Predicting Assessment of Near-Earth Comets' Hazard Potential Using Machine Learning

Peter Santana
Case Studies in Machine Learning
AI395T
University of Texas - Austin

Table of Contents

Abstract

Classifying near earth objects as hazardous to Earth or not, specially for comets, is a critical task in planetary defense. These celestial bodies have the potential to intersect Earth's orbit and cause significant impact events and great damage. Traditionally, the Minimum Orbit Intersection Distance (MOID) is used to assess hazard levels, but it is often unavailable for newly discovered and observed comets, leading to delays in assessing how hazardous they are. In this study, we will be implementing a machine learning-based approach to classify potentially hazardous comets without relying on MOID with the use of alternative orbital parameters. By training multiple machine learning models (Logistic Regression, Support Vector Machine, and Random Forest) on a dataset of comet orbital elements, we optimized a Random Forest model that achieved high classification accuracy. This paper details the data preprocessing, feature engineering, model selection, and evaluation processes. Our results demonstrate that the Random Forest model, using optimized hyper parameters, can accurately classify potentially hazardous comets with an AUC score of 0.91. This MOID-independent model provides a rapid and effective solution for hazard classification.

1. Introduction

1.1 Background

Near-Earth objects (NEOs), are celestial bodies whose orbimsbring them close to Earth's orbit, this includes comets and asteroids (National Research Council, 2010). In the past recent years, monitoring and analyzing these objects has gained traction due to them being a potential threat to Earth and humans (Harris & D'Abramo, 2015). The impact of an NEO can have catastrophic consequences, going from as small as regional destruction to having the potential of causing irreparable global climate effects (Chapman, 2004). If we go back in history, we can see events such as the Tunguska event in 1908, where an NEO exploded over Siberia (Gladman et al., 1997) or The Chelyabinsk meteor in 2013 (Brown et al., 2013). These events highlight the importance of early detection and classification of NEOs.

Advancements in telescope technology and sky surveys have led to the discovery of a growing number of NEOS (Jedicke et al., 2015). Space agencies and organizations such as NASA and the European Space Agency (ESA) have initiated programs and and divisions for detecting, tracking, and cataloging NEOs (Mainzer et al., 2011). The data collected from these efforts are critical for knowing the risk associated with each object and developing potential mitigation strategies (Harris et al., 2015).

1.2 Importance of Hazard Assessment

The classification of NEOs into hazardous and non-hazardous categories is a vital component of planetary defense (Reddy et al., 2015). Accurate and timely hazard assessment allows to prioritize monitoring resources and informs decisions regarding possible deflection missions or evacuation plans (Johnson et al. 2015). The need for efficient and effective classification methods has increased parallel to the number of detected NEOs (Valsecchi & Gronchi, 2015).

Current hazard assessment relies heavily on calculating the Minimum Orbit Intersection Distance (MOID), which indicates the smallest possible distance between the object's orbit and Earth's (Sitarski, 1968). Objects that have a MOID less than 0.05 astronomical units (AU) are typically considered potentially hazardous (Chesley & Milani, 1999). However, calculating the MOID requires precise orbital data and complex computations, which may not always be readily available for newly observed or discovered NEOs (Giorgini et al., 2008).

1.3 Challenges with Current Methods

While MOID is a highly valuable asset to determine the risk of collision, it does come with some current limitations:

- **Data Availability:** MOID calculations require precise orbital data, which may be missing or incomplete for newly observed or discovered comets (Mueller et al., 2007). This lack of data can delay the assessment of a comet, potentially leaving Earth vulnerable if the comet is found to be hazardous during the early observation period.
- **Computational Complexity:** Determining MOID involves iterative calculations in orbital mechanics, making it computationally demanding, especially for large datasets of comets or if done in real time applications (Farnocchia et al., 2015). The computational resources required can limit the ability to process data from numerous NEOs quickly.
- **Orbital Dynamics of Comets:** Comets present unique challenges due to their highly eccentric and inclined orbits. They are also subject to non gravitational forces such as outgassing when near the Sun, which can alter their trajectories unpredictably (Yeomans et al., 2004). These factors complicate MOID calculations and hazard assessments for comets compared to asteroids.

