Statistical Learning With Applications To Asteroid Data

Nam Nguyen, Qian Chen, Feni Pandya, Willie Lo

Department of Statistics, Rice University

Introduction

The data is obtained from the JPL Small-Body Database Search Engine, which is the most comprehensive database about small objects in the Solar System. The database is maintained by the Jet Propulsion Laboratory and NASA, and is updated on a daily basis. It contains orbital and physical parameters of over 800,000 objects, most of which are asteroids (plus a small number of comets).

Questions

In our studies, we attempt to accomplish the following using machine learning:

1. Characterize the orbital type in terms of semi-major axis and eccentricity
2. Predict if an asteroid poses a threat to the Earth

Exploratory Data Analysis

To explore the data, we look at the characteristics of asteroids with different orbital categories. More than 10 orbital categories are documented in the data set, and the following are the most common:

- MBA: Main Belt Asteroids
- MCA: Mars Crossing Asteroids
- OMB: Outer Main-belt Asteroids
- IMB: Inner Main-belt Asteroids
- AMO: Near-Earth asteroids with orbits similar to 1221 Amor
- APO: Near-Earth asteroids with orbits similar to 1862 Apollo
- TJN: Jupiter Trojan. Asteroids traped in Jupiter's L4/L5 Lagrange points

For each orbital category, I construct a correlation matrix. Please refer to the table below for description of the parameters.

Variable	Description
e	eccentricity
a	semi-major axis (au)
q	perihelion distance (au)
i	inclination; angle with respect to x-y elliptic plane (deg)
om	longitude of the ascending node (deg)
w	argument of perihelion (deg)
ma	mean anomaly (deg)
ad	aphelion distance (au)
n	mean motion (deg/d)
tp	time of perihelion passage (TDB)
tp_cal	time of perihelion passage (ET)
per	sidereal orbital period (d)
per_y	sidereal orbital period (years)
moid	Earth Minimum Orbit Intersection Distance (au)
moid_id	Earth Minimum Orbit Intersection Distance (LD)
moid_jup	Jupiter Minimum Orbit Intersection Distance (au)
t_jup	Jupiter Tisserand Invariant

