Under Attack - Binary Classification of Hazardous Objects from Space

DSE511 - Introduction to Data Science and Computing I

Katie Knight, Anna-Maria Nau, and Christoph Metzner

Introduction

- **Edmond Halley** first to voice concerns about the impact of extraterrestrial objects (18th century)
- **Alvarez et al.** published hypothesis that such events caused several mass extinctions

- US Congress ordered NASA to develop mitigation strategies
- One Strategy: Detection of such **N**ear **E**arth **O**bjects (NEOs) utilizing statistical models, i.e., machine learning based on properties (e.g., shape, size, velocity)

Previous Work concerning Risk Assessment of NEOs

- Using scales such as *Torino Scale* or *Palermo Scale* → easily interpretable by laymen
- VNIR spectroscopy to characterize compositional and physical structure of NEOs
- Applying neural networks to analyze the energy deposition curves

Research Hypothesis and Questions

Research Hypothesis

"We hypothesize that using machine learning algorithms can play a vital part in the detection of hazardous objects."

Research Question

"Can standard 'off-the-shelf' machine learning algorithms achieve outstanding performance classifying such objects?"

Data

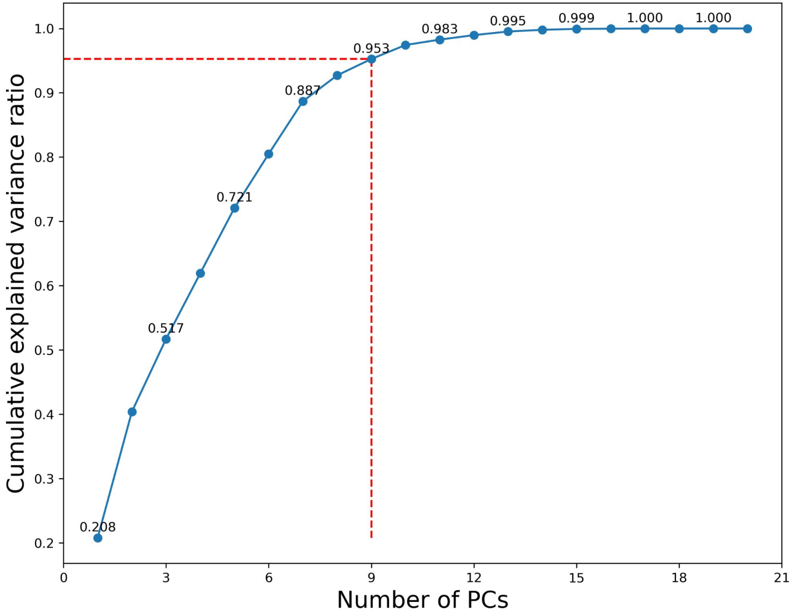
- Data source: NASA API called NeoWS (Near Earth Object Web Service) and available on Kaggle.
- Dataset contains 4687 rows and 40 columns/features
  - 39 input features
  - 1 binary target feature (hazardous or nonhazardous)
- Input features include typically measured information of an asteroid such as absolute magnitude, estimated diameter, relative velocity, distance measures from the sun, or the shape of the objects orbit known as eccentricity.
  - No missing values, a lot of redundancy

Methods - Data Preprocessing

The following data preprocessing steps were performed:

- Feature names were converted to snake case (e.g., Absolute Magnitude -> absolute_magnitude)
- Redundant and time related features were removed (39 -> 20 input features)
- 80:20 train/test split
  - 3749 train samples
  - 938 test samples
- Min-max normalization to transform all features in the range [0, 1]
- Principal component analysis (PCA) for dimensionality reduction
  - Total explained variance ratio of 0.95 result in 9 principal components (20 - 9 input features)

Methods - Data Preprocessing Continued…



Methods - Classification

- Algorithms
  - Naive Bayes (Baseline)
  - Support Vector Machine
  - Decision Tree
  - Random Forest
- Hyperparameter tuning via sklearn's GridSearchCV (k=5) function on training data
- Model Evaluation on testing data
  - F1-Score
  - Recall
  - Precision

