Potentially Hazardous Asteroid Identification

DS203 Programming for Data Science

Course Project

Anupam Nayak

19d070010

Darin Jeff

190260016

Instructors:

Prof. Amit Sethi

Prof. Manjesh Hanawal

Centre for Machine Intelligence and Data Science

INDIAN INSTITUTE OF TECHNOLOGY BOMBAY

2020

Contents

Abstract

In terms of orbital elements, NEOs (Near Earth Objects) are asteroids and comets with perihelion distance less than 1.3 astronomical units.Since their orbital paths often cross that of the Earth, collisions with near-Earth objects have occurred in the past and we should remain alert to the possibility of future close Earth approaches.The orbit of the asteroid is not very easy to determine considering the number of celestial bodies that influence their trajectory through gravitational forces. As far as, asteroids are concerned, there are many asteroids called near-earth asteroids, but all are not hazardous. Potentially Hazardous Asteroids (PHAs) are currently defined based on parameters that measure the asteroid's potential to make threatening close approaches to the Earth. In these report we try a data driven approach to study the parameters the make an asteroid as a PHA. We considered various parameters from the data collected online through multiple sources. These parameters were analysed for the effect they have on they have on the asteroids level of hazard and redundant parameters were eliminated. Different machine learning models we trained on this data in order to identify an optimal classifier that can used for the purpose of Potentially Hazardous Asteroid Identification and prediction models were identified to predict the of minimum orbit intersection distance.

Keywords: orbital parameters, potentially hazardous asteroids, near-Earth asteroids, minimum orbit intersection distance, machine learning models,Exploratory data analysis, PHA, NEO .

Chapter 1

Introduction

1.1 Background

Near-Earth Objects (NEOs) are comets and asteroids that have been nudged by the gravitational attraction of nearby planets into orbits that allow them to enter the Earth's neighborhood.The vast majority of NEOs are asteroids, referred to as Near-Earth Asteroids (NEAs). Asteroids larger than about 100 meters when they reach the Earth's surface and cause local disasters or produce the tidal waves that can inundate low lying coastal areas. Generally, all asteroids with an Earth Minimum Orbit Intersection Distance (MOID) of 0.05 au or less and a suitable absolute magnitude can be considered as PHAs.NEAs are divided into classes (Atira, Aten, Apollo and Amor) according to their perihelion distance (q), aphelion distance (Q) and their semi-major axes (a).

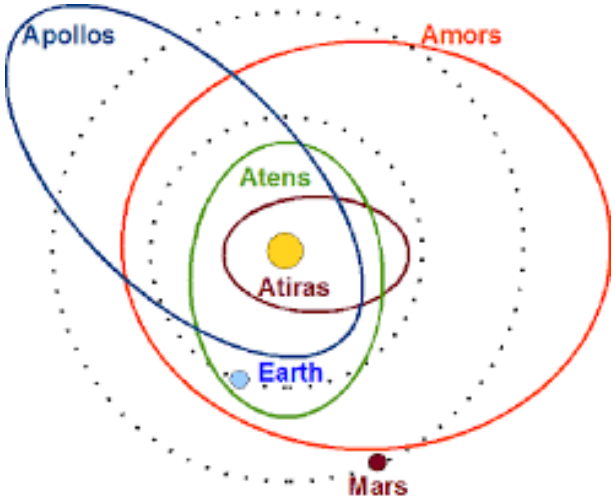

Figure 1.1: Asteroid Classes.

1. **Amors** ($a > 1.0$au, $1.017 < q < 1.3$au): Earth-approaching NEAs with orbits exterior to

Earth's but interior to Mars' (named after asteroid 1221 Amor).

2. **Apollos** ($a > 1.0$au, $q < 1.017$au): Earth-crossing NEAs with semi-major axes larger than Earth's (named after asteroid 1862 Apollo).

3. **Atens** ($a < 1.0$au, and $Q > 0.983$au): Earth-crossing NEAs with semi-major axes smaller than Earth's (named after asteroid 2062 Aten).

4. **Atiras** ($a < 1.0$au, and $Q < 0.983$au): NEAs whose orbits are contained entirely with the orbit of the Earth (named after asteroid 163693 Atira).

