

NASA ASTEROID CLASSIFICATION PROJECT

Student: Alison Jing March
University of Colorado, Boulder





AGENDA

- I. Abstract
- II. Introduction
- III. Related Work
- IV. Proposed Work
- V. Evaluation
- VI. Project Timeline

I. Abstract

- The project is about NASA Asteroids classification using its NeoWs(Near Earth Object Web Service) Restful web service for near-Earth Asteroid information. This dataset is directly available on Kaggle(<https://www.kaggle.com/datasets/shrutimehta/nasa-asteroids-classification>). **The primary goal of analyzing this project is to identify whether an asteroid is potentially hazardous and to classify hazardous and non-hazardous ones.** In addition, we are studying the features that are responsible for identifying the hazardous asteroid, and most importantly summarizing the Data Science methodology and process to accomplish these goals.
- **This project is important because NASA hasn't ruled out the risk of an asteroid collision hitting Earth soon.** It's imperative for NASA to develop a rapid and accurate data monitoring and analysis system such as NEO Observations Program(<https://www.nasa.gov/planetarydefense/neo>) for future collision prevention.



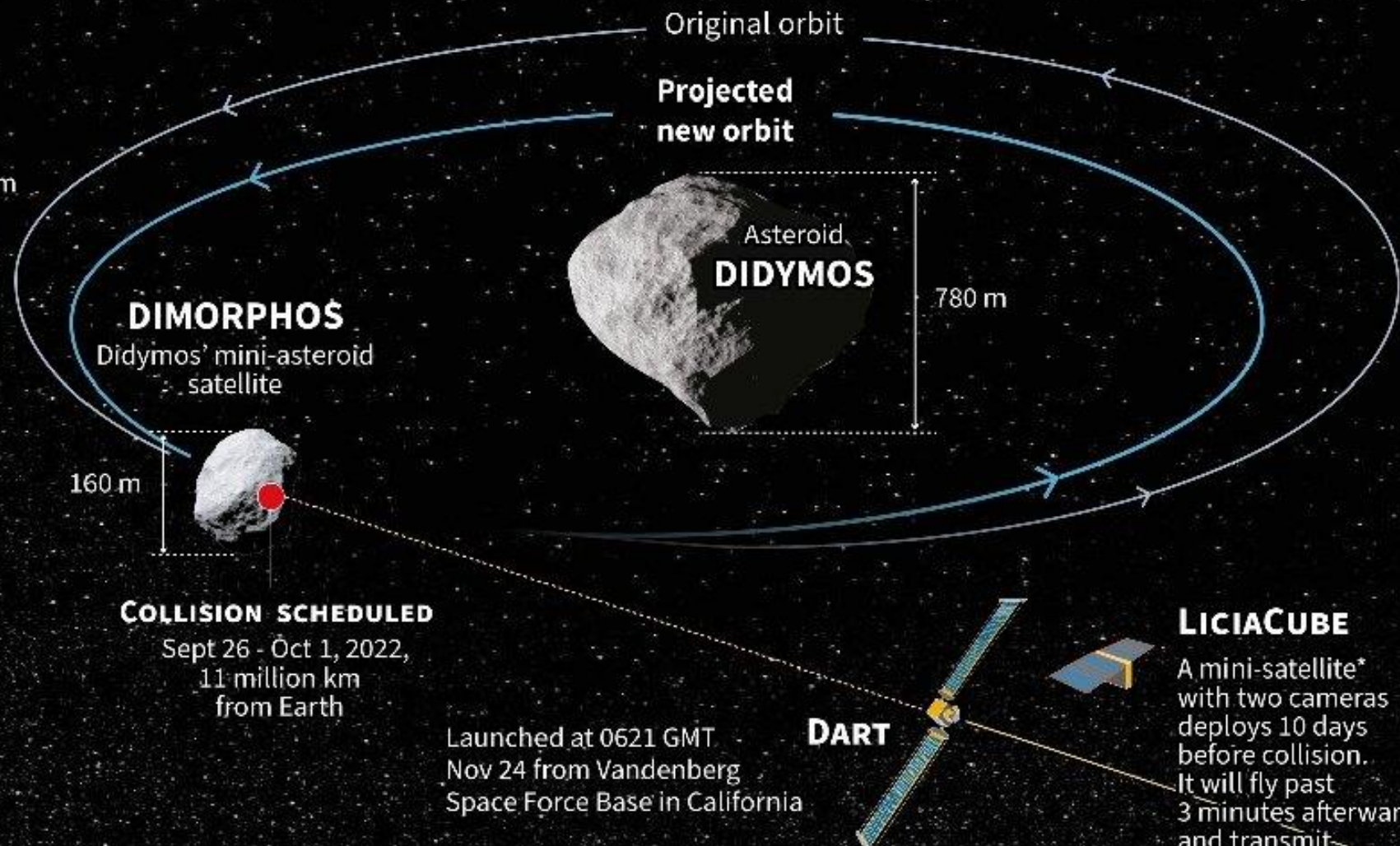
NASA mission to test method to prevent major asteroids impacting Earth