Given these challenges, there is a clear need for alternative methods that can provide rapid and reliable hazard assessments without relying only on MOID. Machine learning offers a promising solution for development of methods to predict hazardous NEOs (Michel et al., 2018).

2. Problem Definition

2.1 Limitations of MOID

The reliance on MOID as the primary metric for hazard assessment has several pushbacks:

- **Delayed Hazard Classification:** For newly discovered comets, it can take time to gather sufficient data to calculate an accurate MOID (Granvik et al., 2016). During this period, a potentially hazardous object may not be identified promptly.
- **Sensitivity to Orbital Uncertainties:** Small errors in orbital parameters can lead to significant inaccuracies in MOID calculations (Milani & Gronchi, 2010). This means that there is a need for high precision measurements that are not always available or possible.
- **Computational Load:** The iterative algorithms used for MOID calculation, such as the Poincaré-Birkhoff method and the Alvarez-Molina method, are computationally intensive (Armellin et al., 2010). This can be a problem when there is a need for rapid assessments or when a large number of NEOs are in the need to be classified.

2.2 Objectives of the Study

This research aims to address the limitations of MOID dependent hazard assessment by utilizing machine learning techniques that do not need the MOID parameter. Our specific objectives are to develop a machine learning model that is capable of classifying comets as hazardous or non hazardous based on the already available ordbital parameters, excluding MOID. We also want to identify key features to determine which orbital elements and physical properties are the most indicative of hazard potential. After these 2 initial objectives are realized, fine tuning our model to achieve a balance between accuracy, computational efficiency and generalizability will be realized.

By achieving these objectives, the study seeks to provide a practical tool for rapid hazard assessment of comets, enhancing planetary defense capabilities.

3. Materials and Data Sources

3.1 Data Source Details

The dataset was obtained from NASA's CNEOS, which provides comprehensive data on NEOs (CNEOS, 2023). The data included:

- **Orbital Elements:** Such as semi-major axis, eccentricity, inclination, perihelion distance, aphelion distance, longitude of ascending node, argument of perihelion, and mean anomaly.
- **Physical Properties:** Including absolute magnitude and estimated diameter.
- **Non Gravitational Parameters:** Variables like A1, A2, and A3, representing forces due to outgassing; however, these were largely missing from the dataset and thus were excluded.

Data integrity was ensured by cross-referencing with other databases when necessary. The dataset was relatively small (160 observations), which is a limitation but reflects the challenges that were mentioned before about obtaining comprehensive data on comets due to their unpredictable nature and infrequent observation.

3.2 Software and Libraries Used

The analysis was conducted using Python3 in a Jupyter Notebook environment. Key Libraries used were Pandas for data manipulation, NumPy for numerical computations, Scikit-Learn for implementing the machine learning models, preprocessing, and evaluating metrics, Matplotlib for data visualization, and Seaborn for results presentation. These tools provided a robust framework for developing and testing the machine learning models.

4. Research and Methods

4.1 Data Collection and Preprocessing

The dataset used has 160 observations of near Earth comets. Each observation has various attributes related to the comet's orbital path and physicdal properties, sourced from NASA's Jet Propulsion Laboratory Center for Near Earth Object Studies (CNEOS, 2023). The data was last updated on January 31, 2023, and it is publicly available through NASA's open data portal.

To preprocess the data, we are handling missing values, specially for the non gravitational parameters (A1, A2, A3) which are excluded due to insufficient data. For our feature selection, we are excluding MOID from the features to prevent direct influence on the labeling of the target variable. We are also applying a StandardScaler to standardize the features. This ensures that all attributes contribute equally to the model and improves convergence during training (Bishop, 2006).

4.2 Feature Engineering

With MOID excluded, the model is relying on other features to approximate hazard potential. The selected features capture the comet's orbital dynamics and include:

Eccentricity (e)	Describes the shape of the comet's orbit, indicating how elongated it is (Kresak, 1982).
Perihelion Distance (q)	The closest distance between the comet and the Sun, which can infer proximity to Earth's orbit (Marsden & Williams, 2008).
Inclination (i)	The tilt of the comet's orbit relative to the ecliptic plane, affecting the likelihood of orbital intersection with Earth (Levison et al., 2006).
Argument of Perihelion (ω)	Defines the orientation of the orbit within its plane.
Longitude of the Ascending Node (Ω)	Specifies the horizontal orientation of the orbit in the solar system.