1.2 Motivation

With increasing regularity, we are discovering asteroids and comets with unusual orbits – ones that take them close to Earth and the Sun. Just a very few of these bodies are potential hazards to Earth.Most of their orbits become very hard to determine mainly because their orbital paths are influenced by the gravitational tug of several celestial bodies, which cause their paths to alter. Even the diameter of most asteroids is not well determined, as it is usually only estimated based on their brightness and distance, rather than directly measured, e.g. from radar observations. For this reason NASA and the Jet Propulsion Laboratory use the more practical measure of absolute magnitude Any asteroid with an absolute magnitude of 22.0 or brighter can pose a threat depending on the trajectory.

An effort of analyzing orbital distributions of PHAs has already been made by Development of new asteroid-hunting tools for planetary defense, both ground-based and space-based, requires sophisticated analysis of asteroid trajectories, which can be used for efficient survey planning, softening requirements to space-based hardware, and thus reducing the total cost of future asteroid-hunting space missions.Due to such constraint in the prior estimation of Trajectory and size we have no direct function that can be used to classify an approaching NEA into a threat.

Some of the feasible methods that are employed/are undergoing significant research are deflecting an asteroid,this includes nuclear fusion weapons set off above the surface to slightly change the asteroid's velocity without fracturing it. High speed neutrons from the explosion would irradiate a shell of material on the surface of the asteroid facing the explosion.A very modest velocity change in the asteroid's motion (only a few millimeters per second), acting over several

years, can cause the asteroid to miss the Earth entirely. However, the trick is to gently nudge the asteroid out of harm's way and not to blow it up. Another option that has been discussed includes the establishment of large solar sails on a small threatening object so that the pressure of sunlight could eventually redirect the object away from its predicted Earth collision.

1.3 Approach

The approach presented in the report differs from the common statistical treatment of NEAs and relies mainly on the application of machine learning techniques.In general, machine learning frameworks are particularly well-suited for function approximation and recognizing complex patterns and hidden in multidimensional datasets.

In our particular case, because we are no longer reliant on calculations that attempt to estimate the asteroids position at a particular point in time, the machine learning based approach is more resilient to perturbations of the initial conditions, that is, chaotic motion.We did not use artificially generated data with perturbations because this was causing many models as the parameters are of different scales of magnitudes and may have a complicated non linear relationship that could cause slight perturbations of different parameters in the data to defy the underlying physics that is involved in the problem to different scales, may end up confusing the classifier ie the classifier may adopt a method where it classifies the artificial data correctly and the real data would be wrongly classified.

The problem at hand is a discrete binary classification task, where the two mutually exclusive classes for the observed objects. Several models have been analysed ranging from Linear regression and Random forests to support Vector machines and Neural networks.The cleaned data used in the consisted of 16 element uncorrelated fields which were passed on to the potential classifiers.

Chapter 2

Datasets

For this project, we collected our datasets from multiple sources. The major sources of data for this project are:

https://ssd.jpl.nasa.gov/sbdb$_{q}$*uery.cgi*

https://cneos.jpl.nasa.gov/

https://theskylive.com/near-earth-objects

https://sbn.psi.edu/pds/archive/asteroids.html

European Space Agency (ESA)

Some more datasets used for this project are:

https://data.world/texas-whiskey/discoverable-datasets/workspace

https://data.world/qwofford/hazardous-asteroid-orbits

The data from various datasets was initially analysed and cleaned externally and then programmatically concatenated based on the column - *Neo Reference ID*, which uniquely determine an asteroid.

Columns that were sparse or redundant were programmatically removed and columns that contained categorical variables were one-hot-encoded for the machine-learning part. Some sparse columns such as diameter were kept despite their sparsity due to their potential significance in determining the risk level of an asteroid.

The dataset obtained after these steps for used for the further analysis.

Chapter 3

Analysis Pipeline

3.1 Pre-processing

We begin our analysis pipeline with the pre-processing for the dataset. Firstly, we drop the columns with only null entries and give appropriate datatypes to those columns in which the data is stored in an incorrect format. We then remove redundant columns that store the same value in different units such as *'Est diameter in Miles'* and *'Est diameter in Feet'*, fields that contained the standard deviation of a number of observations of the parameters of the same asteroid.We also averaged the minimum and maximum estimated diameter and considered that as the parameter for analysis. We then rename the columns from the different datasets, to standardise the nomenclature used and aid in the concatenation of the datasets. The datasets are then concatenated based on the column - *'Neo Reference ID'* which is unique to an asteroid. Finally, we remove the columns that are very sparse and save the final dataset for further analysis.

