Potentially Hazardous Asteroids & Diameter Prediction

Edoardo Di Paolo, 1728334
Big Data, A.Y. 2020-2021

01

Introduction

A brief introduction to the topic and why it is important.

02

Dataset

Everything you need to know about the dataset used in the project.

03

Feature Engineering

How the dataset was modified to be used in the project.

04

Machine Learning Models

The machine learning models applied for the tasks.

05

Results

An overview about the results obtained in the tasks.

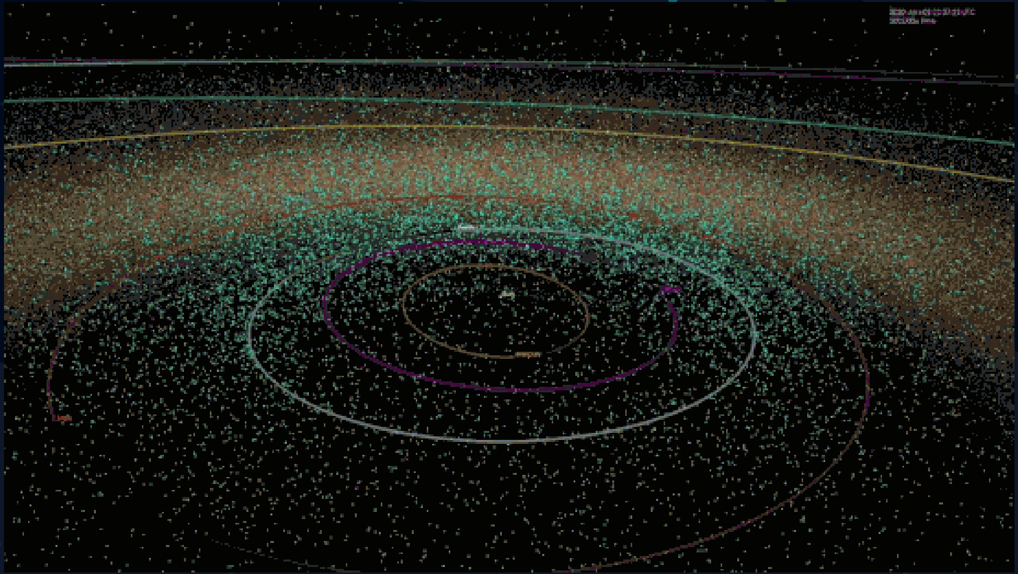
06

Conclusions & future works

Conclusions and possible future works on those topics.

What are PHAs?

- Potentially Hazardous Asteroids.
- Their orbits can make close approaches to the Earth.
- They are large enough to cause significant regional damage in the event of impact.
- NASA astronomers reported that 5 to 10 years of preparation may be needed to avoid a potential impactor.



Known Near Earth Objects in 2018

Dataset

~ 450 MB
of data

+ 900k
objects in the dataset

45
different columns

+ 20k
NEOs, near Earth objects

+ 2k
PHAs

Feature Engineering

- I removed useless columns and columns with multiple missing values.

- In particular, for the PHAs classification task, I removed these columns: *id, name, prefix, spkid, full_name, diameter, albedo, diameter_sigma, pdes and equinox.*

- I added a column, called *Q_ph*: it is the *aphelion distance* ($Q = a(1 - \mathrm{e})$), where a is the semi-major axis and e is the eccentricity.

- For the *diameter prediction* task, I used a subdataset with a valid value in the diameter column (~ 135k samples).

Feature Engineering

- In the PHAs classification task, the original dataset is **highly** unbalanced.
- There are about 2066 PHAs and 930269 not PHAs.
- I oversampled the dataset, with the *sample* method provided by Spark, and I used also a dataset with only the NEOs that contains about 22k samples.
- The oversampling is performed on the training dataset.
- I used the 70% of the dataset to train the model and the remaining 30% to test the model.
- I didn’t use cross validation for this task due to limits of Databricks.

Machine Learning Models

PHAs classification task

- ❖ Logistic regression
- ❖ Decision tree
- ❖ Random forest

Asteroids diameter prediction

- ❖ Linear regression
- ❖ Random forest regressor
- ❖ Gradient-boosted tree

Results

Logistic Regression

All the dataset

- Best results on the dataset not oversampled without the scaler.
- Accuracy: ***0.99***
- Precision: ***0.70***
- Recall: ***0.58***
- F1-Score: ***0.64***

Only dataset with NEOs

- Best results on the dataset not oversampled and with the scaler.
- Accuracy: ***0.95***
- Precision: ***0.85***
- Recall: ***0.86***
- F1-Score: ***0.86***

Results

Decision tree

All the dataset

- Best results on the dataset oversampled without the scaler.
- Accuracy: ***0.99***
- Precision: ***0.61***
- Recall: ***0.99***
- F1-Score: ***0.75***

Only dataset with NEOs

- Best results on the dataset not oversampled and without the scaler.
- Accuracy: ***0.99***
- Precision: ***0.98***
- Recall: ***0.98***
- F1-Score: ***0.98***

Results

Random forest

All the dataset

- Best results on the dataset oversampled without the scaler.
- Accuracy: ***0.97***
- Precision: ***0.54***
- Recall: ***0.98***
- F1-Score: ***0.70***

Only dataset with NEOs

- Best results on the dataset oversampled and with the scaler.
- Accuracy: ***0.91***
- Precision: ***0.74***
- Recall: ***0.90***
- F1-Score: ***0.81***

Results

Linear regression, Random forest regressor and Gradient boosted tree

Model	Train set	Test set
Linear regression	5.832	6.58
Random forest regressor	3.32	6.3
Gradient boosted tree	3.198	7.0

Results are expressed in the RMSE metrics, lower is better.

Conclusions & future works

- I trained 6 different models on the two tasks
- Obtained good results
- Add more models like SVM and Naïve Bayes
- Use the cross validation on the classification task
- Use SMOTE to oversample the datasets