DART (Double Asteroid Redirection Test) aims to crash a spacecraft into a mini-asteroid, orbiting a larger asteroid, and change its trajectory. Neither asteroid poses a threat and there are currently no known asteroids on an impact course with Earth.

DART SPACECRAFT

- Weight: 610 kg
- Crash speed: 24,000 kph
- Dimensions: $1.2 \times 1.3 \times 1.3$ m
- 8.5 m with solar panels
- On board camera to take images on approach

Collision to be observed by telescopes on Earth to measure change in mini-asteroid's orbit



II. Introduction

NASA ASTEROID CLASSIFICATION

The raw dataset consists of 4,687 rows(data instances) and 40 columns(features). The features are listed below:

1. **Neo Reference ID:** reference ID assigned to an asteroid.
2. **Name:** the name assigned to an asteroid
3. **Absolute Magnitude:** refers to the magnitude of an asteroid. An absolute magnitude is a visual magnitude an observer would record if the asteroid were placed 1 Astronomical Unit(au) away, and 1 au from the Sun and at a zero-phase angle.
4. **Est Dia in KM (min):** refers to the estimated minimum diameter of the asteroid in kilometers(KM)
5. **Est Dia in KM (max):** refers to the estimated maximum diameter of the asteroid in kilometers(KM)
6. **Est Dia in M(min):** refers to the estimated minimum diameter of the asteroid in meters(M)
7. **Est Dia in M(max):** refers to the estimated maximum diameter of the asteroid in meters(M)
8. **Est Dia in Miles(min):** refers to the estimated minimum diameter of the asteroid in miles.
9. **Est Dia in Miles(max):** refers to the estimated maximum diameter of the asteroid in miles.
10. **Est Dia in Feet(min):** refers to the estimated minimum diameter of the asteroid in feet.
11. **Est Dia in Feet(max):**
12. **Close Approach Date:** refers to the actual date the asteroid is approaching Earth
13. **Epoch Date Close Approach:** refers to an arbitrary fixed instant of time/date used as a chronological reference approaching date of the asteroid to Earth.
14. **Relative Velocity km per sec:** is the relative velocity of the asteroid in kilometers per second
15. **Relative Velocity km per hour:** is the relative velocity of the asteroid in kilometers per hour
16. **Miles per hour:** relative velocity of the asteroid in miles per hour
17. **Miss Dist.(Astronomical):** the miss distance between an asteroid to Earth in the Astronomical unit(Au)
18. **Miss Dist.(lunar):** the miss distance between an asteroid to Earth in lunar distance unit
19. **Miss Dist.(kilometers):** the miss distance between an asteroid to Earth in kilometers
20. **Miss Dist.(miles):** the miss distance between an asteroid to Earth in miles
21. **Orbiting Body:** the planet the asteroid is revolving around is Earth.
22. **Orbit ID:** the ID # assigned to an asteroid that is revolving around planet Earth.
23. **Orbit Determination Date:** the date and time a brand-new asteroid is verified and tracked using the Orbit Determination process.
24. **Orbit Uncertainty:** a number on a scale from 0-9 to label the orbit uncertainty of an asteroid.
25. **Minimum Orbit Intersection:** the distance between the closest points of the orbit of the asteroid and the Earth's orbit.
26. **Jupiter Tisserand Invariant:** this feature denotes the Tisserand's parameter for the asteroid. Tisserand's parameter is a value calculated from several orbital elements (semi-major axis, orbital
27. **Epoch Osculation:** refers to an asteroid's orbiting osculating time in Epoch time. Judging from the number of digits in decimal format, this could be in Julian Day with fractional day format.
28. **Eccentricity:** refers to the value of eccentricity of the asteroid's orbit. It measures "how circular" an orbit is. A perfect circle has an eccentricity of 0 and an ellipse has an eccentricity approaching but never reaches 1. Eccentricity does not have units.
29. **Semi-Major Axis:** refers to the semi-major axis of the asteroid's elliptical realm. The semi-major axis is measured in astronomical units. One AU is equal to Earth's average distance from the Sun.
30. **Inclination:** measures the angle between the plane of an asteroid's orbit and the plane of Earth's orbit around the Sun.