Table 1: Description of Variables

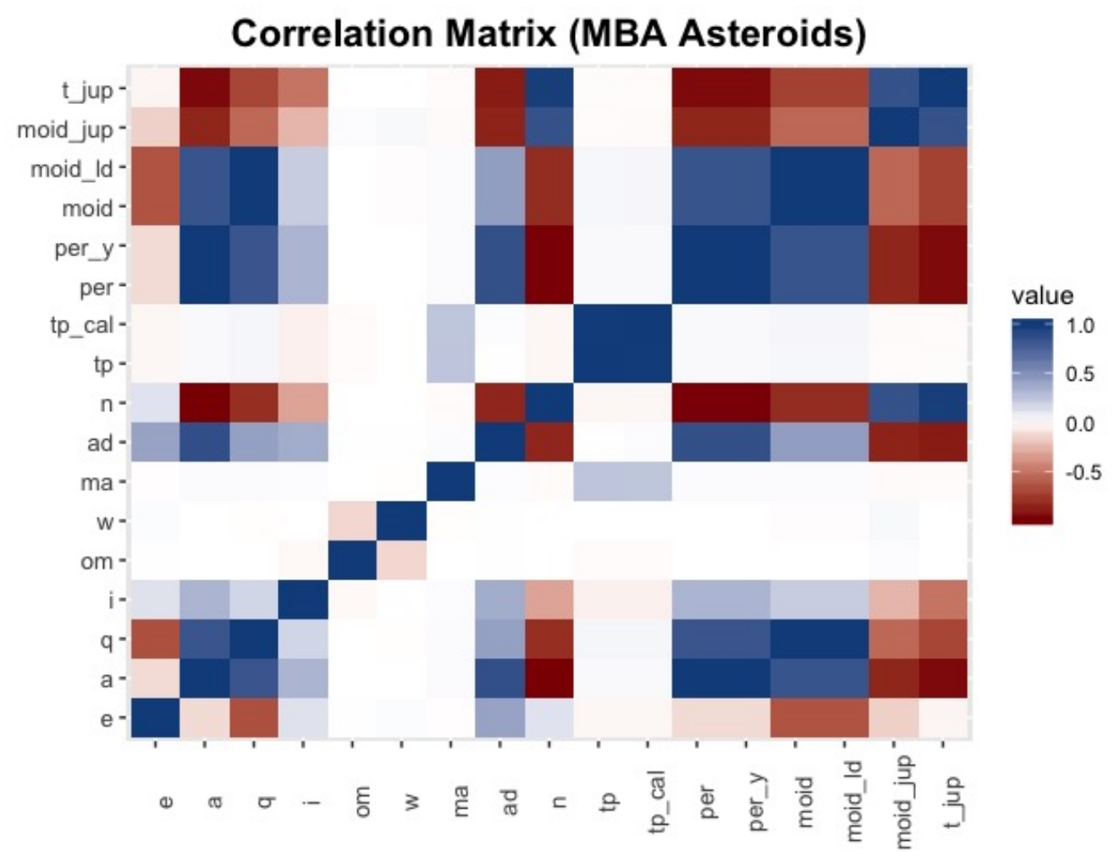

Figure 1: Correlation matrix for MBA asteroids

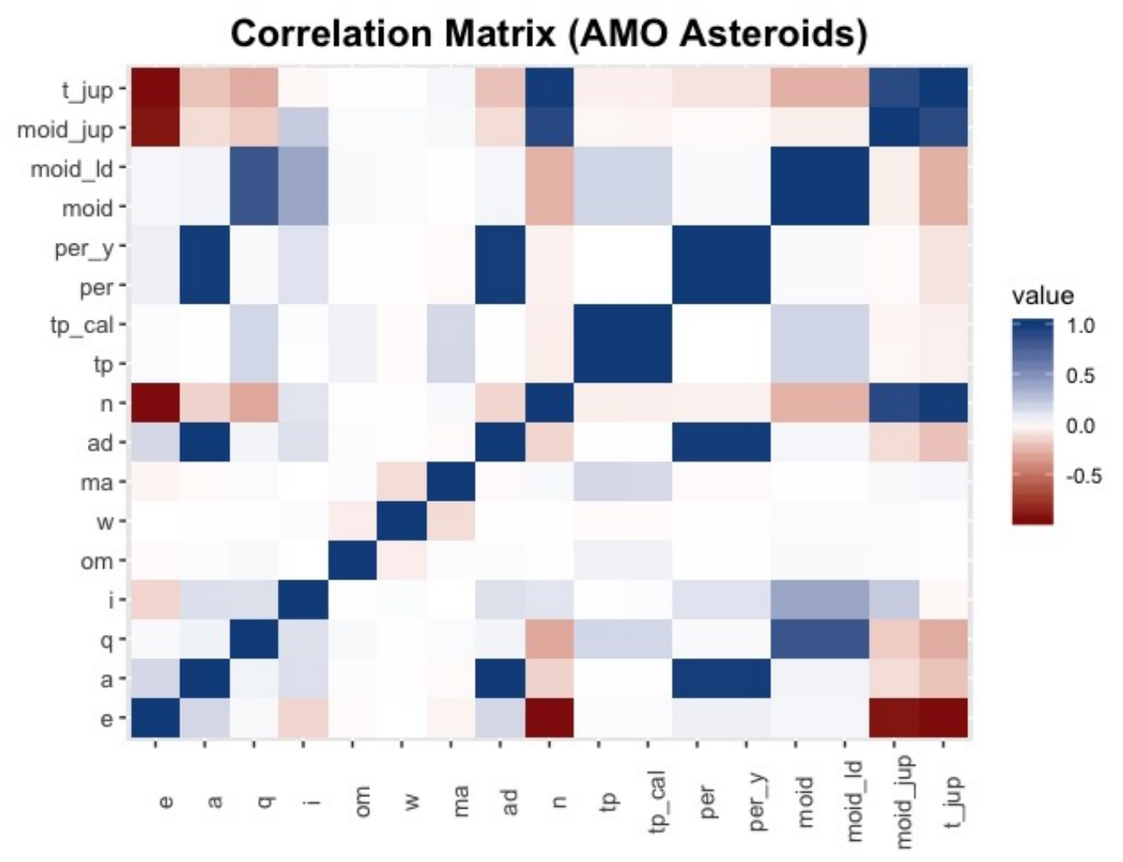

Figure 2: Correlation matrix for AMO asteroids

Machine Learning: Orbit Classification

From the exploratory data analysis, we see that orbits of different types show distinct characteristics. To investigate this problem further, we employ machine learning algorithms to predict orbit classification based on eccentricity and semi-major axis.

