# ASTEROID SCIENCE: A STUDY OF ASTEROID FEATURES

*Hrushikesh Akhade, Paritosh Sabade, Deepak Singh Bhadoria*

Indiana University Bloomington

## ABSTRACT

In this paper, we use asteroid data to evaluate the efficiency of different models in predicting the required parameter – asteroid diameter – based on provided parameters. We read data from the CSV file before processing it for training, cleaning it and handling missing information, and performing one-hot encoding on categorical data as required. The data is divided into training and testing sets and is processed using Random Forest, Linear regression, KNN, XGBoost, and Neural Network as training methods. The performance of each training model is compared and presented in the paper.

***Index Terms—*** Random Forest Algorithm, XGBoost, KNN, Linear Regression, Artificial Neural Networks

## 1. INTRODUCTION

Scientists around the world have been observing space objects for their composition, trajectories, and impact on nearby entities for decades now. The most abundant objects – the asteroids – are present in an unbelievably huge number and are drifting between planetary orbits, often crossing paths. This makes them an object of keen interest among the scientific community, and data about different features of asteroids is a highly sought knowledge. Scientists have been using rotation, inclination, eccentricity, semi-major axis, etc. to determine the trajectories, size, and subsequently the composition of the asteroids. In this experiment we are trying to determine some of these features with the help of machine learning models, that read a set of available asteroid features and process it to predict other features – starting with the diameter in this paper. The diameter or the size of the asteroid is a crucial feature of an asteroid. It can help us understand the impact it has on nearby entities and how much of an impact will a collision with such an asteroid cause. It also gives us an idea about the composition, for instance, we know that there is a significant chance of discovering tungsten on an asteroid with a diameter between 70 and 140 miles.

We have identified a Kaggle open-source dataset – Open Asteroid Dataset [1] – to train our model, which provides a set of 31 features for numerous asteroids collected from NASA Jet Propulsion Labs

## 2. PRIOR WORK

There has not been plenty of work done on this research topic due to a lack of data. Since the Jet Propulsion Laboratory of California Institute of Technology which is an organization under NASA released the 'Open Asteroid Dataset', there has been little research on solving the most interesting problems like predicting the diameter of the asteroid and its impact on earth using AI and Machine Learning. Previous approaches to solving this problem with the least error and higher accuracy include Random Forest Algorithm, XGBoost, and Linear Regression. We did a comparative study, by applying various algorithms and tried to increase the accuracy of the model by introducing neural networks.

## 3. APPROACH

We used different supervised machine learning algorithms to calculate the accuracy.

### 1. Linear Regression

It is a linear model that assumes a linear relationship between the input variables (train) and a single output variable(test). It shows the linear relationship between the dependent variable and the independent variable.