3.2 Exploratory Data Analysis

We begin by tabulating the mean, median, skew, max, min, inter-quartile range, and standard deviation for the columns with numerical values. We then plot a heat-map of the correlation matrix of the relevant columns from the dataset to gain deeper insights into the interdependence between different columns. The underlying distributions of the columns are graphically expressed through frequency plots and violin plots. Q-Q plots are also plotted to see the Gaussian nature of the given data. Scatter and Bar plots are also used to graphically highlight the relationship between different columns of the dataset. The plots for this project were made with the seaborn and matplotlib libraries of python.

3.3 Descriptive Analysis

We now try to verify some intuitive relationships that one might expect the columns to have through hypothesis testing. Since the Minimum Orbit Intersection distance is an important determinant of the potential threat level of the asteroid, we tested the following basic hypotheses of the relationship between the *Minimum Orbit Intersection distance* and the *asteroid diameter, aphelion distance, perihelion distance, class* based on the data collected by us.

1. H_0 : Larger asteroids have the same mean Minimum Orbit Intersection with respect to smaller asteroids
   H_1 : Larger asteroids have higher Minimum Orbit Intersection (a measure of is an asteroid is hazardous or not)

2. H_0 : Asteroids with larger Aphelion Distances have the same mean Minimum Orbit Intersection with respect to asteroids with smaller Aphelion Distances
   H_1 : Asteroids with larger Aphelion Distances have the greater mean Minimum Orbit Intersection with respect to asteroids with smaller Aphelion Distances

3. H_0 : Asteroids with larger Perihelion Distances have the same mean Minimum Orbit Intersection with respect to asteroids with smaller Perihelion Distances
   H_1 : Asteroids with larger Perihelion Distances have the greater mean Minimum Orbit Intersection with respect to asteroids with smaller Perihelion Distances

4. H_0 : Asteroids of class Amors same mean Minimum Orbit Intersection with respect to asteroids of class Atiras
   H_1 : Asteroids of class Amors greater Minimum Orbit Intersection with respect to asteroids of class Atiras

5. H_0 : Asteroids with orbits with Eccentricity more than 0.5 have the same mean Perihelion Distance compared to those with Eccentricity less than 0.5.
   H_1 : Asteroids with orbits with Eccentricity more than 0.5 have lesser mean Perihelion Distance than those with Eccentricity less than 0.5.

The statistic tests we use verify our hypotheses are Welch's t test and Wilcoxon signed-rank test.We ended up rejecting the null hypothesis in every case except the case 4.

3.4 Predictive Analysis

We try out different models for our predictive analysis to find the best model that captures the underlying relationship between the *Minimum Orbit Intersection distance* or *pha nature* and the other columns of our table. We initially use a simple linear regression model and even try adding an L_1 (Ridge) and L_2 (Lasso) penalty to improve the results. However, the poor R^2 scores obtained from these models suggest that there is an underlying higher-order non-linear dependence. We then proceed to try out a Random forest model, to predict the *Minimum Orbit Intersection distance* from the other columns of the dataset. This model performed much better than the linear model and provided much higher R^2 scores and a significantly smaller mean square error. Finally, we use a neural networks and support vector machines to predict the *Minimum Orbit Intersection distance* and if the asteroid is hazardous or nor (*pha*) from the other columns. We tried different values for the various hyper-parameters to fine tune the model.

