NASA : Asteroid Classification

Vidit Goyal

Machine Learning Engineer Intern
AI Technology and Systems
www.ai-techsystems.com
vidit2011998@gmail.com

Abstract: Asteroids are rocky-metallic things which range in size from approximately the size of stones to about 600 miles (~1,000 km) across. Most of the undiscovered asteroids are the smaller ones (less than 100 km across) which are harder to detect. It is expected that there are over a million of these smaller asteroids.

Using The dataset from NeoWs. NeoWs [1] (Near Earth Object Web Service) that is a RESTful web service for near earth Asteroid information, the authors were able to classify the asteroid as hazardous or Non-Hazardous with an accuracy of more than 99% using Machine Learning.

Keywords : RESTful, Asteroid, Machine Learning.

1. INTRODUCTION

There is growing concern in identifying asteroids whose orbits cross Earth's, and that could, given enough time, collide with Earth . The most important clubs of near-Earth asteroids are the Apollos, Amors, and Atens. Several asteroid deflection strategies have been introduced, as early as the 1960s. [2]

The smallest asteroids found (based on **absolute magnitude** *H*) are 2008 TS26 with *H* = 33.2 and 2011 CQ1 with *H* = 32.1 both with an approximated size of about 1 meter. [3].

The bulk of identified asteroids orbit within the asteroid belt within the orbits of Mars and Jupiter, usually in relatively **low-eccentricity** (i.e. not very elongated) orbits. This belt is now measured to contain between 1.1 and 1.9 million asteroids greater than 1 km (0.6 mi) in diameter,[56] and millions of smaller ones. [4]

Near-Earth asteroids, or NEAs, are asteroids that hold orbits that pass near to that of Earth. Asteroids that cross Earth's orbital path are called as *Earth-crossers*. As of June 2016, 14,464 near-Earth asteroids are identified and the number over one kilometer in diameter is expected to be 900–1,000. [5]

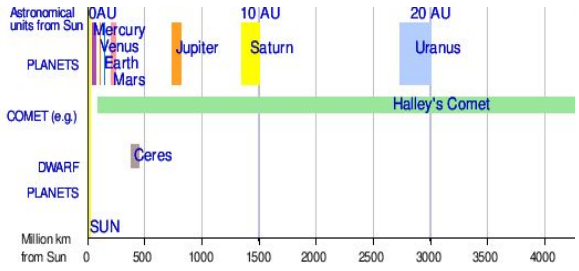


Fig 1: Distances of chosen bodies of the Solar System from the Sun. The left and right ends of each bar correspond to the perihelion and aphelion of the body, respectively, hence long bars indicate high orbital eccentricity. The radius of the Sun is 0.7 million km, and the radius of Jupiter (the biggest planet) is 0.07 million km, both too small to decide on this image. [5]

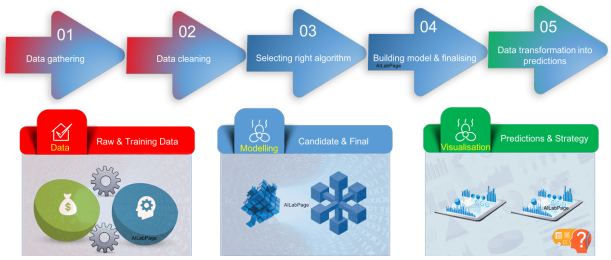
Table 1 examines the related research work that has been done on the similar data. I have identified the various salient features and scope for improvement in these kernels and tried to improve upon these in my research work.

Table 1 : Background study

Kernels studied	Salient features	Scope for improvement
[6]	Data cleaning is done. Data distribution is very well defined.	Only bell curves are shown. Dist-plots and box-plots are not used.
[7]	XGB classifier is used that has an accuracy of 99.68%.	Accuracy can be further increased and notebook is not explained properly.

Figure 2 shows the approach adopted to classify an asteroid as Hazardous or Non-Hazardous. The three Machine learning models used in this paper are Decision Tree, XGBoost and Gradient Boosting Classifier.