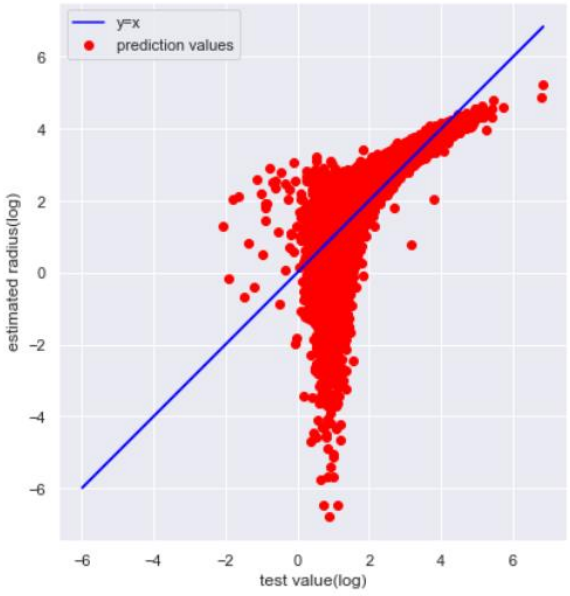


**Figure 1: Test log vs Estimated Radius**

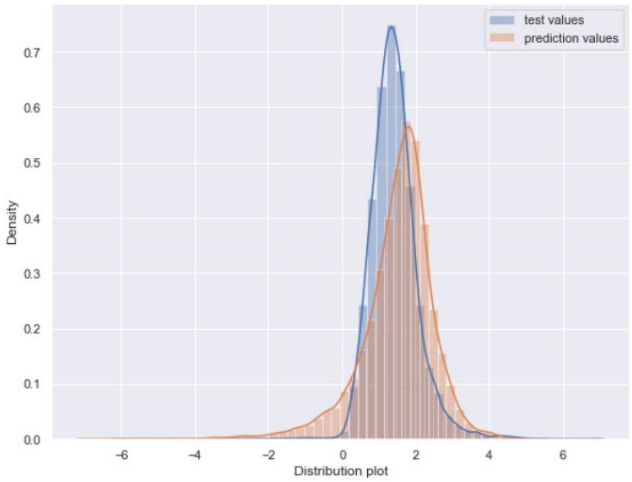


**Figure 2: Distribution plot vs Density**

Model Evaluation for Linear Regression is as follows.
$R^2$ score: 0.5011
Mean Square Error: 62.9633
Mean absolute Error: 2.4620

**2. Random Forest**

It is an ensemble learning method used for both classification and regression. For regression, the mean or average prediction of the tree is returned. We applied hyperparameter tuning to choose the optimal set of parameters for the learning algorithm. We used three different values for max_depth and n_estimators and applied search to find out the best hyperparameters for the random forest Algorithm are Max_depth=30, N_ estimators = 50, and Bootstrap=True.

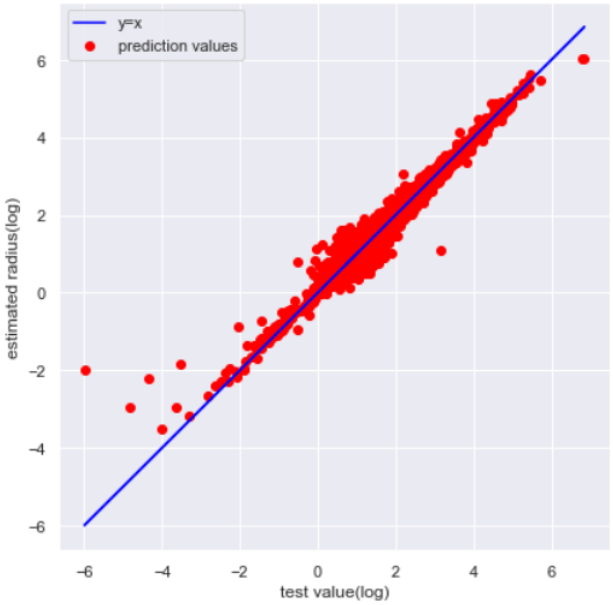


**Figure 3: Test log vs estimated radius**

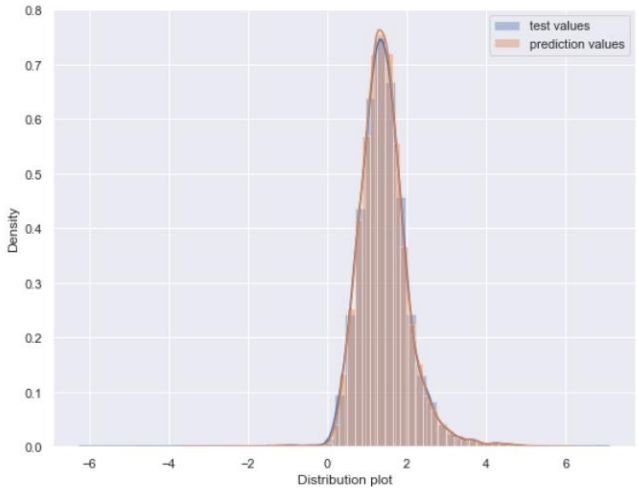


**Figure 4: Distribution plot vs Density**

Model Evaluation for the Random Forest algorithm is as follows:

$R^2$ score = 0.8428
Mean Square Error = 19.83
Mean Absolute Error = 0.4252

**3. K-Nearest Neighbors Regressor**

In the KNN algorithm, the mean or median of continuous values is assigned to K-nearest neighbors from the training dataset and predicts a continuous value for our new data point. In our case, we performed grid search algorithms on three different K- values to find out K with value 4 performs the best among K=3 and K=5.

