***Asteroid Threat Evaluation: Classification and feature importance investigation on an imbalanced dataset***

Tom Grubb – **09029648**

# Introduction

Asteroid collisions, although a low-probability event, pose a significant threat to our planet, carrying the potential for catastrophic consequences. Unlike many natural disasters, the trajectories of asteroids can be predicted, offering a critical window for detection and response. Leveraging technological advancements and data analysis, scientists endeavour to identify and characterise asteroids to enable early detection and mitigation efforts. [Jamschon MacGarry et al. 2024]

This report delves into the classification and regression of "hazardous" asteroids, utilising advanced machine learning techniques to analyse asteroid data. The project's objectives are twofold:

**Development of a Robust Classification Model:** The primary goal is to construct a machine learning model capable of accurately classifying hazardous asteroids with a high level of precision and accuracy.

**Exploration of Feature Importance:** An investigation into the importance of various features in the classification of hazardous asteroids, shedding light on the key factors influencing their categorisation.

# dataset description and Initial dimension reduction

The dataset used in this analysis comprises data on 4687 asteroids, commonly referred to as "near Earth objects" (NEOs), obtained from the Kaggle website. Each NEO entry encompasses 40 dimensions of data, with no instances of null values. Prior to analysis, cleaning of the dataset was imperative, resulting in the removal of 20 dimensions and 995 records. Below, I explain the rationale behind these exclusions:

1. **Unique Identifiers**: The initial two dimensions, "Neo Reference ID" and "Name," function as unique identifiers for each NEO. Upon review, it was observed that 3692 unique objects exist, necessitating the elimination of duplicate entries alongside these identification dimensions. Additionally, "Orbit ID," serving as an identification number for the orbit of the NEO, was deemed redundant and thus removed.
2. **Dimension Conversion**: Eleven dimensions containing measurements converted into alternative units were omitted. Specifically, eight dimensions provided estimated minimum and maximum diameters of NEOs, while three dimensions delineated relative velocities of NEOs concerning Earth, and four dimensions described the miss distance of NEOs from Earth. To maintain consistency, dimensions utilising SI units were retained:

* Est Dia in KM(min)
* Est Dia in KM(max)
* Relative Velocity km per sec
* Miss Dist.(kilometers)

1. **Time Information**: Five dimensions containing time-related data irrelevant to the models were discarded, including:

* Perihelion Time
* Epoch Date Close Approach
* Orbit Determination Date
* Epoch Osculation
* Close Approach Date

1. **Single-Value Dimensions**: Two dimensions with only one unique value each were excluded:

* Orbiting Body: All objects were found to be orbiting Earth.
* Equinox: All values referenced 'J2000' as the equinox.

Through these exclusions, the dataset was reduced to 3692 records and 20 dimensions.

# dimension Analysis and Further reduction

Given the intricacies of the dataset, it is reasonable to assume that several dimensions are interconnected by fundamental principles of physics and astronomy. Upon examination of the correlations depicted in the heatmap (Figure 1), it becomes apparent that certain dimensions exhibit notably high correlations. Specifically, the dimensions "Jupiter Tisserand Invariant," "Semi Major Axis," "Orbital Period," "Aphelion Dist," and "Mean Motion" demonstrate particularly strong correlations.

A colorful grid with black text

Description automatically generated with medium confidence

1. Heatmap showing the correlation of all 20 remaining features

One notable relationship within this set of dimensions is governed by Kepler's third law of motion, which establishes a connection between the semi-major axis and the orbital period:

Where is the orbital period and is the semi-major axis.

Considering the risk of multicollinearity and lacking comprehensive knowledge of astrophysics, a decision was made to remove four dimensions from this set, leaving solely the orbital period. While any of these dimensions could have been left, preference was given to a value directly related to the primary celestial body, namely Earth.

Additionally, attention was directed towards the estimations of the diameter of the NEOs. Given that the size of asteroids can only be approximated, a range of values is provided. Recognising the high correlation between these estimations, it was deemed appropriate to replace them with the mean average, thereby consolidating them into a new dimension titled 'Est Dia in KM(AVG)'.

A screenshot of a data analysis

Description automatically generated

1. Heatmap showing the correlation of the remaining 14 features

# out of the box modelling results

The initial investigation conducted on the dataset revolves around a supervised binary classification task, predicting the categorisation of asteroids as either 'Hazardous' or 'Non-Hazardous'. Each NEO is currently designated with a value of either "True" or "False", signifying its hazardous nature to Earth. Before fitting the models, the target feature undergoes binary encoding, with "True" values replaced by 1 and "False" values by 0. Notably, the dataset exhibits a substantial imbalance, comprising 3116 non-hazardous NEOs and 576 hazardous ones, resulting in a difference of 2540 instances.