Figure 2 : Approach Diagram



2. APPROACH AND IMPLEMENTATION

The dataset used in this paper were obtained from the NASA Website [8]. This API is maintained by SpaceRocks Team: David Greenfield, Arezu Sarvestani, Jason English and Peter Baunach.
Initially, the data had 40 attributes with 4687 samples. It consists of integer, float and object values. The object values were Close Approach Date, Orbiting Body ,Orbit Determination Date and Equinox. These were not required as some of them were single valued attributes and others were date attributes. Therefore, these values were dropped.
Secondly, the average value of attributes Est Diameter in km (min) and Est Diameter in km (max) were taken and inserted into the table with column name avg_diameter. All the other values i.e. Neo Reference ID, Name, Close Approach Date, Orbit ID, Orbit Determination Date, Est Dia in KM(min), Est Dia in KM(max), Est Dia in M(min), Est Dia in M(max), Est Dia in Miles(min), Est Dia in Miles(max), Est Dia in Feet(min), Est Dia in Feet(max), Epoch Date Close Approach, Relative Velocity km per sec, Miles per hour, Miss Dist.(Astronomical), Miss Dist.(lunar) ,Miss Dist.(miles) were dropped as they were not necessary.

Secondly, for data analysis, heatmap was made that showed the correlation between the attributes. Figure 3 shows a heatmap showing correlation between the attributes of the data. Strong correlation is shown by dark colour whereas weak correlation is shown by lighter colours.
A countplot and pi chart was made to analyse the Hazardous and Non-Hazardous asteroids. It can be seen from Figure 4 that The total number of asteroids which are not hazardous are approx 3900 and hazardous are 700.

Figure 3 : Heatmap

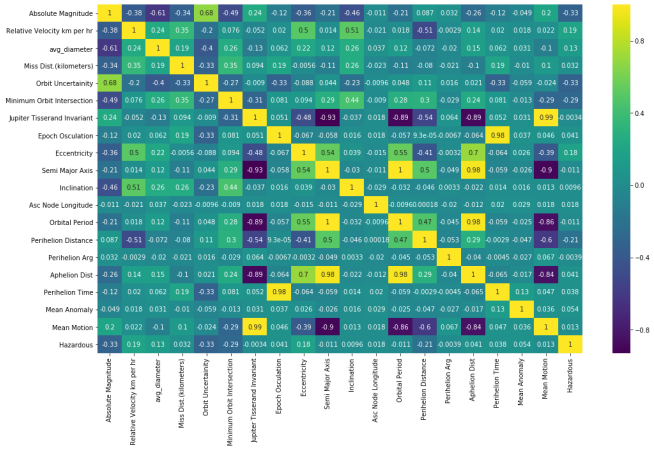
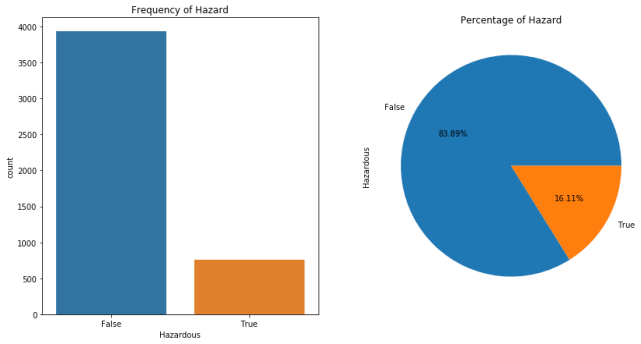


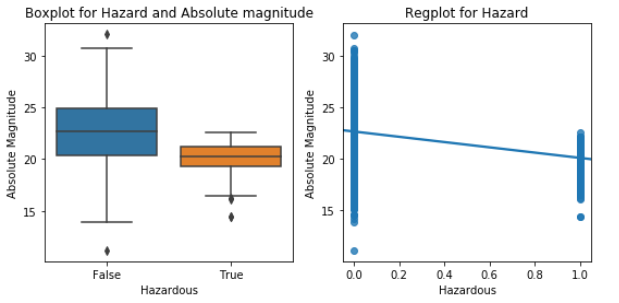
Figure 4 : Countplot and Pi-chart for target value column