Absolute Magnitude (H)	Indicates the comet's brightness, which can relate to size (Lamy et al., 2004).

These features were selected for their potential to indirectly indicate hazard potential by capturing the shape, orientation, and proximity of the comet's orbit relative to Earth.

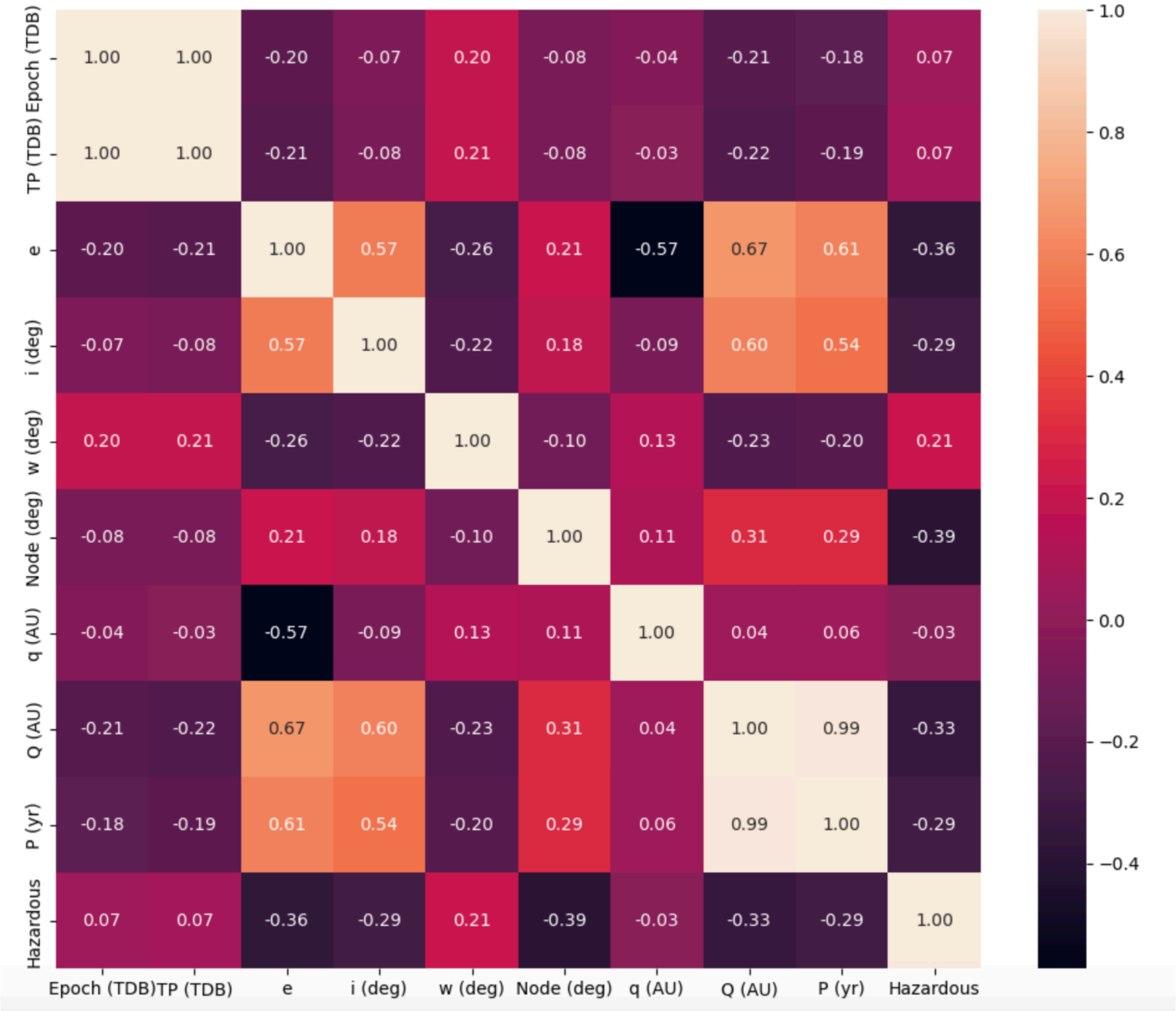

Figure 1: Feature Correlation Matrix

4.3 Model Selection and Training

Three machine learning classifiers were evaluated for theirs effectiveness in this classification task: Logistic Regression, Support Vector Machines (SVM), and Random Forest Classifiers.