31. **Asc Node Longitude:** refers to an Asteroid's ascending node longitude, this is an angle in the ecliptic plane between the inertial-frame x-axis and the line through the ascending node.
32. **Orbital Period:** the value of the orbital period of an asteroid, aka the time taken by the asteroid to make one full revolution around its orbiting body.
33. **Perihelion Distance:** refers to the value of the Perihelion distance of the asteroid. Perihelion means the object's closest distance to the Sun measured in AU.
34. **Perihelion Arg:** refers to the argument of the perihelion of the asteroid. It is an angle in the orbit plane between the ascending node and the perihelion point.
35. **Aphelion Dist:** refers to the value of the Aphelion distance of the asteroid. Aphelion is the object's farthest distance from the Sun measure in AU.
36. **Perihelion Time:** the time it takes for the asteroid to reach the closest distance to the sun.
37. **Mean Anomaly:** refers to the fraction of an Asteroid's elliptical orbit's period that has elapsed since the orbiting body passed periapsis; it is expressed as an angle that can be used in calculating the position of the asteroid.
38. **Mean Motion:** refers to the angular speed required for an asteroid to complete one orbit,
39. **Equinox:** all data entries use J2000 which refers to the Julian date 2451545.0 in the concerned time system. In the Gregorian calendar, this is 2000/1/1 12:00:00
40. **Hazardous:** denotes whether the asteroid is hazardous or not.


```
In [20]: 1 nasa.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4687 entries, 0 to 4686
Data columns (total 40 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Neo Reference ID                      4687 non-null   int64
1   Name                                  4687 non-null   int64
2   Absolute Magnitude                   4687 non-null   float64
3   Est Dia in KM(min)                   4687 non-null   float64
4   Est Dia in KM(max)                   4687 non-null   float64
5   Est Dia in M(min)                    4687 non-null   float64
6   Est Dia in M(max)                    4687 non-null   float64
7   Est Dia in Miles(min)                4687 non-null   float64
8   Est Dia in Miles(max)                4687 non-null   float64
9   Est Dia in Feet(min)                 4687 non-null   float64
10  Est Dia in Feet(max)                 4687 non-null   float64
11  Close Approach Date                  4687 non-null   object
12  Epoch Date Close Approach            4687 non-null   int64
13  Relative Velocity km per sec          4687 non-null   float64
14  Relative Velocity km per hr           4687 non-null   float64
15  Miles per hour                        4687 non-null   float64
16  Miss Dist.(Astronomical)              4687 non-null   float64
17  Miss Dist.(lunar)                    4687 non-null   float64
18  Miss Dist.(kilometers)                4687 non-null   float64
19  Miss Dist.(miles)                    4687 non-null   float64
20  Orbiting Body                         4687 non-null   object
21  Orbit ID                             4687 non-null   int64
22  Orbit Determination Date              4687 non-null   object
23  Orbit Uncertainty                     4687 non-null   int64
24  Minimum Orbit Intersection            4687 non-null   float64
25  Jupiter Tisserand Invariant           4687 non-null   float64
26  Epoch Osculation                      4687 non-null   float64
27  Eccentricity                          4687 non-null   float64
28  Semi Major Axis                      4687 non-null   float64
29  Inclination                           4687 non-null   float64
30  Asc Node Longitude                   4687 non-null   float64
31  Orbital Period                       4687 non-null   float64
32  Perihelion Distance                  4687 non-null   float64
33  Perihelion Arg                       4687 non-null   float64
34  Aphelion Dist                        4687 non-null   float64
35  Perihelion Time                      4687 non-null   float64
36  Mean Anomaly                         4687 non-null   float64
37  Mean Motion                          4687 non-null   float64
38  Equinox                              4687 non-null   object
39  Hazardous                            4687 non-null   bool
dtypes: bool(1), float64(30), int64(5), object(4)
memory usage: 1.4+ MB
```