- Due to computational limitation, we cannot work with the entire data set (more than 800,000 asteroids). Instead, we train the models on a random subset of the data. The smallest subset has 1,400 observations and the largest has 14,000 observations.
- To test the out-of-sample performance of the models, we also extract a test sample with 13,000 observations.
- The algorithms to be considered are:
  - Support Vector Machine (SVM)
  - K-Nearest Neighbors (KNN)
  - Linear Discriminant Analysis (LDA)
  - Quadratic Discriminant Analysis (QDA)
  - Generalized Logistic Regression (GLR)
- Once the models have been trained, we plot the decision boundaries for each model. It is expected that LDA and Logistic Regression should display linear boundaries whereas decision boundaries for SVM and KNN can be highly non-linear. To do this conveniently, we make a Shiny app, where the user can select the sample size and the algorithm to be used.
- We compute and compare the predictive accuracy of the models. The models are statistically compared using the Stuart-Maxwell test, which is a generalization of the McNemar's test

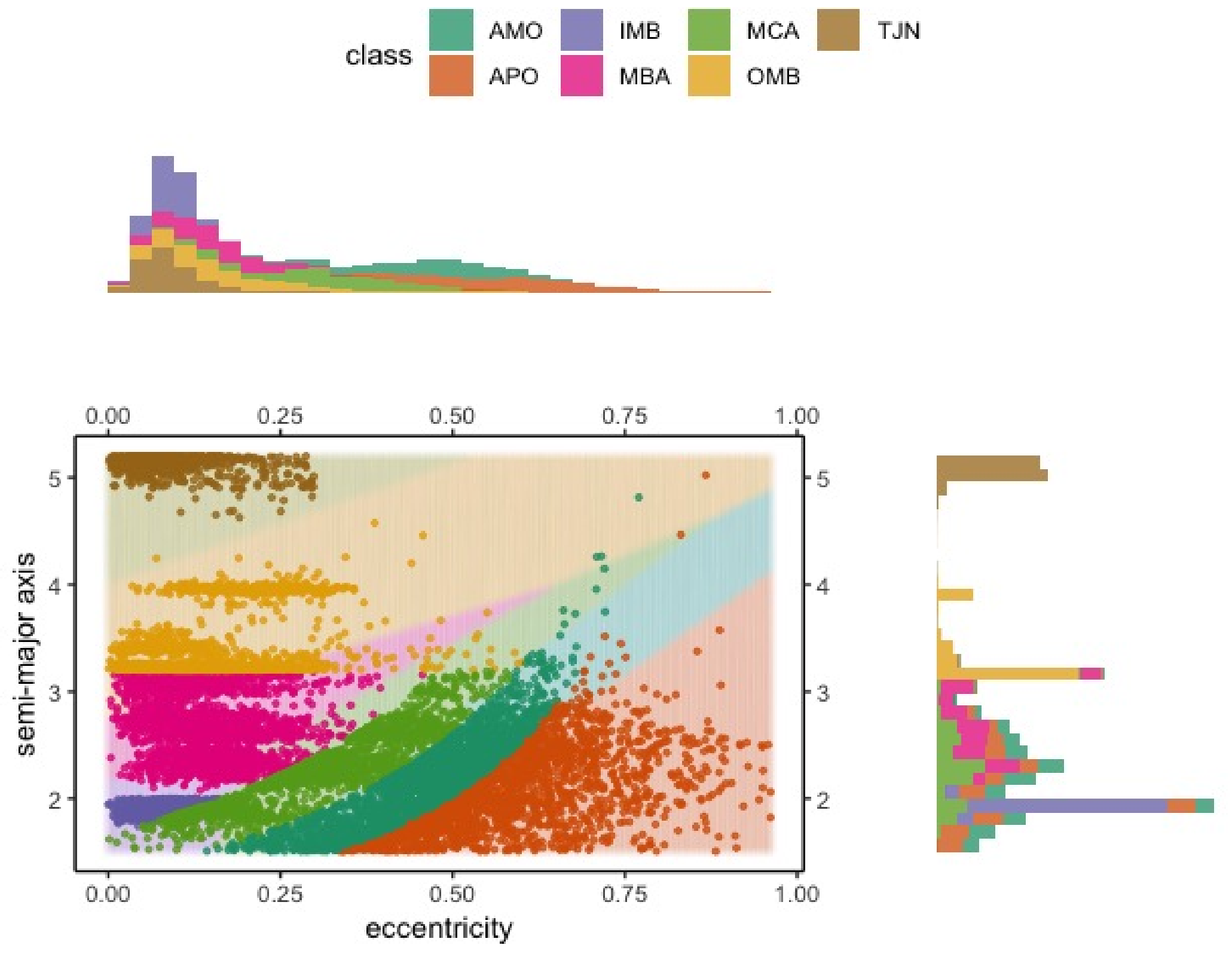

Figure 3: LDA Classification of Orbital Types (n = 2,000)