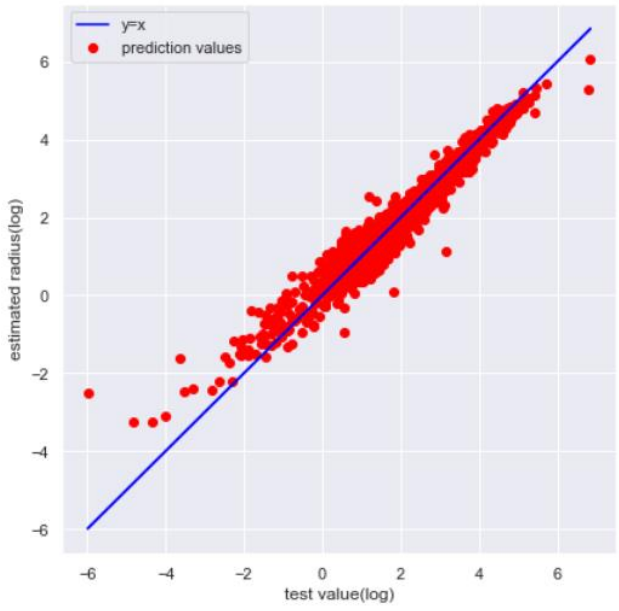


**Figure 5: Test log vs Estimated radius**

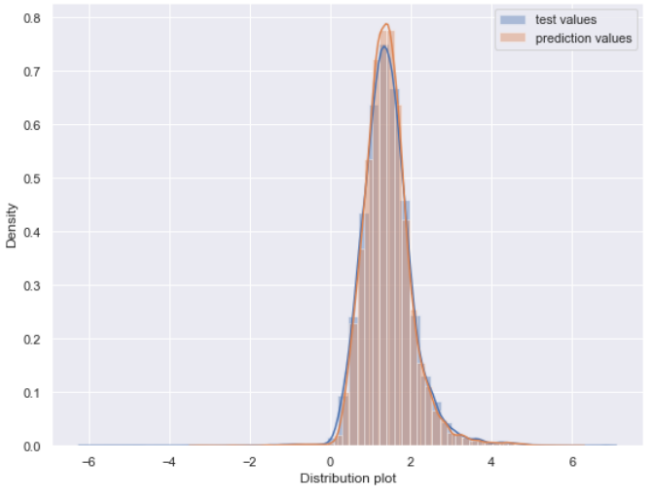


**Figure 6: Distribution plot vs Density**

Model Evaluation for the KNN is as follows.

$R^2$ score: 0.7577
Mean Square Error: 30.57
Mean Absolute Error: 0.6521

**4.XGBoost**

XGBoost, eXtreme Gradient Boosting is a library used for higher computational speed and better model performance. We used different numbers of values in each of the parameters like learning rate, maximum depth, n_estimators and colsample_bytree for hyperparameter tuning to find out the model performs the best when n_estimators are set to 600, max_depth to 3, learning_rate to 0.08 and colsample_bytree to 0.4.