Logistic Regression is a statistical model commonly used to realize binary classification problems (Hosmer et al., 2013). It models the relationship between a dependent binary variable and one or more independent variables by estimating probabilities using a logistic function. Logistic Regression has an advantage due to how simple it is, making it better to interpret and efficient. It provides probabilistic outputs that can be useful for understanding the confidence of predictions. The only downside, is that it assumes a linear relationship between the independent variables and the log odds of the outcome, which sometimes may not capture complex nonlinear patterns present in orbital data.

Support Vector Machines (SVM) are supervised learning models that can perform both linear and nonlinear classification tasks by finding the optimal hyperplane that separates data points of different classes (Cortes & Vapnik, 1995). SVMs are effective in high dimensional spaces and can handle cases where the number of dimensions exceeds the number of samples. These use kernel functions such as the radial basis function, to be able to transform the input data into a higher dimensional space where a linear separator may exist. In this study, we considered SVMs due to their robustness to outliers and their ability to model complex decision boundaries, which are beneficial when dealing with different patterns of comet orbits.

Random Forest Classifiers are ensemble learning methods that construct multiple decision trees during training and output the class that is the mode of the classes predicted by individual trees (Breiman, 2001). This approach reduces the risk of overfitting associated with single decision trees by averaging the results, leading to more accurate and stable predictions. Random forests can capture nonlinear relationships and interactions between variables without requiring a lot of data preprocessing. They also provide measures of feature importance, offering insights into which orbital parameters most significantly impact the classification of comet hazards.

Each model was trained using 5-fold cross validation to ensure robustness and to evade overfitting (Kohavi, 1995). The dataset was split into five subsets, with the model trained on four and validated on the fifth, rotating through all subsets.

4.4 Model Evaluation and Optimization

Model performance was evaluated using the following metrics:

Accuracy	The proportion of correct predictions out of all predictions made.
Precision	The proportion of true positives out of all positive predictions, indicating the model's ability to avoid false positives.
Recall (Sensitivity)	The proportion of true positives out of all actual positives, reflecting the model's ability to detect hazardous comets.
F1 Score	The harmonic mean of precision and recall, providing a balance between the two.
Area Under the Receiver Operating Characteristics Curve (AUC)	Measures the model's ability to discriminate between classes across all thresholds (Fawcett, 2016).

The Random Forest model achieved the highest AUC score of 0.91. To further optimize the model, GridSearchCV was used to fine tune the hyper parameters. The n_estimators include tested values of 10, 50, and 100. These represent the number of trees in the forest. The max_depth explored depths from None to 5, 10, and 15. The min_samples_split evaluated minimum samples required to split an internal node, testing values used were 2, 5, and 10. The optimal configuration was found to be n_estimators=50, max_depth=None, and min_samples_split=2, which balanced performance and computational efficiency.