https://github.com/radr44/ds203_proj

Chapter 4

Results and Discussions

	Field_name	mean	median	skew	min	max	25th-percentile	75th-percentile	std_dev
0	Absolute Magnitude	22.93123	23.16950	Left	9.40000	33.20000	20.60000	25.20000	2.97193
1	Relative Velocity km per sec	13.94835	12.81399	Right	0.33550	44.24469	8.31080	18.04763	7.42641
2	Miss Dist.(kilometers)	37784317.98626	38651850.00000	Left	26609.88672	74781600.00000	18385234.50000	57389857.00000	22062578.70788
3	orbit_id	20.25154	9.00000	Right	1.00000	8629.00000	5.00000	20.00000	66.05178
4	Orbit Uncertainity	3.69691	4.00000	Left	0.00000	9.00000	0.00000	7.00000	3.12860
5	Minimum Orbit Intersection	0.09300	0.05257	Right	0.00000	0.70772	0.01551	0.14218	0.10181
6	Jupiter Tisserand Invariant	4.25322	4.01200	Right	-1.20500	9.98700	3.42400	4.87925	1.06894
7	Eccentricity	0.44254	0.45612	Left	0.00285	0.99648	0.31189	0.56838	0.17691
8	Inclination	12.50624	8.82124	Right	0.01451	165.54089	4.57860	17.72210	11.11307
9	Asc Node Longitude	172.89218	173.09038	Left	0.00194	359.97659	82.19231	254.05792	103.69879
10	Orbital Period	993.98208	816.04634	Right	151.19451	2418657.15700	545.38452	1185.63706	15451.08570
11	Perihelion Distance	0.91463	0.96457	Left	0.07051	1.29999	0.78492	1.06844	0.23225
12	Perihelion Arg	182.12661	184.13622	Left	0.00692	359.99310	92.42219	271.77673	104.28794
13	Aphelion Dist	2.64166	2.47198	Right	0.65374	704.01413	1.69992	3.38802	4.67145
14	Mean Motion	0.52028	0.44115	Right	0.00015	2.38104	0.30363	0.66008	0.28114
15	pha	0.08850	0.00000	Right	0.00000	1.00000	0.00000	0.00000	0.28401
16	diameter	0.49437	0.23634	Right	0.00164	37.67500	0.07137	0.56695	1.02944
17	albedo	0.17234	0.13700	Right	0.00900	0.85600	0.04775	0.25325	0.14387
18	rot_per	16.36760	4.61450	Right	0.00330	1880.00000	2.00000	10.43750	65.66563
19	ma	175.45266	170.73634	Right	0.00521	359.99804	53.30233	298.08024	122.53465
20	data_arc	1408.28143	29.00000	Right	1.00000	46330.00000	7.00000	1377.00000	3071.01400
21	condition_code	5.03395	6.00000	Left	0.00000	9.00000	1.00000	8.00000	3.24381
22	rms	0.48182	0.47242	Right	0.03719	1.75200	0.39669	0.55502	0.12599

The above table gives an idea of the distribution of all the major parameters involved in some part of analysis. Most of the models used in the predictive analysis were trained on the non sparse fields in of these parameters as a result of which we were left with just 17 fields namely Absolute Magnitude,orbit id, Eccentricity, Inclination, Asc Node Longitude, Orbital Period, Perihelion Distance, Perihelion Arg, Aphelion Dist, ma, rms, Mean Motion, Minimum Orbit

Intersection, Jupiter Tisserand Invariant, data arc, class, pha' that were uncorrelated and had no null entries the number of rows in the dataset were 24185, We used Principal Component Analysis and K-Nearest neighbours to identify significant dependency and clusters among the data The dataset was divided into different sizes to test validation and train set for analysis purposes.

We examined various relations in the exploratory phases of analysis using scatter plots bar graphs pie charts etc,even the effect of the sparse columns on the pha index/Minimum orbit intersection in order to examine if the data missed entries of a parameter that significantly affects our data. We also made background checks to ensure the parameters used in the analysis are easily estimated with significant accuracy using the existing hardware used by space agencies

	Model	Task	Target Parameter	Critical hyperparameters	Performance
0	Linear Regression	Regression	Minimum Orbit Intersection	-	R-2 score: 0.541
1	Linear Regression with ridge loss	Regression	Minimum Orbit Intersection	lambda = 0.01	R-2 score: 0.541
2	Linear Regression with Lasso loss	Regression	Minimum Orbit Intersection	lambda = 0.001	R-2 score: 0.510
3	Random Forest	Regression	Minimum Orbit Intersection	n_estimators = 150	R-2 score: 0.940
4	Support vector Regressor	Regression	Minimum Orbit Intersection	Kernel = RBF	R-2 score:0.792
5	Support vector Regressor	Regression	Jupiter Tisserand Invariant	Kernel = RBF	R-2 score:0.993
6	Support vector Classifier	Classification	PHA	Kernel = Linear	Accuracy 91.20%
7	Neural Network	Classification	PHA	Hidden layer size = 12, SGD learning rate = 0.001	Accuracy 91.475%
8	Neural Network	Classification	Minimum Orbit Intersection	Hidden layer size = 8, SGD learning rate = 0.001	Mean square error loss 0.34