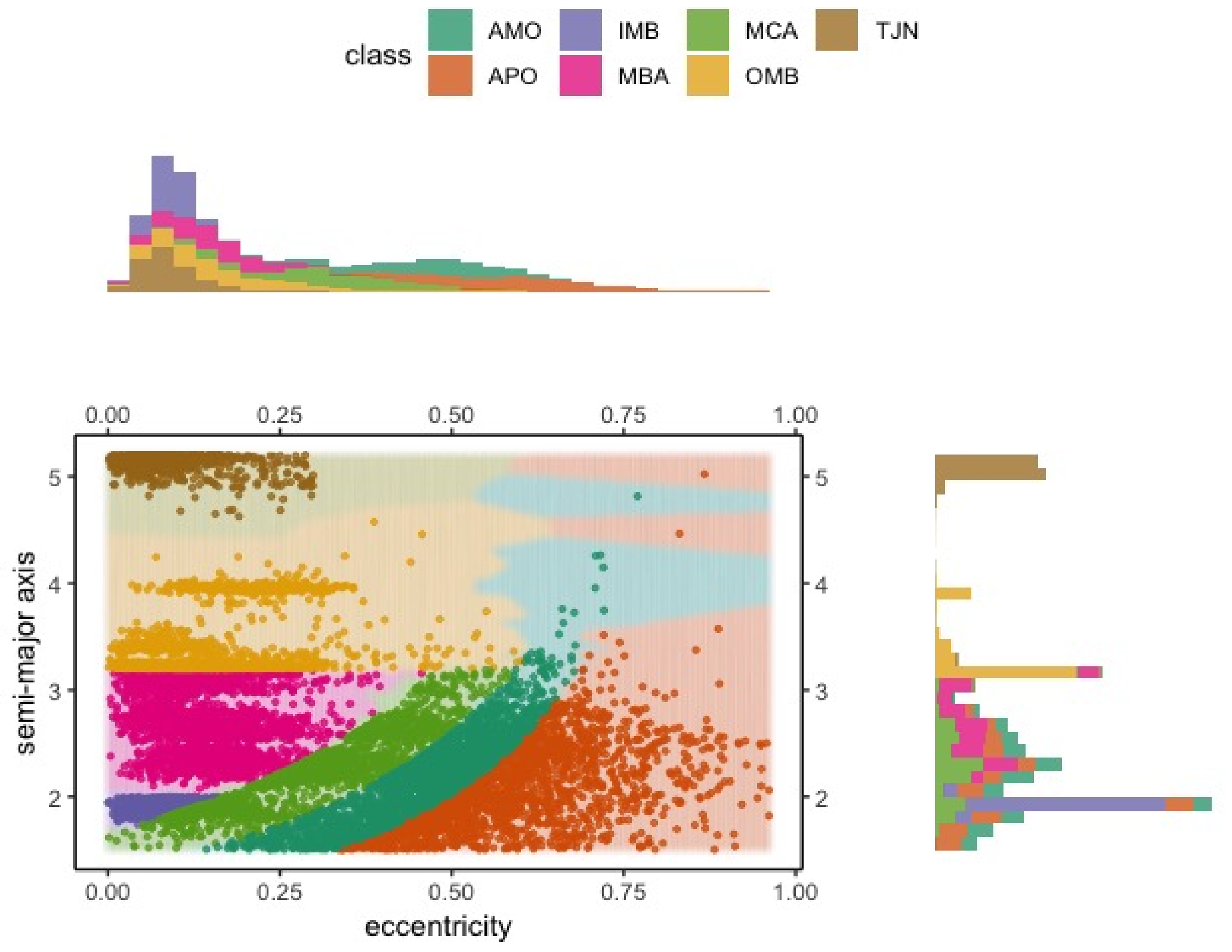

Figure 4: KNN Classification of Orbital Types (n = 2,000)

Although orbital types are not defined explicitly in terms of semi-major axis and eccentricity, it seems that these features characterize the orbits very well (the points are separated into distinct clusters).

Result

Algorithm	In-sample (%)	Out-of-sample (%)
SVM	95.312	86.732
KNN	100	95.527
LDA	91.312	73.973
QDA	94.531	85.036
GLR	96.031	88.384

Table 2: Predictive accuracy of algorithms for n = 1,400

Algorithm	In-sample (%)	Out-of-sample (%)
SVM	97.293	97.046
KNN	100	98.854
LDA	85.664	85.246
QDA	92.414	92.346
GLR	93.857	94.286

Table 3: Predictive accuracy of algorithms for n = 14,000

KNN produces the best performance, followed closely by SVM and GLR. LDA has the lowest predictive accuracy on both the test set and the training set. This is expected because LDA assumes linear decision boundaries, which clearly do not hold for this particular problem. We can tune the KNN model (by finding the optimal number of nearest neighbors and the best distance metric) to achieve better result.

We compare the two best models statistically using the Stuart-Maxwell test. The test statistic follows a χ^2 distribution with $K - 1$ degrees of freedom, where K is the number of classes (i.e. df = 6 in this case).

- n = 1,400, KNN / GLR: $\chi^2 = 374.81$, p-value $< 2.2 \times 10^{-16}$
- n = 14,000, SVM / KNN: $\chi^2 = 276.03$, p-value $< 2.2 \times 10^{-16}$

We conclude that KNN is the best model for this problem.

Machine Learning: PHA Asteroids

A potentially hazardous object is a near-Earth object with an orbit that can make close approaches to the Earth and large enough to cause significant regional damage in the event of impact. Most of these objects are potentially hazardous asteroids (PHAs), and a few are comets. As of October 2019, there are 2,018 known PHAs (about 10% of the total near-Earth population), of which 156 are estimated to be larger than one kilometer in diameter. Most of the discovered PHAs are Apollo asteroids (1,601) and fewer belong to the group of Aten asteroids (169).

