# DSC 520 Final: Classifying Asteroids

Rahul Rajeev

2023-02-23

## Part 1

### Introduction

Potential Hazard Asteroids or PHAs are currently defined based on parameters that meausre the asteroid's potential to make dangerously close approaches to Earth. One of the parameters called Earth Minimum Orbit Intersection Distance or MOID is measured in astronomical units (AU). Absolute magnitude is another parameter that is the brightness of an object if it were to be seen at a standard distance of 10 parsecs. Asteroids with MOIDS of 0.05 AU or less and an absolute magnitude of 22 or less are considered PHAs.

Detecting PHAs is a data science problem primarily because of the danger large PHAs pose to the Earth. Also, in order for the asteroiud to be a PHA, it must enter Earth's orbit. There is much value in being able to catch asteroids and study anchoring of spaceships, structural characterization, and dust environment of NEOs. Being able to predict any asteroid based on parameters when others might be difficult to find is quite beneficial on all ends.

### Research Questions

1. What other parameters affect PHAs?
2. What are some extreme cases and outliers in the data, and how would it effect the model?
3. Are there specific classes of asteroids more dangerous than others?
4. How are each of the parameters required for this model calculated?
5. Can the same parameters be used to predict other near earth objects like meteors?

### Approach

The approach will be to find datasets of varying parameters, most likely one or be able to combine multiple datasets if any asteroids are missing in one avoiding the duplicates. The datasets should have multiple parameters that can be used in a classification model using training data and then testing data against it from newly collected asteroid data.

### How your approach addresses (fully or partially) the problem.

Creating a model that predicts asteroids as PHAs or close enough to intersect Earth's orbit will help avoid possible collisions and help study asteroid capturing and its benefits.

## Data

NASA Dataset This is the primary dataset updated by NASA 6 years ago that has a majority of the parameters I am interested in, but needs updating, so won't be the only dataset I can use.

JPL Dataset 1 This is a data set shared by MIR Sakhawat Hossain, an astronomy and astrophysics researcher collected from JPL. This dataset has more parameters, but not every parameter may be needed.

JPL Dataset 2 Another dataset of asteroids provided again by JPL, but updated a bit more recently. The set of parameters are the exact same as the previous one, but could have more or less points than the previous.

## Required Packages

I will need packages to read the datasets, which include readxl, or readcsv. To manipulate the datasets, I can use plyr libraries such as plyr, dplyr, tidyr, and purrr. To plot any histograms, correlation plots, and residual plots, I will use ggplot2. Additionally, any packages that will be necessary to machine learning, classification models, creating random samples, and testing data against the training data.

## Plots and Table Needs

As mentioned previously, I can perform some early data analysis and study histograms, residual plots, and correlation plots based on existing important parameters compared to others. As far as tables go, I believe the dataset exists as a csv or a xl file, so it can be loaded as a dataframe into R that can be manipulated and analyzed.

## Questions for future steps

We haven't started discussing how to implement machine learning models in R, so once we do, the additional questions are:

1. What type of machine learning models could be used to classify asteroids?
2. Based on the residual plots, are there any extreme points that could affect the results of the model?
3. How can I change the model to classify other NEOs?

# Part 2

## How to import and clean my data

I have four potential data files I could import and work with. PHA1-1 and 1-2 which are datasets collected by Sentry with asteroids at risk of potential impacts and other asteroids classified as NEOs with their characteristics. I could use the impacts csv to understand possible impacts, and maybe use them as test data for a model created from the second data set. The second dataset contains 958524 asteroid data points with labels on being NEO and a PHA with just information on absolute magnitude and asteroid diameter. The third dataset also contains a subset of phas with specific information about the asteroids.

My approach is to combine the datasets matching by Asteroid name, and removing duplicates of measurements and rows. Since only a small portion of known asteroids are actually NEOs and PHAs, it does narrow down the search quite a bit. With the new detailed and filtered database, I can proceed to the next steps.