The results were as shown above(all the inputs were converted to zero mean unit standard deviation for analysis, Performance was evaluated on the test sets).Any combination of hyperparameters donot yield a satisfactory performance while using linear regression models indicating that there is some complicated non linear relation involved in determining whether the asteroid is a PHA or not.

From the table below the best models for classification of asteroids into PHA and non PHA is the Neural network based classifier which was trained using stochastic gradient descent for 7 epochs ,also since the minimum orbit intersection and jupiter tisserand invariant are major factor in determining the whether an asteroid is PHA or not The task is done best using Random forests and Support vector machine with parameters as shown above

The hypotheses that were tested mainly using Wilcoxon signed rank test and Welch t test that resulted in the following data driven conclusions.

1. Larger asteroids have higher Minimum Orbit Intersection (a measure of is an asteroid is hazardous or not)

2. Asteroids with larger Aphelion Distances have the greater mean Minimum Orbit Intersection with respect to asteroids with smaller Aphelion Distances

3. Asteroids with larger Perihelion Distances have the greater mean Minimum Orbit Intersection with respect to asteroids with smaller Perihelion Distances

4. Asteroids of class Amors same mean Minimum Orbit Intersection with respect to asteroids of class Atiras

5. Asteroids with orbits with Eccentricity more than 0.5 have lesser mean Perihelion Distance than those with Eccentricity less than 0.5.

Most of these hypotheses match with our intuition ,The first one in particular where inspite of average diameter being used instead of estimated minimum/maximum diameter.However we didn't consider diameter in our final analysis as most of the estimates of diameter are vary significantly with cases wherein the minimum and maximum estimate of the mean diameter of the asteroid varied as much as thrice the minimum also Absolute magnitude in most cases is considered as a positive correlate of the size of the asteroid. The difficulty in the estimation of diameter is indicated by several sparse rows in the diameter section of the dataset.

PHA classification can be useful for planning future PHA discovery surveys and future asteroid-hunting space missions, yet some work has to be done to verify obtained results. Particularly, the original dataset of NEAs contains survey biases that may influence the true results. This issue can be addressed in the follow up work and obtained results can be tested against debiased model of NEA orbital distribution

Chapter 5

Summary and Conclusions

In this project, we have use machine learning technique to gain deeper insights into the orbital parameters of asteroids and predict the threat these asteroids pose to our planet.

With the help of some graphs and plots, we were able to graphically represent the underlying relationships between different columns of the dataset. Using statistical tests, we were able to accept or reject several intuitive hypothesises of certain relationships that one might expect the columns to have.

We try out different models for our predictive analysis to accurately predict if an asteroid was a threat or not. The parameters we predicted are the *Jupiter Tisserand Invariant*, *Minimum Orbit Intersection* and *pha* (if the asteroid was a threat or not).

Initially, we used a simple linear regression model and even tried adding an L_1 (Ridge) and L_2 (Lasso) penalty to improve the results. However, the poor R^2 scores obtained from these models suggest that there is an underlying higher-order non-linear dependence. We also used a Support Vector Machine (SVM) model which yielded better but sill unsatisfactory results for the *Minimum Orbit Intersection* but gave accurate results for the *Jupiter Tisserand Invariant* when using 'linear' and 'rbf' kernels.

Finally, we turned to Random forest and neural network models, to predict the *Minimum Orbit Intersection distance* and *pha* from the other columns of the dataset. These models performed much better than the linear and SVM models and provided much higher R^2 scores and accuracy respectively. We also plotted graphs that indicate how the loss of our model varied as we fine tuned the hyper-parameters.

For each of the models used, we fine tuned the hyper-parameters so that our model so that it best suits the problem at hand.