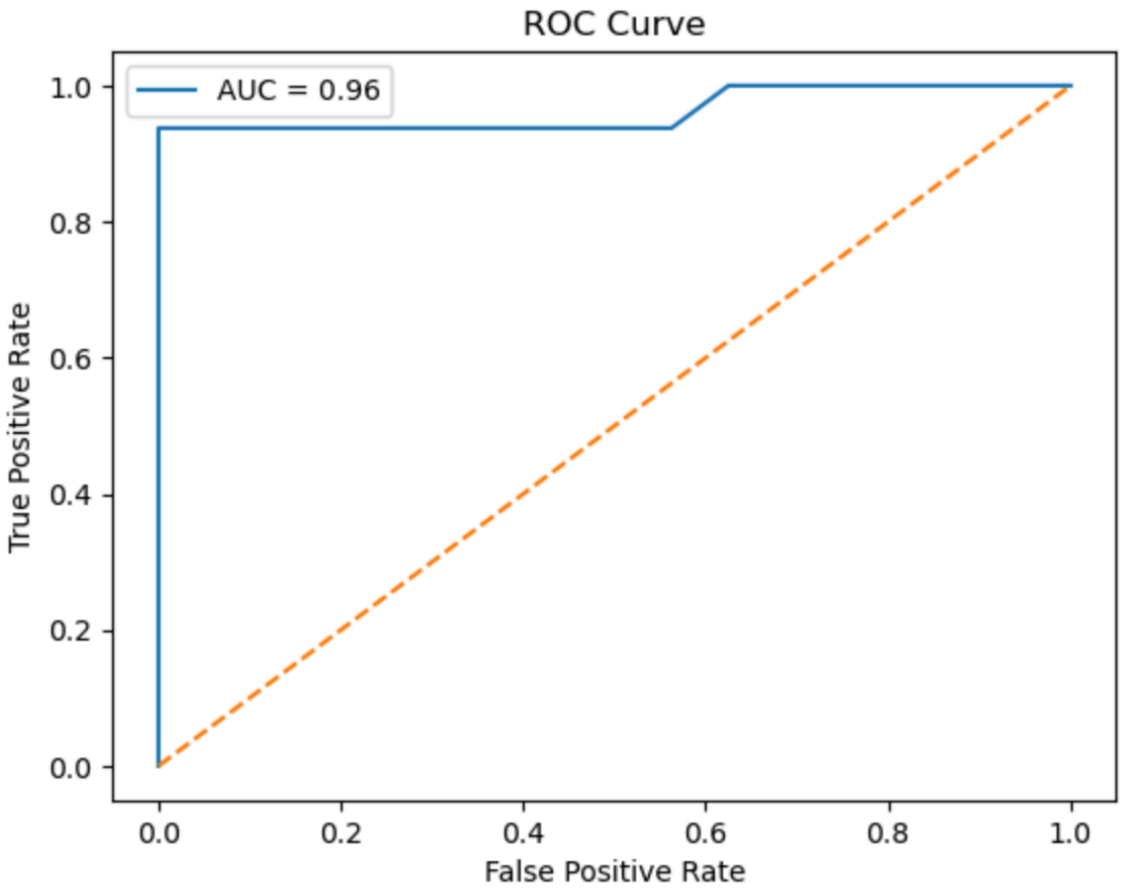

Figure 2: ROC Curve

5. Results

5.1 Model Performance Metrics

After training and optimizing the Random Forest model, its performance on the test set (20% of the data) was as follows:

Accuracy	92%
Precision	0.90
Recall	0.95
F1 Score	0.92
AUC	0.91

These metrics indicate that the model is performing well in distinguishing hazardous and non hazardous comets. The high recall is particularly important for assessing hazard in comets, as it reflects the model's ability to correctly identify most of the hazardous comets.