Results: Model Performance

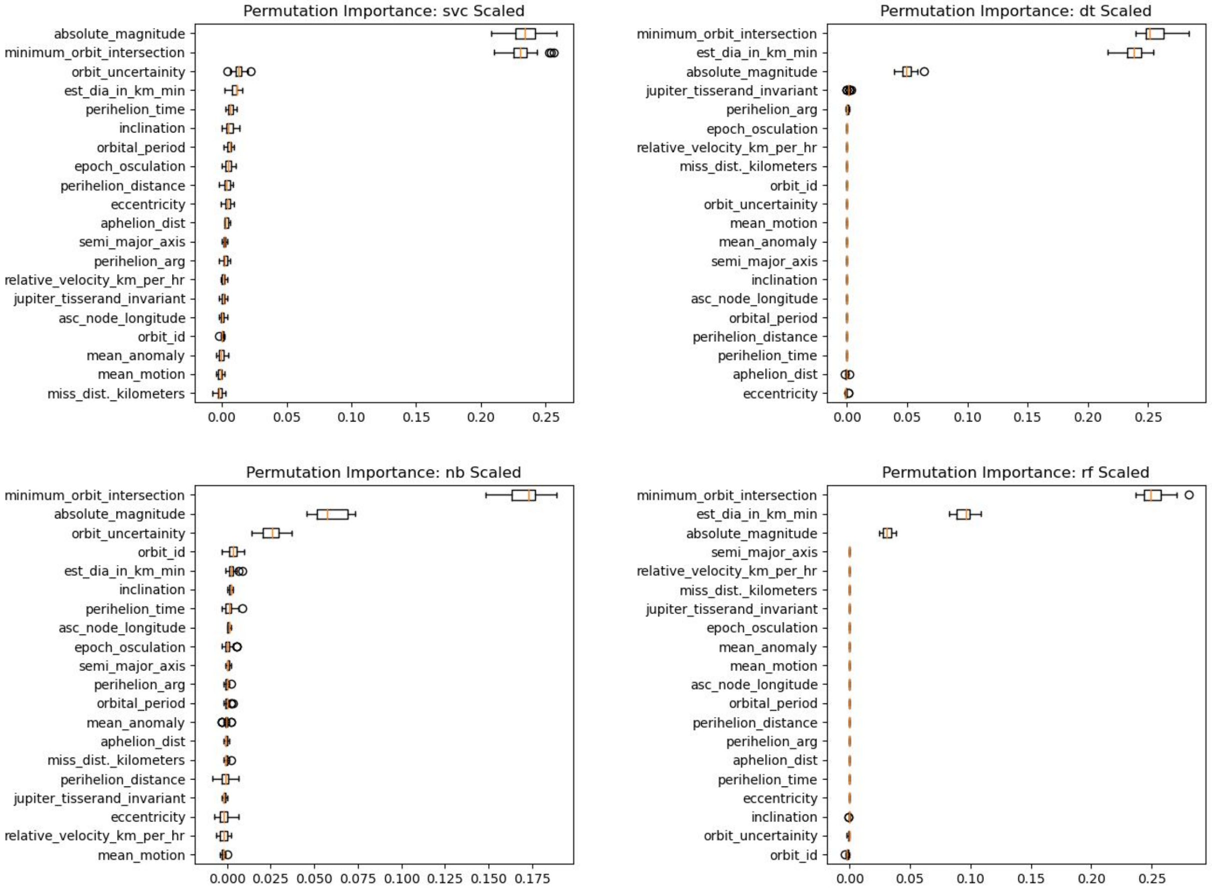
- Best Models: Decision Tree and Random Forest
- Decision Tree very high performance given the very fast computation
- Preprocessing of data matters → *min-max scaled* outperforms *PCA*

Algorithm	Data	F1-Score	Recall	Precision	Time [s]
NB	Scaled	0.8464	0.8552	0.8378	0.0160
NB	PCA	0.515	0.4138	0.6818	**0.0104**
SVM	Scaled	0.8865	0.8621	0.9124	2.544
SVM	PCA	0.5959	0.5034	0.73	5.376
DT	Scaled	**0.979**	**0.9655**	**0.9929**	0.02973
DT	PCA	0.4758	0.4069	0.5728	0.03461
RF	Scaled	0.9718	0.9517	0.9928	3.7710
RF	PCA	0.5641	0.4552	0.7416	5.4340

Results: Confusion Matrix

Algorithm	Data	TP	FP	FN	TN
NB	Scaled	769	24	21	124
NB	PCA	765	28	85	60
SVM	Scaled	781	12	20	125
SVM	PCA	766	27	72	73
DT	Scaled	792	1	5	140
DT	PCA	749	44	86	59
RF	Scaled	792	1	7	138
RF	PCA	770	23	79	66

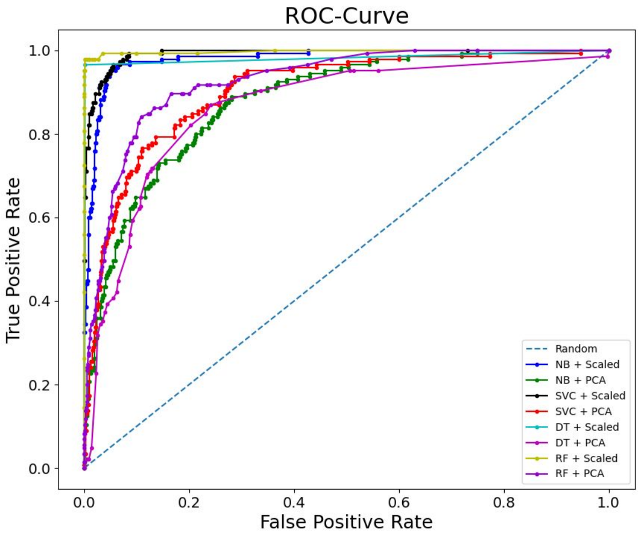
Results: Permutation Importance -Katie



- Minimum orbit intersection significant for all algorithms;
- SVC favored absolute magnitude slightly more
- SVC and NB favor similar features; DT and RF favor similar features