III. Related Work

The following steps have been implemented in Jupyter notebook:

1. Data Cleaning + Preprocessing
 - Check data format + missing values
 - Data Visualization to gain insights
 - Use Heatmap to identify features correlations + outliers(see slide # 8)
 - Use the *OneHotEncoder* method to transform categorical data features(see highlighted features)

IV. Proposed Work

Project Checklist:

1. Frame the problem and look at the big picture
2. Get the data
3. Explore the data to gain insights
4. Prepare the data to better expose the underlying data patterns to Machine Learning algorithms
5. Explore many different models and shortlist the best ones.
6. Fine-tune the models and combine them into a great solution
7. Explore solutions using Tensorflow*
8. Present the solutions
9. Launch, monitor, and maintain the system

References:

Aurelien Geron. 2019. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd. ed.). O'Reilly Media, Inc.

IV. Proposed Work

current progress as of 5.19.23

1. Data Cleaning + Preprocessing

- Check data format + missing values
- Data Visualization to gain insights
- Use Heatmap to identify features correlations + outliers
- Use the **OneHotEncoder** method to transform categorical data features

2. Dimensionality Reduction

- Feature selection
 - Drop features with correlation > 90% to prevent multicollinearity
 - Drop Features like 'Orbiting Body' and 'Equinox'
- Feature Scaling
 - Use **Min-max scaling(normalization) method**: Scikit-Learn **MinMaxScaler** transformer
- Check and Resolve the Imbalance

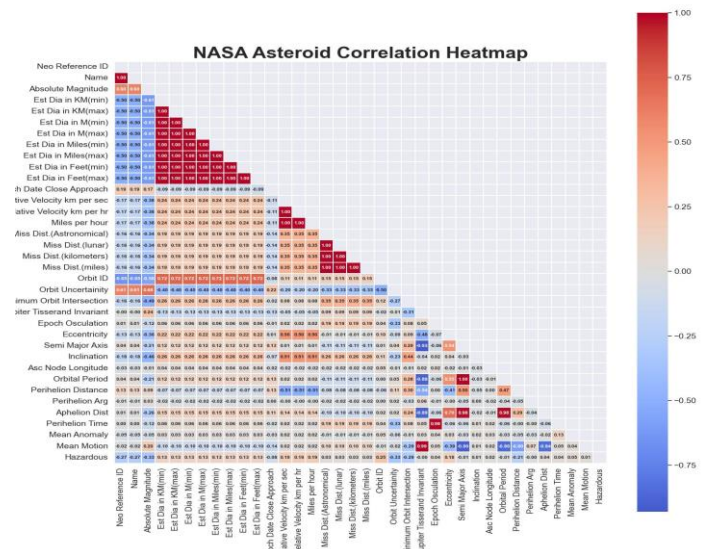
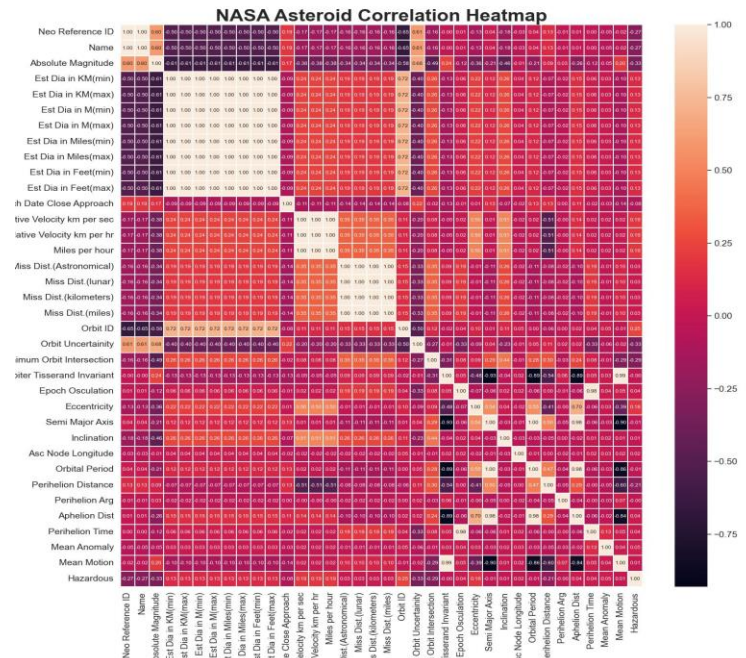
3. Supervised Machine Learning Models