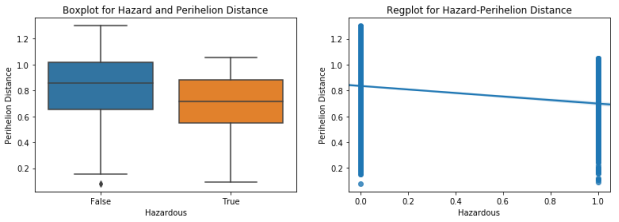


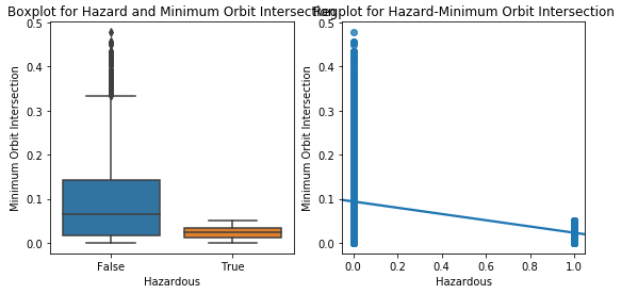
Box plots and regplots were plotted for different values that showed high correlation i.e. between Absolute Magnitude and target value, Perihelion distance and target value, Minimum Orbit Intersection and target value, Orbital Period and target value, Mean Motion and target value. Figure 5 shows the boxplots and regplots for different attributes. After this, bell-curves were plotted for the values Absolute Magnitude, Relative Velocity km per hr, Miss Dist.(kilometers), Eccentricity, avg_diameter, Orbit Uncertainty, Minimum Orbit Intersection, Epoch Osculation.

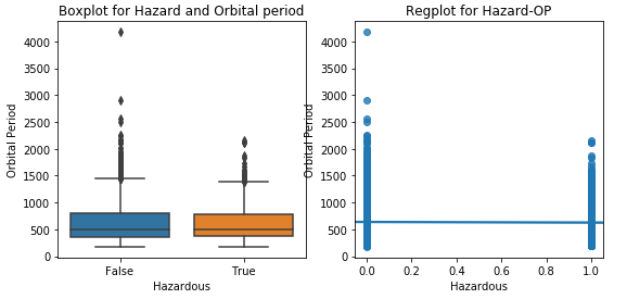
Finally 20 columns were left after the pre-processing and the model was trained on these value using Decision Tree, XGBoost and Gradient Boosting Classifier. Figure 6 shows the bell curves.

Figure 5: Various boxplots and reg plots









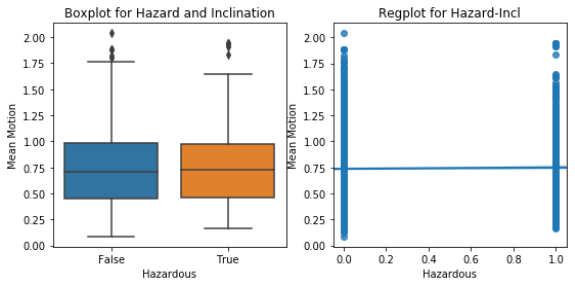


Figure 6: Various bell curves

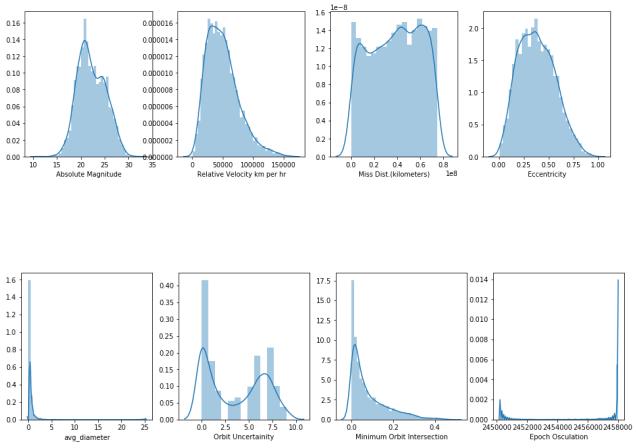
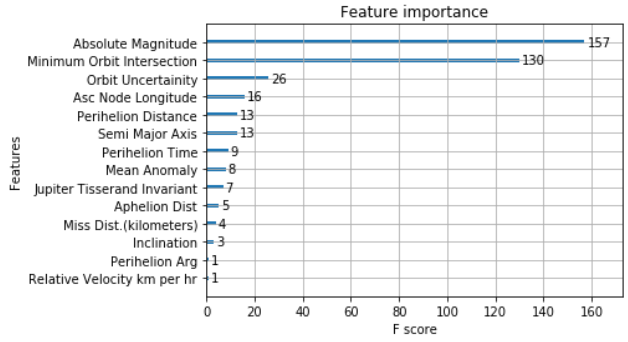
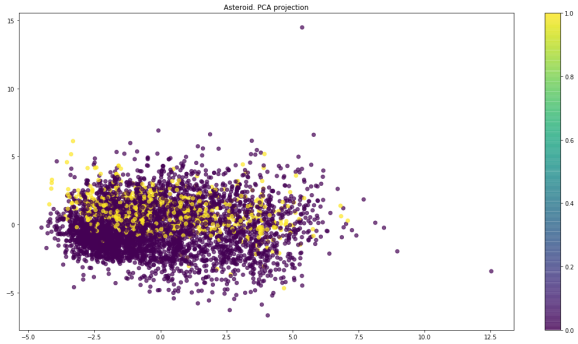


Figure 7: Feature Importance graph



Principal Component Analysis: Principal component analysis helps make data easier to explore and visualize. It is a simple non-parametric technique for extracting information from complex and confusing data sets. Principal component analysis is focused on the maximum variance amount with the fewest number of principal components. Below is a 19-dimensional data in 2D.



