TORRENS UNIVERSITY AUSTRALIA

TORRENS

Final Report : Predicting Asteroid Diameters

DIBYAM RAJ BISTA

A00106712

PREPARED FOR :

Rubina Sarkki

TABLE OF CONTENTS

TORRENS UNIVERSITY AUSTRALIA

ACKNOWLEDGEMENT

I wish to extend my sincere thanks to miss Rubina Sarki for her engaging instruction and expert guidance in the CLR204 unit. This report, and the practical understanding of the graphics pipeline it represents, would not have been possible without his support and the excellent learning resources provided.

Dibyam Raj Bista

Student ID: A00106712

TORRENS UNIVERSITY AUSTRALIA

INTRODUCTION - The Data Set

This project was undertaken to develop a regression model capable of predicting asteroid diameters from observational data sourced from the NASA Jet Propulsion Laboratory (JPL). The initial exploration of the dataset, which contains 126,497 records, confirmed its completeness and structural integrity, with no missing values present across its 23 features.

A key finding from the preliminary statistical analysis was the right-skewed nature of the target variable, diameter. The mean diameter (4.16 km) was found to be larger than the median (3.79 km), a statistical indicator of a distribution with a long tail of larger, less frequent values. This characteristic, visually confirmed by the histogram in Figure 1, was a critical consideration for the subsequent modeling strategy. To effectively analyze feature relationships within this large dataset while mitigating the common issue of overplotting, density-based jointplots were employed. These visualizations successfully uncovered the complex, non-linear patterns between key features—such as albedo and moid—and the target variable, reinforcing the need for non-linear modeling techniques.

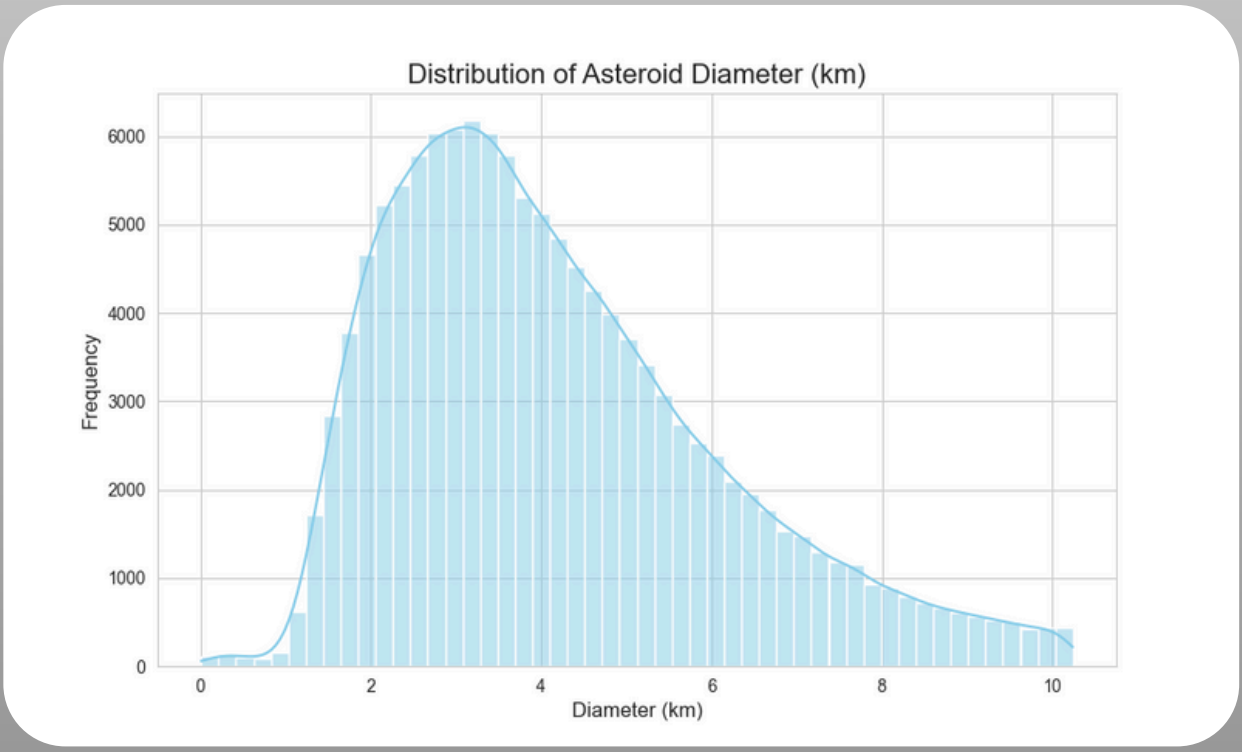

Figure 1: Distribution of Asteroid Diameter (km)

TORRENS UNIVERSITY AUSTRALIA

Model Implementation

A sophisticated and methodical modeling strategy was implemented to ensure both high performance and reproducibility. The preprocessing workflow began with a logarithmic transformation of the skewed diameter variable to achieve a more normal distribution, a crucial step for improving model stability. Additionally, a new feature, kepler_ratio, was engineered from existing orbital parameters to capture deeper physical relationships. The entire process, including one-hot encoding for categorical features and standardization for numerical ones, was encapsulated within a scikit-learn Pipeline to ensure a robust and leak-proof workflow.