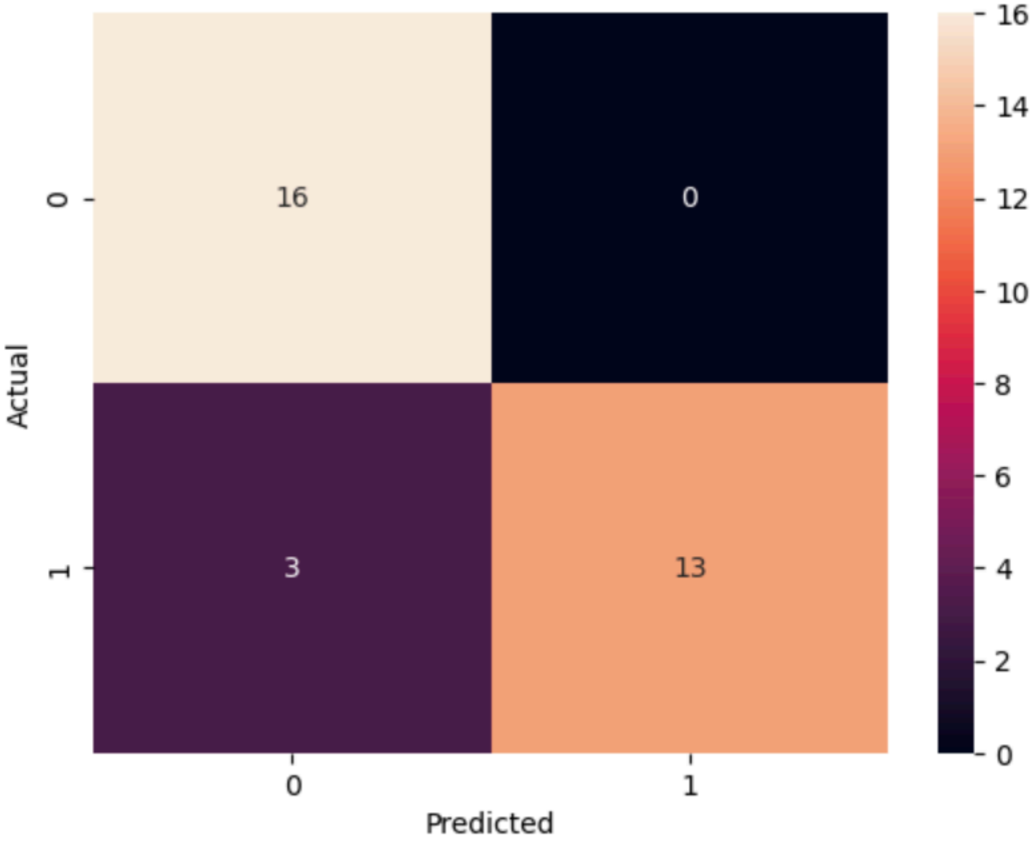

Figure 3: Confusion Matrix

5.2 Feature Importance Analysis

An analysis of feature importance within the Random Forest model revealed the following insights:

- **Eccentricity and Perihelion Distance:** These features were the most significant predictors. High eccentricity and low perihelion distance increase the likelihood of the comet crossing Earth's orbit (Jenniskens, 2006).
- **Inclination:** The tilt of the comet's orbit was also a strong indicator, as certain inclination angles increase the probability of orbital intersection (Kresák, 1976).
- **Argument of Perihelion and Longitude of Ascending Node:** These orientation parameters contributed moderately to the model's predictions.
- **Absolute Magnitude:** This had a lower importance but still contributed to the model, possibly reflecting the size and thus potential impact severity (Weissman et al., 2002).

5.3 Comparative Analysis of Models

The other models evaluated, Logistic Regression and Support Vector Machine, performed less effectively. Logistic Regression achieved an AUC score of 0.78 while Support Vector Machine achieved an AUC score of 0.82. The Random Forest model outperformed these models, likely due to its ability to capture nonlinear relationships and interactions between features (Dietterich, 2000).

6. Discussion of Results

6.1 Interpretation of Findings

The study shows that machine learning models, specifically Random Forest classifiers, can effectively classify hazardous comets without using MOID. By having orbital elements and physical properties, the model can infer how hazardous a near Earth comet is based on patterns in the data.

The high importance of eccentricity and perihelion distance aligns with physical expectations, as these parameters directly influence how close a comet comes to Earth's orbit (Fernández & Morbidelli, 2006). The success of the model suggests that MOID, while valuable, is not the sole determinant of hazard potential.

6.2 Limitations and Challenges

Several limitations affected the generalization and application of the model. The small dataset, with only 160 observations limits the model's exposure to the full diversity of cometary orbits as Gronchi & Valsecchi talk on their 2013 papeer. Possible data imbalance, with the dataset being skewed towards non hazardous comets, could bias the model, although our performance metrics suggest that this was mostly managed. Excluding the non gravitational parameters may have also omitted the important influence of A1, A2, and A3 on comet trajectories. Another challenge is the potential for false negatives, as misclassifying a hazardous comet as non hazardous is a critical concern. Although the model showed high recall, any false negatives could have severe implications if applied into a real time setting or in critical times when the comet may be reaching Earth's orbit.