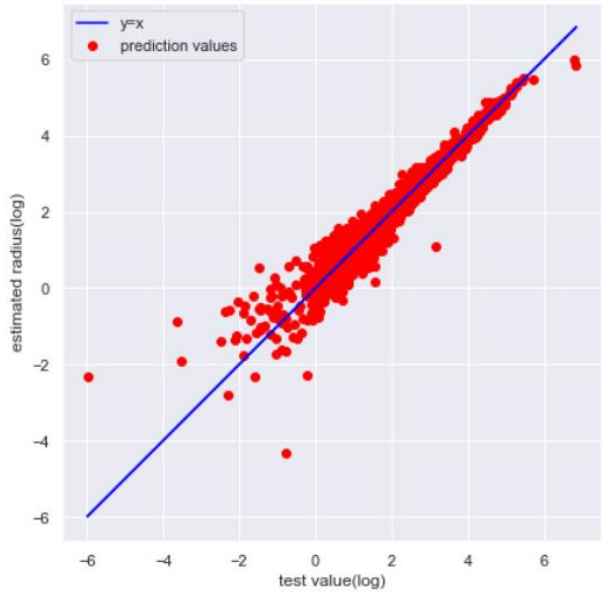


**Figure 7: Test log vs Estimated radius**

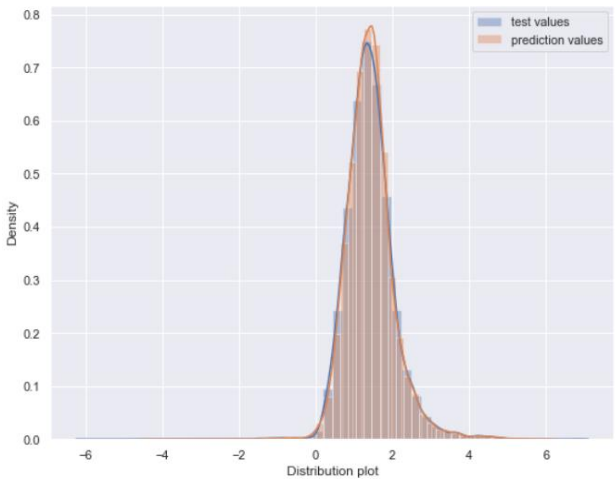


**Figure 8: Distribution plot vs Density**

Model evaluation for XGBoost is as follows:
$R^2$ score: 0.8192
Mean Square Error: 22.80
Mean Absolute Error: 0.4973

**5. Artificial Neural Networks (ANN)**

A neural network is a set of algorithms that attempt to find relationships in a data set using a technique that mimics the human brain. We used the activation function for the first hidden layer, using 24 dimensions and 12 dimensions for the second hidden layer using the relu activation function. The output layer dimension is 1 which holds the output of the problem. For the training purpose, we are using an epoch value of 200 and a batch size of 256, so that it gives us the best result.

