***Asteroid Threat Evaluation: Classification and Regression Modelling***

Tom Grubb – **09029648**

# Introduction

# Asteroid collisions although a low probability event, pose a unique threat to the planet, with potentially catastrophic consequences. Unlike many natural disasters, the trajectories of asteroids are predictable, offering a window of opportunity for detection and response. Using technological advancements and data analysis, scientists aim to identify and characterise asteroids to enable early detection and mitigation.

# This report explores “hazardous” asteroid classification and regression, employing advanced machine learning techniques to analyse asteroid data. The aim of my project is threefold:

1. Creation of a machine learning model that will be able to classify hazardous asteroids with a high degree of accuracy.
2. Investigate feature importance in the classification of hazardous asteroids.
3. Creation of regression models for the prediction of the most notable features.

# dataset description and Initial dimension reduction

The dataset used in this analysis comprises information on 4687 asteroids, also known as "near Earth objects" (NEOs), sourced from the Kaggle website. Each NEO entry includes 40 dimensions of data, with no null values present. Prior to analysis, it was necessary to refine the dataset by removing 20 dimensions and 995 records. Below, I outline the rationale for these exclusions:

1. **Unique Identifiers**: The initial two dimensions, "Neo Reference ID" and "Name," serve as unique identifiers for each NEO. Upon inspection, I saw that there are 3692 unique objects, requiring the removal of duplicate entries along with these identification dimensions. Additional to this “Orbit ID” can also be removed being an identification number given to the orbit of the NEO.
2. **Dimension Conversion**: Eleven dimensions containing measurements converted into different units were removed. Specifically, eight dimensions provided estimated minimum and maximum diameters of NEOs, three dimensions detailed relative velocities of NEOs with respect to Earth, and four dimensions described the miss distance of NEOs from Earth. I retained dimensions using SI units for consistency:

* 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 information irrelevant to the models were also removed, 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.

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

# dimension Analysis and Further reduction

Given the characteristics of the dataset, it is reasonable to assume that several dimensions are interconnected through fundamental physics and astronomy principles. Looking at the correlations shown in the heatmap (figure 1). You can see that there are very high correlations between the following dimensions:

* Jupiter Tisserand Invariant
* Semi Major Axis
* Orbital Period
* Aphelion Dist
* Mean Motion

One such equation is Keplar’s third law of motion which links the semi major axis with the orbital period

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

A colorful grid with black text

Description automatically generated with medium confidence

1. Heatmap showing the correlation of all 20 remaining features

Due to the risk of multicollinearity and without sufficient knowledge of astrophysics I decided to remove 4 of these dimensions leaving just the orbital period. I could have kept any of these dimensions but decided to keep a value that has a direct relationship with the primary body in this case Earth.

Finally the last two dimensions that require attention are the estimations of the diameter of the NEO. As the size of the asteroid can only be estimated, a range of values is given. Again, as these values are highly correlated it is suitable to remove these replacing them with the mean average, in a new dimension named ‘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 and feature importance

The first investigation I will be carrying out on the dataset is a supervised binary classification problem predicting the value of the ‘Hazardous’ category. Each NEO is currently assigned a value of “True” or “False” categorising it as either hazardous to the Earth or not. Before fitting the models, this target feature must undergo binary encoding where all “True” values are replaced with 1 and “False” values 0. The dataset is very imbalanced with 3116 non-hazardous NEOs and 576 Hazardous a difference of 2540.

A blue and orange rectangular bars

Description automatically generated

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

To gain some baseline accuracy figures I will be fitting the data to models that have been known to work well with imbalanced datasets. Ensemble learning models tend to have better results when the data is imbalanced and so I will be using the random forest classifier an ensemble decision tree classifier, bagging (bootstrap aggregation) classifier used to prevent overfitting. [Singh and Jain, 2022] Adaboost and Gradientboosting are two further ensemble models that I will test

In order to evaluate the models I measured accuracy, F1-score and Matthews correlation coefficent(MCC). Descriptions of each are listed below.

