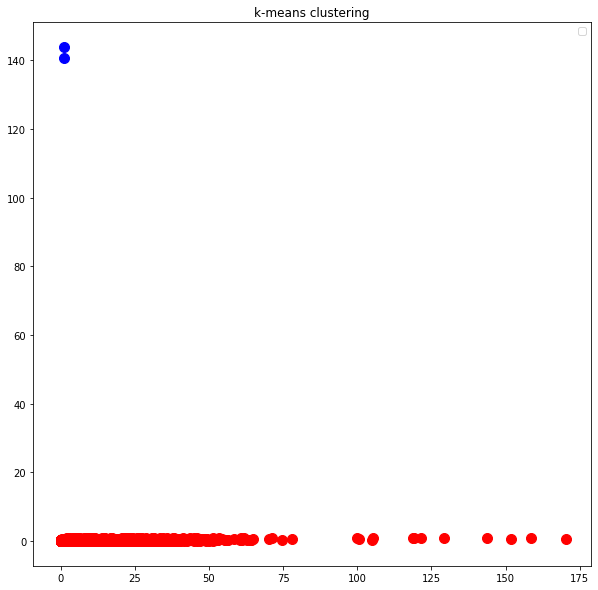
Figure 3. K-means clustering of all independent variables

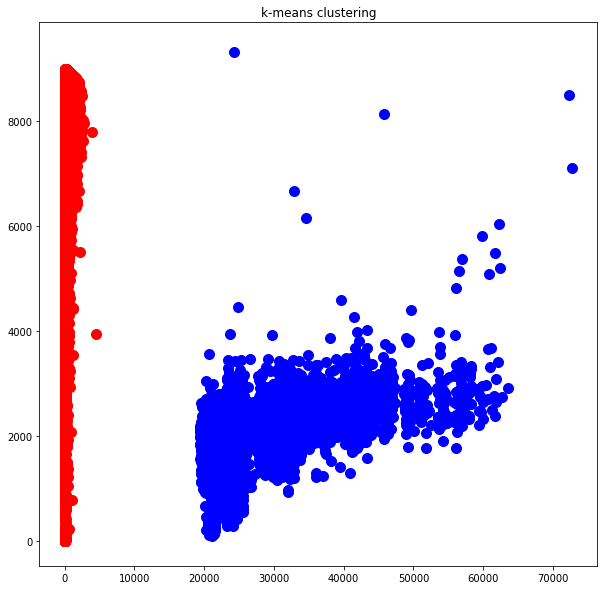
Figure 4. K-means clustering of all independent variables highly correlated to diameter

Figure 5. The mean diameter by orbit class 