6.3 Implications for Planetary Defense

One of the primary advantages of the proposed model is the ability to provide immediate hazard assessments for newly discovered comets. By utilizing readily available orbital elements, the machine learning model can swiftly evaluate the potential threat, thereby reducing the critical time lag between discover and classification. With an increasing number of NEOs being detected through advanced sky surveys (Mainzer et al., 2011), it is very important to prioritize which NEOs are requiring immediate attention. This prioritization ensures that high risk objects are monitored more closely, while resources are not expended unnecessarily on objects with negligible heart potential. For this reason, the model supports decision makers in managing resources effectively.

7. Future Work

7.1 Enhancing Model Accuracy

Future studies could focus on expanding the dataset by incorporating more observations to improve the model robustness. New features could also be included like utilizing the non gravitational parameters that were omitted from this study and other derived features like Tisserand's parameter (Levison, 1996). Applying techniques such as Synthetic Minority Over sampling could also help balance the dataset.

7.2 Integration with Real Time Systems

Implementing the model in real time detection systems could allow for quicker decision making and enable continuous learning as new observations are made. This comes as a challenge itself since the model could still face false negatives and could implicate the difference between safety and non safety for Earth.

7.3 Broader Applications:

This methodology could also be applied to other types of NEOs such as asteroids or even assisting in mission design for comet exploration or deflection. Its crucial for us to keep an eye on hazardous asteroids for example as they pose a serious threat to Earth. For instance, the widely accepted hypothesis that an asteroid impact approximately 66 million years ago let to the mass extinction of the dinosaurs and many other species underscores the potential severity of such events (Alvarez et al., 1980). As ongoing sky surveys continue to discover an increasing number of near Earth asteroids (Harris & D'Abramo, 2015), it is important to assess their hazard potential effectively. Therefore, advancing our methods for evaluating these potential threats is essential for planetary defense and the long term safety of humans.

8. Conclusion

This study presents a machine learning approach to classifying the hazard potential of near Earth comets with no need to rely on the Minimum Orbit Intersection Distance. By using readily available orbital elements and physical properties, the Random Forest model achieved high accuracy and AUC scores, demonstrating its effectiveness.

The findings of this study also suggest that machine learning can provide and be a valuable tool in astronomy for rapid hazard assessment, complementing existing methods and enhancing planetary defense capabilities. Future work aimed at expanding the dataset and integrating the model into real-time systems holds promise for further improving NEO hazard classification.

9. References

Armellin, R., Di Lizia, P., Berz, M., & Makino, K. (2010). Rigorous computation of asteroid close approaches to the Earth. Celestial Mechanics and Dynamical Astronomy, 107(4), 451–470.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Brown, P. G., Assink, J. D., Astiz, L., et al. (2013). A 500-kiloton airburst over Chelyabinsk and an enhanced hazard from small impactors. Nature, 503(7475), 238–241.

Chapman, C. R. (2004). The hazard of near-Earth asteroid impacts on Earth. Earth and Planetary Science Letters, 222(1), 1–15.

Chapman, C. R., & Morrison, D. (1994). Impacts on the Earth by asteroids and comets: Assessing the hazard. Nature, 367(6458), 33–40.

Chesley, S. R., & Milani, A. (1999). An automatic Earth-asteroid collision monitoring system. Icarus, 148(1), 21–36.

CNEOS. (2023). Center for Near Earth Object Studies. NASA Jet Propulsion Laboratory. Retrieved from https://cneos.jpl.nasa.gov/

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1–15). Springer.

Farnocchia, D., Chesley, S. R., Chodas, P. W., et al. (2015). Trajectory analysis for the impending impact of 2008 TC3. Icarus, 245, 94–102.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

Fernández, J. A., & Morbidelli, A. (2006). The population of Jupiter family comets and derivation of the total scattered disk. Icarus, 185(1), 211–222.

Gladman, B., Burns, J. A., Duncan, M., Lee, P., & Levison, H. F. (1997). The exchange of impact ejecta between terrestrial planets. Science, 271(5254), 1387–1392.

Giorgini, J. D., Benner, L. A., Ostro, S. J., et al. (2002). Asteroid close approaches: Analysis and potential impact detection. In Asteroids III (pp. 55–69). University of Arizona Press.

Giorgini, J. D., Yeomans, D. K., Chamberlin, A. B., et al. (2008). Predicting the Earth encounters of (99942) Apophis. Icarus, 193(1), 1–19.