**Accuracy** measures the proportion of correctly classified instances among all instances.

The **F1-score** is the harmonic mean of precision and recall. It provides a balance between precision and recall and is useful when the class distribution is imbalanced. It considers both false positives (FP) and false negatives (FN) together with true postitives (TP). In asteroid classification, the F1-score provides a single metric that captures the model's performance in terms of both precision and recall.

where

and

**Matthews Correlation Coefficient is** a measure of the quality of binary classifications, considering both true and false positives and negatives. MCC is particularly useful when dealing with imbalanced datasets, as it also takes into account the true negatives.

It is important to note that due to the inbalance of the dataset if a model were to classify all records as non-hazardous then it would gain an 82% accuarcy rating but would achieve 0% in F1-Score and MCC. This is why it is so importnant to include these measures in my evaluation.

|  |  |  |  |
| --- | --- | --- | --- |
| 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%** |

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

As you can see we are already achieving a 99.59% accurate results using Adaboost without the need for hyper parameter tuning. All models performed well in all areas including F1-scaore and MCC. In order to fully evaluate the models kfold cross validation is performed to test the consistency of the models with the following results.

|  |  |  |
| --- | --- | --- |
| 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 results above show that in fact the random forest classifier returns not only the highest average accuracy but is also the most consistent.

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the file “MSW\_USltr\_format”.

## Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

# Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

## Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Units

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive.”
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter,” not “webers/m2.” Spell units when they appear in text: “...a few henries,” not “...a few H.”

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors).*

* Use a zero before decimal points: “0.25,” not “.25.” Use “cm3,” not “cc.” (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in

*a**b*    

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1),” not “Eq. (1)” or “equation (1),” except at the beginning of a sentence: “Equation (1) is ...”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o.”
* In American English, commas, semi-/colons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset,” not an “insert.” The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively.”
* In your paper title, if the words “that uses” can accurately replace the word using, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect,” “complement” and “compliment,” “discreet” and “discrete,” “principal” and “principle.”
* Do not confuse “imply” and “infer.”
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”
* The abbreviation “i.e.” means “that is,” and the abbreviation “e.g.” means “for example.”

An excellent style manual for science writers is [7].

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

### For author/s of only one affiliation (Heading 3): To change the default, adjust the template as follows.

#### Selection (Heading 4): Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

#### Deletion: Delete the author and affiliation lines for the second affiliation.

### For author/s of more than two affiliations: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the “Columns” icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

#### Highlight author and affiliation lines of affiliation 1 and copy this selection.

#### Formatting: Insert one hard return immediately after the last character of the last affiliation line. Then paste down the copy of affiliation 1. Repeat as necessary for each additional affiliation.

#### Reassign number of columns: Place your cursor to the right of the last character of the last affiliation line of an even numbered affiliation (e.g., if there are five affiliations, place your cursor at end of fourth affiliation). Drag the cursor up to highlight all of the above author and affiliation lines. Go to Column icon and select “2 Columns”. If you have an odd number of affiliations, the final affiliation will be centered on the page; all previous will be in two columns.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include ACKNOWLEDGMENTS and REFERENCES, and for these, the correct style to use is “Heading 5.” Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract,” will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1,” “Heading 2,” “Heading 3,” and “Heading 4” are prescribed.

## Figures and Tables

### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1,” even at the beginning of a sentence.

1. Table Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. *(Table footnote)*
2. Example of a figure caption. *(figure caption)*

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization,” or “Magnetization, M,” not just “M.” If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization (A ( m(1),” not just “A/m.” Do not label axes with a ratio of quantities and units. For example, write “Temperature (K),” not “Temperature/K.”

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. L. Jamschon Mac Garry, R. Albrecht, and S. Camacho-Lara, ‘Diplomatic, geopolitical and economic consequences of an impending asteroid threat’, *Acta Astronautica*, vol. 214, pp. 496–504, Jan. 2024, doi: [10.1016/j.actaastro.2023.10.052](https://doi.org/10.1016/j.actaastro.2023.10.052).