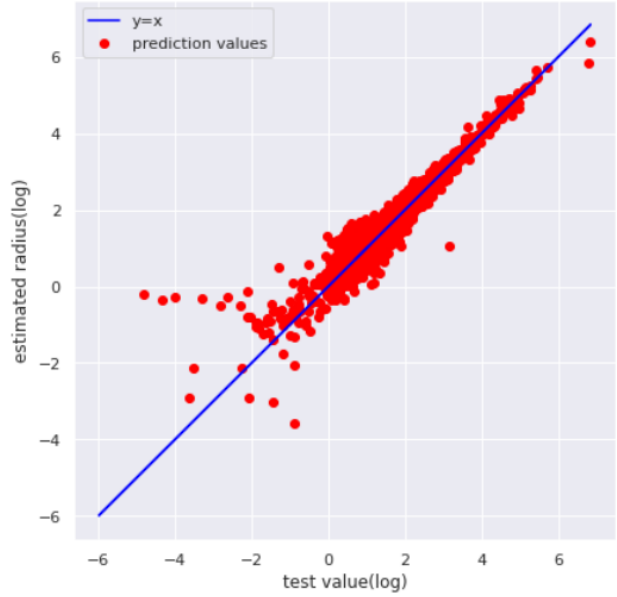


**Figure 9:Test log vs estimated radius**

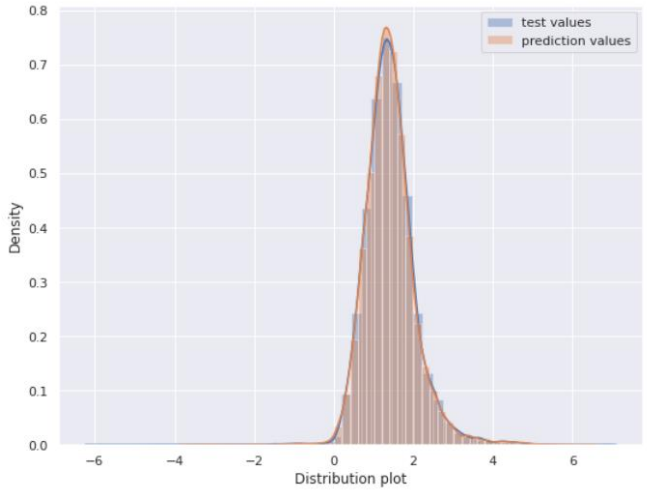


**Figure 10:Distribution vs Density plot**

Model evaluation for ANN is as follows:
$R^2$ score: 0.8683
Mean Square Error: 16.620
Mean Absolute Error: 0.4485

## 4. RESULTS

**1. Comparing the $R^2$ score of all the algorithms used**

$R^2$ interprets how well the regression model explains observed data. A Higher R-square value indicates more variability is explained by the model. A good R-square score can be higher such as 0.9 or above, which shows a higher level of correlation.

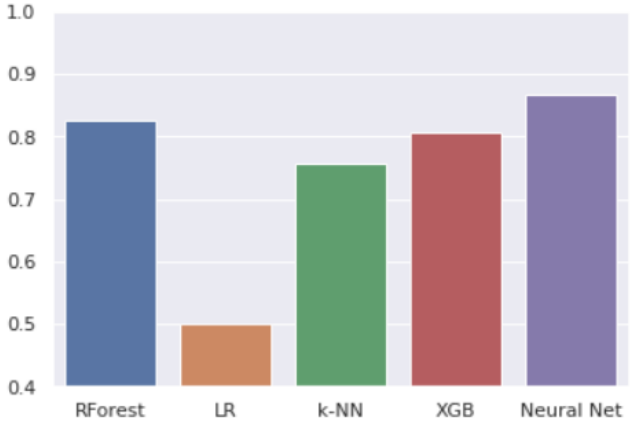


**Figure 11:Bar plot of Algorithms and R-Square score**

As you can observe from the Bar plot given above, Neural Net has the highest R-squared value, followed by Random Forest, XGBoost, KNN, and Linear Regression.

**2. Comparing the Mean Squared Error of all the algorithms used**

Mean squared error (MSE) measures how close a fitted line is to data points. It calculates the distance between the points and regression line and squares it, in order to remove any negative values. Lower the MSE better the forecast. The error-free model gives zero MSE

.

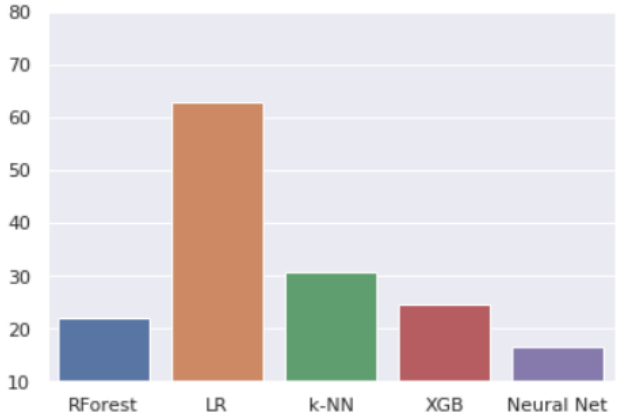


**Figure 12: Bar plot of Algorithms and MSE**

As you can observe from the above bar plot for MSE and algorithms used, Linear Regression gives the highest MSE, giving us the largest error, followed by KNN, XGBoost, Random Forest, and Neural Network. Artificial Neural Network gives us the least MSE, i.e. least error among all other algorithms used.

**3. Comparing the Mean Absolute Error of all the algorithms used.**

Absolute error is the amount of error in our measurements. It is a mean of the difference between the actual value and the measured value in a set of predictions. Mean absolute error measures how far predicted values are away from the actual values. The lower the MAE percentage, the more accurate the model.