A comparative analysis was then conducted on three distinct regression models to identify the most suitable algorithm. Performance was evaluated using Root Mean Squared Error (RMSE) on the original, untransformed scale. The tiered results clearly demonstrated the superiority of non-linear ensemble methods for this problem:

Model	RMSE(km)
XGBoost	0.5621
Random Forest	0.5798
Linear Regression	0.894

The XGBoost Regressor emerged as the top-performing model, achieving a final test RMSE of 0.5621 km, a significant improvement over both the Random Forest and the Linear Regression baseline.

Model Insights

To ensure the final model was not merely a "black box," a deep interpretation was conducted on the winning XGBoost model using SHAP (SHapley Additive exPlanations). The analysis of global feature importance identified n_obs_used (the number of observations) and albedo (surface reflectivity) as the two most influential predictors of an asteroid's diameter.

The SHAP summary plot provided granular insight into the directional impact of these features. Notably, higher albedo values consistently push the predicted diameter down, an observation that aligns with physical principles where smaller bodies can appear bright if they are highly reflective. Conversely, a higher n_obs_used was associated with a larger predicted diameter. The SHAP waterfall plot for a single instance provided a transparent, step-by-step breakdown of how individual feature values contributed to a specific prediction, moving it from the baseline to its final output. This level of interpretability is critical for building trust in the model's outputs and for generating scientifically plausible insights.

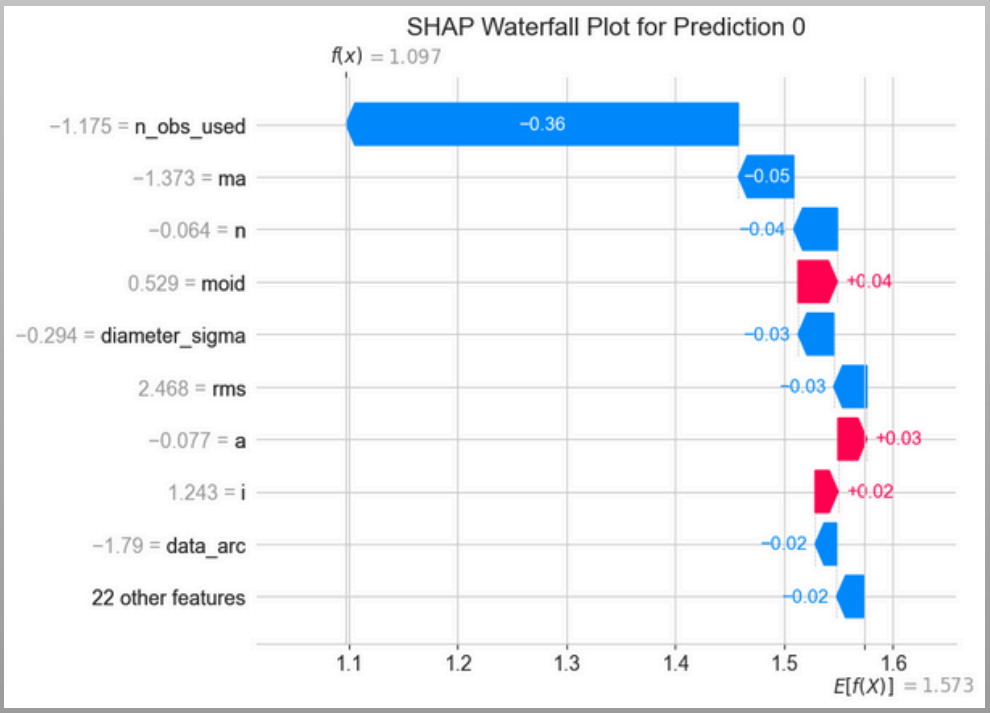

Figure 2: SHAP Waterfall Plot for a Single Prediction

TORRENS UNIVERSITY AUSTRALIA

Recommendations

This project successfully developed a high-performing and interpretable model for predicting asteroid diameters. The XGBoost model, supported by a robust preprocessing strategy, provides an effective solution. For future work, two key areas for improvement are recommended. Firstly, hyperparameter tuning using a systematic process like GridSearchCV could be employed to further optimize the XGBoost model and potentially achieve an even lower RMSE. Secondly, more advanced feature engineering, informed by deeper orbital mechanics, could be explored to create more powerful predictive features.

It is also pertinent to note an ambiguity in the original assessment brief, which suggested implementing 'Logistic Regression'—a classification algorithm—for what is fundamentally a regression task. This was interpreted as an oversight, and a 'Linear Regression' model was correctly substituted as the appropriate linear baseline for comparison, demonstrating a critical approach to the project requirements.