## What does the final dataset look like?

```
head(selected)

##                         name neo pha    H diameter albedo         e        a
## 1       433 Eros (1898 DQ)   Y   N 10.4   16.840   0.250 0.2229513 1.458046
## 2     719 Albert (1911 MT)   Y   N 15.4       NA      NA 0.5465585 2.638602
## 3     887 Alinda (1918 DB)   Y   N 13.8    4.200   0.310 0.5703317 2.473737
## 4   1036 Ganymed (1924 TD)   Y   N  9.4   37.675   0.238 0.5330461 2.664725
## 5     1221 Amor (1932 EA1)   Y   N 17.7    1.000      NA 0.4352849 1.919498
## 6    1566 Icarus (1949 MA)   Y   Y 16.9    1.000   0.510 0.8270213 1.078169
##           q         i   moid_ld
## 1 1.1329726 10.830543  57.83961
## 2 1.1964518 11.567485  79.18909
## 3 1.0628863  9.393854  31.99635
## 4 1.2443035 26.677643 134.24653
## 5 1.0839696 11.876537  41.81671
## 6 0.1865002 22.822113  13.32701
```

So I actually ended up just using the second dataset because it provides more than enough information for each datapoint and is updated very recently so the asteroids that were labeled as PHAs in the other two older datasets.

The variables I have chosen are described as follows:

1. Name - Name of the asteroid
2. NEO - classification (yes/no), of near earth orbit
3. PHA - classification (yes/no), of potential hazardous asteroid
4. H - numerical, absolute magnitude of asteroid
5. diameter - numerical, diameter of the asteroid
6. albedo - numerical, geometric albedo
7. e - numerical, eccentricity
8. a - numerical, semi-major axis of orbit in AU
9. q - numerical, perihelion distance of orbit in AU
10. i - numerical, inclination with respect to x-y plane
11. moid_ld - numerical, Earth Minimum Orbit Intersection Distance in AU

These are all variables that could have an effect on whether the asteroid is a potential hazardous asteroid or not.

## What information is not self-evident?

Information about the relationships between the different variables, and if that relationship creates an effect on the identifiability of PHAs is not self-evident. For example, the relationship between ‘a’ and ‘q’ could have collinearity that would affect the binary logistic model, and potentially any machine learning models in the future. Therefore, I would need to test for multicollinearity and independence.

## What are different ways you could look at this data?

1. Grouping parameters differently to create unique binomial regression models
2. Using the data, splitting it into training and test data and run through a classification model or a machine learning model that can predict other parameters given one or two measurable parameters.
3. Classification model on the type of asteroid

## Plan to slice and dice the data?

I plan on first cutting the data into asteroids that are labeled as NEO at first for testing the binary logistic regression model. Then I will run a test of multicollinearity and independence of parameters. Then for the classification model I will use most data points, perhaps cleaning the NA values. For the classification model to classify the type of asteroid, I will also keep most of the data points except for the ones with NA values based on the parameters I choose to predict.

## Plots and Tables to Illustrate Findings?

1. Plot a logistic regression curve
2. Residual plots for each variable if it is numerical
3. When testing for accuracy of model against test data, using a confusion matrix and calculating accuracy
4. Plots to show accuracy of training and testing in a machine learning model?

## Incorporating Machine Learning Techniques?

I will have to understand how to incorporate machine learning to identify the class of asteroid, whether it is a pha, or if there are any missing parameters that could add to help identify pha given initial parameters of measurement and perhaps even the classification.

## Questions for future steps

In addition to the questions from part 1, some additional machine learning modeling questions:

1. What type of model will work the best with classifying asteroids and determining they are phas?
2. What type of model will help predict the characteristics of an asteroid given initial values the best?

# Part 3: The Narrative

## Introduction to the Story

I started studying PHAs in hopes of studying ways to predict them because they pose danger to the Earth. In addition to being able to predict the hazard of an asteroid, there are multiple benefits to studying them including anchroing of spaceships, characterization, and understanding the dust environment. Being able to predict any asteroid based on parameters when others might be difficult to find is quite beneficial on all ends.