4. Fine-Tune the System

5. Present the Solution + Write Project Report

6. Launch the Solution for Production

- Deploy Final Code + Documentation on Github
- Deploy model with Heroku using Flask



IV. Proposed Work

current progress as of 5.19.23

Tentative Supervised Learning ML Models:

1. Logistic Regression
2. Decision Trees
3. Support Vector Machines
4. Random Forest
5. XGBoosting
- ...

V. Evaluations

1. Cross-Validation

- K-fold cross-validation with three folds(tentatively)

2. Confusion Matrix

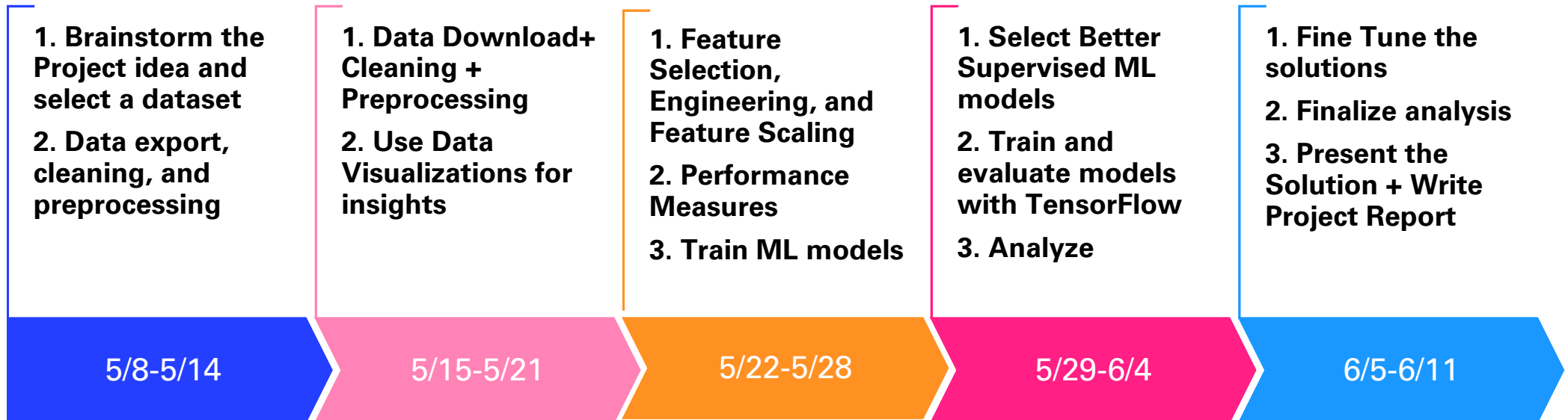
- Precision: $TP/(TP+FP)$
- Recall: $TP/(TP+FN)$
- Specificity: $TN/(TN+FP)$
- F1 score: $TP/(TP+ (FN+FP)/2)$

3. Precision/Recall Trade-off

4. The ROC Curve

5. Other methods(if needed)

VI. Project Timeline



PRESENTATION TITLE

+



o



.



THANK YOU