3. RESULTS

From data analysis, it can be seen that Most asteroids of Absolute Magnitude 20-25 are Not Hazardous whereas of 19-21 are Hazardous. Non hazardous can be as low as 14 and as high as 32 in Magnitude Hazardous can be as low as 16 and as high as 22 in Magnitude.
Most asteroids having a Minimum Orbit Intersection 0-0.11 are Not Hazardous whereas of 0-0.02 are Hazardous. Non hazardous can be as low as 0 and as high as 0.5 in Minimum Orbit Intersection. Hazardous can be as low as 0 and as high as 0.02 in Minimum Orbit Intersection.
Most asteroids having Orbital Period 450-900 are Not Hazardous whereas of 400-800 are Hazardous. Non hazardous can be as low as 0 and as high as 400 in Orbital Period. Hazardous can be as low as 0 and as high as 2000 in Orbital Period

Bell curves for Absolute Magnitude, Relative Velocity km per hr, Miss Dist.(kilometers), Eccentricity shows that these attributes have bell-shaped curve that suggests it is normal.
Bell curves for avg_diameter, Orbit Uncertainty, Minimum Orbit Intersection, Epoch Osculation shows that some are right skewed i.e. not normal and some are are

bi-modal i.e. 2 independent sources of variation.
Figure 7 shows the feature importance graph of the attributes where it can be seen that Absolute Magnitude and Minimum Orbit Intersection are highly important features.

Finally in model training, the results were as follows:

1. **Prediction with XGBoost:**

   Table 2: Accuracy and f1 score table

Accuracy score	0.99488054607
F1 score	0.99490318756

2. Prediction with Gradient Boosting:

Table 3: Training and Validation accuracy at different learning rates

	Learning rate	Train score	Validation score
1.	0.05	0.833	0.849
2.	0.1	0.853	0.865
3.	0.25	0.983	0.978
4.	0.5	0.998	0.995
5.	0.75	0.999	0.997
6.	1	0.999	0.997

Highest train accuracy was 0.999 and validation score was 0.997 at a learning rate of 0.75 and 1.

Table 4 : Confusion Matrix for Gradient bosting

True negatives (TN)	740
False positives (FP)	0
False Negatives(FN)	4
True Positives (TP)	129

Table 5 : Precision recall table

	precision	recall	f1-score	support
0	0.99	1.00	1.00	746
1	1.00	0.97	0.98	133

3. Prediction with Decision tree Classifier:

Table 5 : Training and Validation accuracy for decision tree classifier

Training accuracy	0.9994
Validation accuracy	0.9981

Table 6: Confusion Matrix for Decision tree

True negatives (TN)	990
False positives (FP)	0
False Negatives(FN)	2
True Positives (TP)	180

Table 7: Precision recall table

	precision	recall	f1-score	support
0	0.99	1.00	1.00	990

1	1.00	0.97	0.98	182

4. CONCLUSION

Out of the 3 classifiers, XGB classifier has the highest Training accuracy of 0.9994 and also the highest f1 score of 0.9949. Therefore it is the best performing model for classifying asteroid as Hazardous or Non-Hazardous.

5. REFERENCES

1. https://en.wikipedia.org/wiki/Standard_asteroid_physical_characteristics
2. http://www.physics.sfasu.edu/astro/asteroids/sizemagnitude.html
3. https://spaceplace.nasa.gov/asteroid/en/
4. https://en.wikipedia.org/wiki/Standard_asteroid_physical_characteristics
5. http://coolcosmos.ipac.caltech.edu/ask/184-What-are-asteroids
6. https://www.kaggle.com/shrutimehta/data-preprocessing-and-correlation
7. https://www.kaggle.com/jav1d98/getting-99-68-accuracy-with-xgbclassifier-model
8. https://www.kaggle.com/shrutimehta/nasa-asteroids-classification/downloads/nasa-asteroids-classification.zip/1