Results - Receiver Operating Characteristics Curve - Katie

- Models with scaled data better than with PCA data
- Best Performances
    - (1) Random Forest
    - (2) Decision tree
- Naive Bayes (baseline) has equal performance as decision tree



Conclusion - Katie

- We demonstrated that "off-the-shelf" machine learning algorithms can be used to identify the hazardous nature of extraterrestrial objects.
- Based on the performance metrics, the decision tree slightly outperforms the random forest algorithm.
- In cases that requires to have a flexible threshold for binary classification, one may want to use the random forest.
- The use of principal component analysis to accomplish feature reduction reduced the performance of all algorithms, and is therefore not recommended as a preprocessing step for this prediction task.

References - Katie

1] Luis W Alvarez, Walter Alvarez, Frank Asaro, and Helen V Michel. 1980. Extraterrestrial cause for the Cretaceous-Tertiary extinction. Science 208, 4448 (1980), 1095–1108.

[2] Space Studies Board, National Research Council, et al. 2010. Defending planet earth: Near-Earth-Object surveys and hazard mitigation strategies. National Academies Press.

[3] William F Bottke Jr, Alessandro Morbidelli, Robert Jedicke, Jean-Marc Petit, Harold F Levison, Patrick Michel, and Travis S Metcalfe. 2002. Debiased orbital and absolute magnitude distribution of the near-Earth objects. Icarus 156, 2 (2002), 399–433.

[4] L. Breiman. 2001. Random Forests. Machine Learning 45, 1 (2001), 5–32.

[5] Ronan Collobert and Samy Bengio. 2004. Links between perceptrons, MLPs and SVMs. In Proceedings of the twenty-first international conference on Machine learning. ACM, 23.

[6] Jerome Fan, Suneel Upadhye, and Andrew Worster. 2006. Understanding receiver operating characteristic (ROC) curves. Canadian Journal of Emergency Medicine 8, 1 (2006), 19–20.

[7] Michael J Gaffey, Paul A Abell, and Paul S Hardersen. 2006. COMPOSITIONAL AND PHYSICAL CHARACTERIZATIONS OF NEOs FROM VNIR SPECTROSCOPY. (2006).

[8] Sarah Greenstreet, Henry Ngo, and Brett Gladman. 2012. The orbital distribution of near-Earth objects inside Earth's orbit. Icarus 217, 1 (2012), 355–366.

[9] José M Hedo, Manuel Ruíz, and Jesús Peláez. 2018. On the minimum orbital intersection distance computation: a new effective method. Monthly Notices of the Royal Astronomical Society 479, 3 (2018), 3288–3299.

[10] R. Olshen L. Breiman, J. Friedman and C. Stone. 1984. Classification and Regression Trees. Wadsworth.

References - Katie

[11] Amy Mainzer, T Grav, J Bauer, J Masiero, RS McMillan, RM Cutri, R Walker, E Wright, P Eisenhardt, DJ Tholen, et al. 2011. NEOWISE observations of near-Earth objects: preliminary results. The Astrophysical Journal 743, 2 (2011), 156.

[12] Kevin P Murphy et al. 2006. Naive bayes classifiers. University of British Columbia 18, 60 (2006), 1–8.

[13]NASA. [n. d.]. Near-Earth Object Observations Program. https://www.nasa.gov/planetarydefense/neoo.

[14] NASA. [n. d.]. Torino Impact Scale. https://web.archive.org/web/20070224184143/ http://impact.arc.nasa.gov/torino.cfm.

[15] Carrie R Nugent, John Dailey, Roc M Cutri, Frank J Masci, and Amy K Mainzer. 2017. Machine learning and next-generation asteroid surveys. In AAS/Division for Planetary Sciences Meeting Abstracts# 49, Vol. 49.

[16] Alfred Romer. 1984. Halley's comet. The Physics Teacher 22, 8 (1984), 488–493.

[17] A. M. Tarano, J. Gee, L. Wheeler, S. Close, and D. Mathias. 2020. Automating the Inference of Asteroid Physical Properties and Motion. In AGU Fall Meeting Abstracts, Vol. 2020. Article P008-02, P008-02 pages.

[18] Vladimir Vapnik. 2013. The nature of statistical learning theory. Springer science & business media.