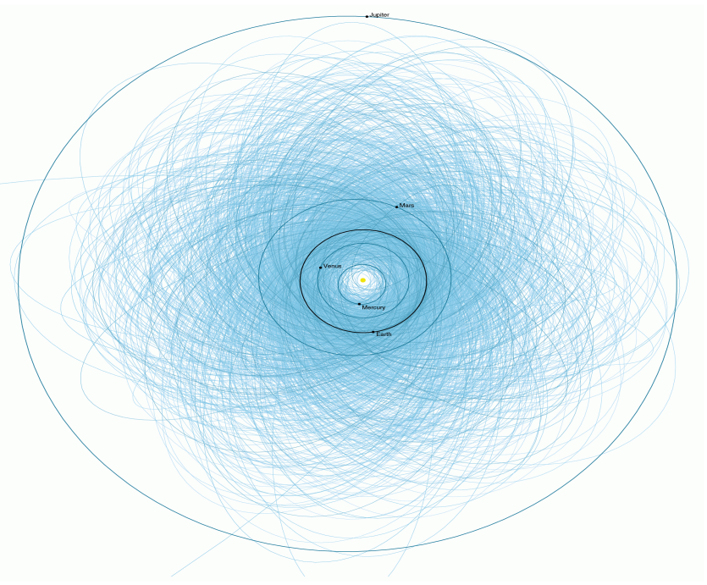

Figure 5: Orbits of over 1,400 known PHAs by the Jet Propulsion Lab at NASA

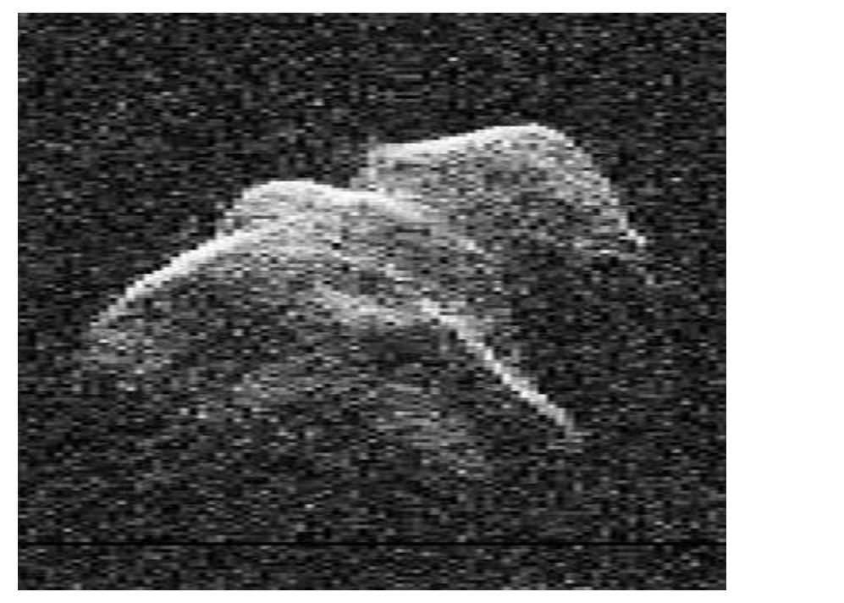

Figure 6: Radar image of the Toutatis asteroid by GDSCC in 1996

In this section, the aim is to predict if an asteroid is hazardous based on its absolute magnitude (the luminosity of a celestial object) and its minimum orbital intersection distance with the Earth.

- Again, due to computational limitation, we subset the full data set randomly to obtain a training set (20,000 observations) and a test set (30,000 observations).
- We train the following classification models:
  - Support Vector Machine (SVM)
  - K-Nearest Neighbors (KNN)
  - Logistic Regression
- Apply the models on the test data set to get the out-of-sample performance.
- Since the data set is highly imbalanced, it makes sense to determine the significance of the models relative to the all-in strategy (labelling everything as non-hazardous). To do this, we use the McNemar's test.

Result

Algorithm	Out-of-sample (%)
SVM	99.733
KNN	99.810
Logit	99.883
All-in	99.733

Table 4: Predictive accuracy of algorithms

Since the data is very imbalanced (most asteroids are non-hazardous), the all-in strategy yields an accuracy of 99.733% on the test set (this should be the baseline to evaluate other models). Logistic regression is the best method in this case with an out-of-sample accuracy of 99.883%.

Is the logistic regression model significantly better?

To answer this question, I use the McNemar's test. The test statistic follows a χ^2 distribution with 1 degree of freedom.

- Logit / All-in: $\chi^2 = 65$, p-value $= 7.5 \times 10^{-16}$

We conclude that the logistic regression model outperforms the all-in strategy (at reasonable significance level).

References

- https://ssd.jpl.nasa.gov/sbdb_query.cgi#x
- http://www.minorplanetcenter.org/iau/lists/Atens.html
- http://www.minorplanetcenter.org/iau/lists/Apollos.html
- http://www.minorplanetcenter.org/iau/lists/Amors.html
- "Earth-Approaching Asteroids as Targets for Exploration," E. Shoemaker and E. Helin; in Asteroids: A Exploration Assessment, pp. 245-256.
- https://en.wikipedia.org/wiki/Potentially_hazardous_object