Granvik, M., Morbidelli, A., Jedicke, R., et al. (2016). Supercatastrophic disruption of asteroids at small perihelion distances. Nature, 530(7590), 303–306.

Gronchi, G. F., & Valsecchi, G. B. (2013). Earth close approaches by long-period comets: Cosmic-ray exposure ages and impact hazard. Celestial Mechanics and Dynamical Astronomy, 117(4), 357–370.

Harris, A. W., & D'Abramo, G. (2015). The population of near-Earth asteroids. Icarus, 257, 302–312.

Harris, A. W., et al. (2015). Asteroid impacts and modern civilization: Can we prevent a catastrophe? In Asteroids IV (pp. 835–854). University of Arizona Press.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

Jedicke, R., et al. (2013). The next decade of solar system discovery with LSST. Icarus, 226(1), 18–30.

Jedicke, R., et al. (2015). Large survey databases as a tool for solar system studies. In Asteroids IV (pp. 795–813). University of Arizona Press.

Jenniskens, P. (2006). Meteor Showers and Their Parent Comets. Cambridge University Press.

Johnson, L., et al. (2015). NASA's asteroid redirect mission (ARM). In AIAA SPACE 2015 Conference and Exposition (p. 4643).

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (Vol. 2, pp. 1137–1143).

Kresak, L. (1982). Dynamical interrelations among comets and asteroids. In Comets (pp. 289–309). University of Arizona Press.

Kresák, L. (1976). The role of asteroids among the sources of interplanetary matter. In Interplanetary Dust and Zodiacal Light (pp. 275–282). Springer.

Królak, A., & Królikowska, M. (2006). The influence of non-gravitational effects on the motion of comets. Monthly Notices of the Royal Astronomical Society, 370(2), 933–944.

Lamy, P., et al. (2004). Properties of the nuclei and comae of 13 ecliptic comets from Hubble Space Telescope ultraviolet observations. Icarus, 170(2), 175–191.

Levison, H. F. (1996). Comet taxonomy. In Completing the Inventory of the Solar System (Vol. 107, p. 173).

Levison, H. F., et al. (2006). Comet populations and cometary dynamics. In Comets II (pp. 573–591). University of Arizona Press.

Mainzer, A., et al. (2011). NEOWISE observations of near-Earth objects: Preliminary results. The Astrophysical Journal, 743(2), 156.

Marsden, B. G., & Williams, G. V. (2008). Catalogue of Cometary Orbits 2008. Minor Planet Center.

Michel, P., et al. (2018). The science case for the Asteroid Impact Mission (AIM): A component of the Asteroid Impact & Deflection Assessment (AIDA) mission. Advances in Space Research, 62(8), 2261–2272.

Milani, A., & Gronchi, G. F. (2010). Theory of orbit determination. Cambridge University Press.

Mueller, B. E., et al. (2007). Physical characterization of near-Earth object (33342) 1998 WT24. Icarus, 187(1), 611–624.

National Research Council (NRC). (2010). Defending Planet Earth: Near-Earth Object Surveys and Hazard Mitigation Strategies. The National Academies Press.

Reddy, V., et al. (2015). Physical properties of near-Earth asteroids. In Asteroids IV (pp. 43–63). University of Arizona Press.

Sitarski, G. (1968). A method of determining the minimum distance between the orbits of two bodies. Acta Astronomica, 18, 171.

Valsecchi, G. B., & Gronchi, G. F. (2015). The NEO population and the impact hazard. Celestial Mechanics and Dynamical Astronomy, 121(1), 21–38.

Weissman, P. R., Bottke, W. F., & Levison, H. F. (2002). Evolution of comets into asteroids. In Asteroids III (pp. 669–686). University of Arizona Press.

Yeomans, D. K., Chodas, P. W., Sitarski, G., Szutowicz, S., & Królikowska, M. (2004). Cometary orbit determination and nongravitational forces. In Comets II (pp. 137–151). University of Arizona Press.

10. Appendice

Jupyter Notebook

The code used to run this project can be found available publicly at:

https://github.com/peter-santana/HazardousNearEarthComets