## Problem Statement

As asteroids enter and leave orbits close to Earth, it will be beneficial to gather certain parameters and based on previous classifications of PHAs, predict whether these asteroids pose a danger to Earth. In addition to the classification of PHAs, creating a successive model that both predicts the classification and other parameters given initial conditions.

## Addressing the Problem

I gathered three datasets of PHAs from several sources including NASA and JPL. My approach was to combine the datasets matching by Asteroid name, and removing duplicates of measurements and rows. Since only a small portion of known asteroids are actually NEOs and PHAs, it narrowed down the search quite a bit. I ended up just using the second dataset as it was the most expansive and recently updated dataset out of the three, and contained enough information for my analysis.

Since I haven't had much time to work with machine learning this quarter, I worked primarily with fitting the data to a binary logistic model and checking out how the results showed. The other two aspects of the problem statement will be saved for later, as part of a future endeavor.

```
selected$phad <- ifelse(selected$pha == 'Y', 1, 0)
binary_log <- glm(phad ~ H + diameter + albedo + e + a + q + i + moid_ld,
                  family = 'binomial', data=selected)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(binary_log)

##
## Call:
## glm(formula = phad ~ H + diameter + albedo + e + a + q + i +
##     moid_ld, family = "binomial", data = selected)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.53683  -0.03440  -0.00001   0.00000   2.35986
##
## Coefficients:
```

```
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept) 40.241793   5.026798   8.005 1.19e-15 ***
## H           -1.600965   0.198244  -8.076 6.71e-16 ***
## diameter    -2.151582   0.354263  -6.073 1.25e-09 ***
## albedo      -3.922549   1.614346  -2.430   0.0151 *  
## e           -1.341158   3.138641  -0.427   0.6692    
## a            0.443205   0.937847   0.473   0.6365    
## q           -0.802820   2.125324  -0.378   0.7056    
## i            0.002844   0.016610   0.171   0.8641    
## moid_ld     -0.378679   0.041807  -9.058  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 891.95  on 797  degrees of freedom
## Residual deviance: 199.38  on 789  degrees of freedom
##   (22097 observations deleted due to missingness)
## AIC: 217.38
## 
## Number of Fisher Scoring iterations: 10
```

## Interesting Insights

Running the binary logistic model, I found that the parameters that influenced the model the most were H (the absolute magnitude), the diameter, and the MOID. The albedo had a smaller effect, and depending on my level of significance, could also influence the model. Based on readings, I did not predict that the magnitude and the albedo would have an influence, but I had no doubt about the MOID and the diameter being significant parameters.

## Implications

The implications of my data analysis imply that further studies of PHAs should include the magnitude and the albedo possibly. As far as the machine learning models go, they could also be essential components to classifying the asteroids by name or predicting other parameters.

## Limitations

The only limitations I faced for the binary logistic model were the number of parameters that can attribute towards whether an asteroid is potentially hazardous or not. The paradox I face here is whether additional parameters could potentially improve or worsen the model. All three datasets shared more or less of the same parameters, so unless there are new measured for each asteroid or some are found to be more important in the future, giving extra parameters even weight with the existing ones is questionable.

## Concluding Remarks

In conclusion, I believe that my initial attempt at modeling the prediction of PHAs was successful while confirming initial thoughts and uncovering new truths. With additional knowledge of machine learning models in the following years I'm sure I will find better approaches to modeling this data in hopes of discovering further information. As far what models I could possibly use, K-means seems like a great introductory for the model. Given each of the parameters, I can cluster the asteroids into k clusters by the important parameters and new asteroids can be assigned clusters given initially calculated parameters. The model can predict the rest of the parameters based on the cluster it is part of, and therefore, be able to predict whether it is a PHA or not. I don't really know what would be the initial centers of each cluster, or whether I will have to run a k-nearest neighbors model first to determine the centroids of clusters given the labels that exist in the dataset, but it seems that there requires a lot more knowledge on the topic.