A blue and orange rectangular bars

Description automatically generated

1. Bar plot showing the number Hazardous and Non-Hazardous NEOs

To establish baseline accuracy metrics, the data is fitted to models known to perform well with imbalanced datasets. Ensemble learning models are particularly adept in such scenarios, hence, I have opted for the Random Forest Classifier—an ensemble decision tree classifier—and the Bagging Classifier, utilised for mitigating overfitting. [Singh and Jain. 2022] Additionally, Adaboost and Gradient Boosting, two further ensemble models, are tested. To compare the performance of the ensemble models in prediction of an imbalanced data set three further models are fitted with proven track records in classification problems, naïve bayes, logistic regression and simple neural networks.

The evaluation of these models includes assessing their accuracy, F1-score, and Matthews Correlation Coefficient (MCC). These metrics collectively provide a thorough understanding of the models' performance:

**Accuracy:** Represents the proportion of correctly classified instances among all instances.

**F1-Score**: The harmonic mean of precision and recall, offering a balanced assessment of the model's performance, especially in imbalanced datasets. It considers both false positives (FP) and false negatives (FN) alongside true positives (TP).

**Matthews Correlation Coefficient** (MCC): A measure of the quality of binary classifications, considering true and false positives and negatives, making it particularly useful for imbalanced datasets.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | F1-Score | MCC |
| Random Forest | 99.32% | 97.67% | 97.28% |
| Gradient Boosting | 99.32% | 97.67% | 97.28% |
| Bagging | 99.19% | 97.22% | 96.75% |
| Adaboost | **99.59%** | **98.59%** | **98.36%** |
| Naïve Bayes | 85.52% | 0% | 0% |
| Logistic Regression | 85.52% | 0% | 0% |
| Neural Network | 85.52% | 0% | 0% |

1. Table showing the initial results of the 4 out of the box models

As evident from the results, ensemble models outperformed all others by a considerable margin. Most notable was Adaboost which achieved an accuracy of 99.59% without the need for hyperparameter tuning. All ensemble models excelled across accuracy, F1-score, and MCC metrics while all others although achiving a respecatble 85.52% accuracy failed to even register a value in both F1-score and MCC. The significance of F1-Score and MCC in assessing an unbalanced dataset is evident from the data. Although accuracy scores of 85.52% in the three non ensemble methods give an unrealistic repesentation of their ability to predict hazardous NEOs all that has happened in these cases was that the model has assigned almost all values as non hazardous, resulting in what appears to be great accuracy but appaling F1-score and MCC. In the context of possible life threatening asteroids hurtaling toward the Earth this would be the worst possible outcome for a classification model.

However, to comprehensively evaluate the models, 10-fold cross-validation was performed to assess their consistency, as Naïve Bayes, logistic regression and neural network were shown to be so unsuccessful they were removed from the results yielding:

|  |  |  |
| --- | --- | --- |
| Model | Mean Accuracy | Standard Deviation |
| Random Forest | **99.51%** | **0.0027** |
| Gradient Boosting | 99.43% | 0.0028 |
| Bagging | 99.40% | 0.0029 |
| Adaboost | 99.38% | 0.0032 |

1. Table showing the results of 10-fold cross validation.

A diagram of different models

Description automatically generated

1. Boxplots showing the results of 10-fold cross validation.

The boxplots illustrate the outcomes of 10-fold cross-validation, revealing that the Random Forest Classifier not only yields the highest average accuracy but also demonstrates the highest consistency among the models.

# Feature importance

As well as being good ensemble classifiers the models used can also be used to investigate feature importance, providing an insight into the way in which asteroids are classified. Interestingly although each classifier showed similarly remarkable performance each ranked the features differently as shown in the (fig.7-14)

A graph with blue bars

Description automatically generated

1. Bar Plot showing the best feature for use in the Random Forest model

|  |  |
| --- | --- |
| Feature | Feature Importance |
| Minimum Orbit Intersection | 0.4682 |
| Est Dia in KM(AVG) | 0.1567 |
| Absolute Magnitude | 0.1500 |
| Orbit Uncertainity | 0.0544 |
| Perihelion Distance | 0.0427 |