Appendix A

Appendix

1. **Near Earth Object:** A Solar System body is a NEO if its closest approach to the Earth (perihelion) is less than 1.3 astronomical units (AU).
2. **Potentially Hazardous Object:** A near-Earth object with a minimum orbital intersection distance with Earth of less than 0.05 astronomical units and an absolute magnitude of 22 or brighter.
3. **NEO Reference ID:** A unique number assigned to every near-Earth asteroid.
4. **Absolute Magnitude:** the visual magnitude an observer would record if the asteroid were placed 1 Astronomical Unit (au) away.
5. **Relative velocity:** The velocity of the asteroid when observed from Earth.
6. **Miss distance:** The distance below which the asteroid will pose a serious threat.
7. **Orbit ID:** A unique number assigned to every observed orbit of the near Earth objects. More than one asteroid can share the same orbit.
8. **Minimum Orbit Intersection:** A measure used in astronomy to assess potential close approaches and collision risks between astronomical objects.
9. **Jupiter Tisserand Invariant:** a value calculated from several orbital elements, used to distinguish different kinds of orbits.
10. **Eccentricity:** A dimensionless parameter that determines the amount by which its orbit around another body deviates from a perfect circle.

11. **Inclination:** The orbital inclination measures the tilt of an object's orbit around a celestial body. It is expressed as the angle between a reference plane and the orbital plane or axis of direction of the orbiting object.

12. **Asc Node Longitude**: The ascending Node Longitude is the angle from the origin of longitude, to the direction of the ascending node, as measured in a specified reference plane.

13. **Orbital Period**: The time a given astronomical object takes to complete one orbit around another object.

14. **Perihelion:** The perihelion is the point in the orbit of a planet, asteroid or comet that is nearest to the sun.

15. **Aphelion:** The aphelion is the point in the orbit of a planet, asteroid or comet that is farthest from the sun.

16. **Perihelion Arg:** It is the angle from the body's ascending node to its perihelion, measured in the direction of motion.

17. **Mean motion:** It is the angular speed required for a body to complete one orbit, assuming constant speed in a circular orbit.

18. **Albedo:** It is a measurement of the amount of light reflected from the surface of a celestial object, such as a planet, satellite, comet or asteroid.

19. **Mean Anomaly(ma):** The fraction of an elliptical orbit's period that has elapsed since the orbiting body passed the nearest point of the orbit to the epicenter, expressed as an angle which can be used in calculating the position of that body in the classical two-body problem.

20. **Rotation Period (rot per):** The time that the object takes to complete a single revolution around its axis of rotation relative to the background stars

References

1. Prediction of Orbital Parameters for Undiscovered Potentially Hazardous Asteroids Using Machine Learning, Vadym Pasko
2. NASA 2018a, Sentry: Earth Impact Monitoring,
3. NASA 2018b, HORIZONS User Manual
4. NASA 2018c, PHA (Potentially Hazardous Asteroid)
5. JMilani, A., Chesley, S., Chodas, P., Valsecchi, G. 2002, in Asteroid Close Approaches: Analysis andPotential Impact Detection eds. W. F. Bottke Jr., A. Cellino, P. Paolicchi, R. P. Binzel (Tucson: Univ. of Arizona Press), Asteroids III, 55
6. Jet Propulsion Laboratory - NASA website
7. Center for Near Earth Object Studies - NASA website
8. The sky live Near Earth Asteroid Database
9. Planetary Data System - NASA website
10. European Space Agency (ESA website)
11. Tricarico P.: The near-Earth asteroid population from two decades of observations, Icarus, Vol. 284, pp. 416–423, (March 2017)

Acknowledgements

We wish to record a deep sense of gratitude to **Prof. Manjesh K Hanawal**, **Prof. Amit Sethi**, **Prof. Sunita Sarawagi** and **Prof. S. Sudarshan** for the opportunity to do this project and their invaluable guidance and council at all stages of this project.

We would also like to thank our parents for their continuous support throughout this project and the course.

A special thanks to our friends, Adit Akarsh and Sumeet kumar Mishra for their ingenious inputs and insightful suggestions.

Finally, we would like to thank the Centre for Machine Intelligence and Data Science department, IIT Bombay for offering this course without which all of this wouldn't be possible.