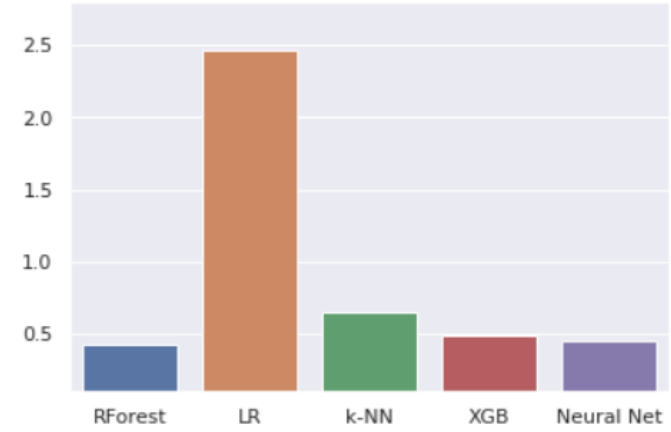


**Figure 13:Bar plot of Algorithms and MAE**

As you can notice from the above bar plot of Algorithms and MAE, Linear Regression has the highest MAE, making it the least accurate model. On other hand, Neural Network has the least MAE and is our best performing model.

## 5. FUTURE SCOPE

We did a comparative study to predict the diameter of the asteroid by applying various supervised machine learning algorithms like Linear Regression, KNN algorithm, Random Forest Algorithm, XGBoost, and Artificial Neural Networks. The Artificial Neural Network performs the best in our case since it gives us higher accuracy, higher R-squared score, and least mean squared error and mean absolute error. While other algorithms produced underfitting models. Using hyperparameter tuning by applying Grid Search Algorithm helped us to find out the best performing parameters and increased our model's performance using Random Forest Algorithm. We tried a different number of input parameters and chose those that gave us the best result in Artificial Neural Network.

In the future, we want to fine-tune neural networks to improve the model efficiency and analyze the best approach by using hyper-parameter tuning. We also want to expand our approach from predicting a single parameter to predicting other important features like the period of revolution, and the composition of the asteroids. The major work would be to fine-tune to models to improve the performance and make models more efficient with hyperparameter tuning and parameter re-selection.

## 6. PROJECT TAKEAWAYS

It was a very fulfilling project with plenty of hands-on experience starting with database selection, data pre-processing, selecting algorithms and mechanisms to train the models, and then performing an analysis of efficiencies of each trained model. Major takeaways would be the experience of being able to implement concepts learned during the semester in class, along with an understanding of how different algorithms perform for the same set of data. We also understood how cleaning data at the earlier iterations is important for efficient training, to save time and resources while training the model.

## 7.SETUP CONFIGURATIONS

**Processor**: Apple M1.8 core CPU, 7 core GPU
**Operating System**: macOS
**System**: 64-bit operating system
**Tools**: Jupyter Notebook

## 12. REFERENCES


[1]Victor Basu," Prediction of Asteroid Diameter with the Help of Multi-Layer Perceptron Regressor," *International Journal of advances in electronics and computer sciences ( IJAECS )*, Volume-9, Issue-2 ( Feb 2022 ).

[2]Alexei Botchkarev, *"Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology"*

[3]Cox Jr, J. E., & Loomis, D. G. (2007), *"A managerial approach to using error measures in the evaluation of forecasting methods"*, International Journal of Business Research, 7(3), 143- 149

[4]Alexei Botchkarev, *"KNN Model-Based Approach in Classification"*, International Journal of Business Research, 7(3), 143- 149, Gongde Guo1, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer

[5]Gerard Biau, *"Analysis of a Random Forests Model"*, Journal of Machine Learning Research 13 (2012) 1063-1095.

[6]Enzo Grossi, Massimo Buscema,"Introduction to artificial neural networks", European Journal of Gastroenterology & Hepatology 19(12):1046-54