1. Random Forest top 5 features with rankings.

A graph with blue bars

Description automatically generated

1. Bar Plot showing the best feature for use in the Bagging model

|  |  |
| --- | --- |
| Feature | Feature Importance |
| Est Dia in KM(AVG) | 0.4739 |
| Absolute Magnitude | 0.3201 |
| Minimum Orbit Intersection | 0.1881 |
| Asc Node Longitude | 0.0052 |
| Mean Anomaly | 0.0034 |

1. Bagging top 5 features with rankings.

A graph with blue bars

Description automatically generated

1. Bar Plot showing the best feature for use in the Gradient Boosting model

|  |  |
| --- | --- |
| Feature | Feature Importance |
| Absolute Magnitude | 0.5740 |
| Est Dia in KM(AVG) | 0.2231 |
| Minimum Orbit Intersection | 0.1963 |
| Mean Anomaly | 0.0022 |
| Inclination | 0.0011 |

1. Gradient Boosting top 5 features with rankings.

A graph with blue bars

Description automatically generated

1. Bar Plot showing the best feature for use in the Adaboost model

|  |  |
| --- | --- |
| Feature | Feature Importance |
| Asc Node Longitude | 0.18 |
| Est Dia in KM(AVG) | 0.1 |
| Mean Anomaly | 0.1 |
| Orbital Period | 0.1 |
| Minimum Orbit Intersection | 0.1 |

1. Adaboost top 5 features with rankings.

Based on the charts and tables, it's evident that the three most influential features for Random Forest, Gradient Boosting, and Bagging classifiers are:

**Est Dia in KM(AVG)**: This feature represents the estimated diameter of the NEO, calculated by averaging the estimated minimum and maximum values.

**Absolute Magnitude:** This feature measures the intrinsic brightness of the NEO.

**Minimum Orbit Intersection**: This feature represents the minimum distance between the NEO and Earth's orbit.

It is worth noting that the AdaBoost model diverges from other classifiers by placing particular emphasis on a different top feature: Asc Node Longitude. This feature signifies a significant point in the NEO's orbital trajectory. The model still considers the estimated diameter and minimum orbit intersection as crucial factors.

The results of the feature importance analysis demonstrate that the size of the NEO emerges as the most influential factor in determining its hazardous classification. It's worth noting that among the features considered, size estimation stands out as the only value reliant on estimation rather than direct measurement. This observation underscores the significance of allocating resources towards the development of improved models capable of providing more accurate estimations of NEO sizes.

# conclusion

In conclusion this report shows the benefits of applying advanced machine learning techniques for the classification of hazardous NEOs. The analysis showed how ensemble models including random forest perform extremely well with an inbalanced dataset with all models performing with above 99% accuracy as well as achieving above 96% accuracy in both F1-scaore and MCC metrics. Random forest was the most consistent model achieving an average accuracy of 99.51% with minimal variation after 10-fold cross validation analysis was performed. Furthermore the models showed how the average size of the NEO was the most significant feature in determining its hazardous classification which avocates for continued investment into developing reliable NEO size monitoring models. Considering future research potential in asteroid detection and classification, it's important to delve deeper into feature relationships and data sources to refine model accuracy. Although the models achieved 99.51% this was done without hyperparamenter tuning and so investigation into the number of trees used by the algorithm or the max depth could result in even more accurate figures. Investigating the connections between features and potentially removing those directly related or calculated from each other can streamline model development and offer deeper insights into NEO characteristics. Datasets created from sensors on telescopes or other NEO detection technologies holds immense promise for enhancing data granularity and prediction precision. Collaborating with experts across disciplines, including astronomy and engineering, will be vital in effectively interpreting and leveraging this data. On top of this further work into producing regression models including non-linear regression models to estimate NEO size or minimum orbiting distance could be adopted to help this area.

# Bibliography

Géron, A. (2017) *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. First edition. Beijing ; Boston: O’Reilly Media.

Jamschon Mac Garry, L., Albrecht, R. and Camacho-Lara, S. (2024) Diplomatic, geopolitical and economic consequences of an impending asteroid threat. *Acta Astronautica* [online]. 214, pp. 496–504.

Kaneko, H. (2022) Cross-validated permutation feature importance considering correlation between features. *Analytical Science Advances* [online]. 3 (9–10), pp. 278–287.

Singh, A. and Jain, A. (2022) An efficient credit card fraud detection approach using cost-sensitive weak learner with imbalanced dataset. *Computational Intelligence* [online]. 38 (6), pp. 2035–2055.

Link to my Github depository:

<https://github.com/tgrubb550/Asteroid-Classifier>