Automated End-to-End Pipeline for Near-Earth Object Hazard Classification Using MLOps Frameworks

1st Subrat Mishra
Independent Researcher / Data Scientist
Email: 3subratmishra1sep@gmail.com

2nd Subasis Mishra
CS/IT, ITER SOA, Bhubaneswar
Email: subasismishra3124@gmail.com

Abstract:

Near-Earth Objects (NEOs) pose potential threats to Earth due to their proximity and impact potential. Accurate classification of NEOs as hazardous or non-hazardous is essential for risk mitigation and planetary defence. This work presents a production-grade machine learning (ML) pipeline for automated NEO hazard classification using NASA datasets. The pipeline encompasses data ingestion, preprocessing, model training, evaluation, monitoring, and deployment. Multiple ML algorithms, including Logistic Regression, Random Forest, and XGBoost, were evaluated, with model performance tracked using MLFlow. The pipeline demonstrates effective hazard prediction and incorporates automated data drift detection to maintain reliability over time.

Introduction

Near-Earth Objects (NEOs), including asteroids and comets whose orbits bring them within 1.3 astronomical units (AU) of the Sun, have gained increasing attention due to their potential threat to Earth. While catastrophic collisions are rare, even relatively small objects can cause significant damage, as exemplified by the Chelyabinsk meteor in 2013, which injured approximately 1,500 people despite being only about 20 meters in diameter. To date, astronomers have discovered over 30,000 NEOs, of which roughly 2,500 are classified as Potentially Hazardous Objects (PHOs) based on their size and proximity to Earth.

Major space agencies, such as NASA through its Planetary Defence Coordination Office (PDCO) and the European Space Agency (ESA) via the Space Situational Awareness program, actively support NEO detection and characterization. These initiatives rely on a combination of telescopic observations, radar systems, and infrared space missions to collect physical and orbital data. Despite their effectiveness, observational limitations—such as atmospheric interference, exposure cycles, and challenges in detecting dark or non-reflective objects—constrain traditional methods. Furthermore, the rapid increase in NEO observations has outpaced the capacity for systematic analysis, introducing challenges of scale, complexity, and uncertainty that demand intelligent, automated processing.

In response, this work presents a **machine learning–based pipeline** for NEO hazard classification. Leveraging a curated dataset of orbital and physical parameters—including absolute magnitude, estimated diameters, close-approach distances, and relative velocities—we develop supervised models to predict the hazard status of each object. The pipeline incorporates comprehensive preprocessing steps such as normalization, missing value handling, categorical encoding, and feature engineering (e.g., deriving diameter_range) to enhance predictive power.

Given the significant class imbalance in NEO datasets, where hazardous objects are rare, we employ **Border Line Synthetic Majority Over-Sampling Technique (SMOTE)** to balance minority class instances, ensuring unbiased model training and improved generalization. A variety of supervised classifiers are evaluated, including Logistic Regression, Decision Trees, k-NN, Support Vector Machines, Random Forests, XGBoost, and CatBoost. Performance is assessed using metrics such as accuracy, precision, recall, and F1-score, providing insight into the suitability of different models for real-world astronomical data.

Beyond classification, our framework incorporates **risk prioritization** by combining model outputs with domain-specific metrics such as estimated impact energy and historical risk scales (e.g., Palermo and Torino scales). This allows for actionable hazard rankings, guiding astronomers and planetary defence agencies in allocating observational resources efficiently. The system is designed to update dynamically as orbital and physical parameters evolve over time.

The remainder of this paper is organized as follows: Section 1 describe about the dataset. Section 2 reviews related work on NEO detection and machine learning approaches. Section 3 describes the dataset, preprocessing steps, and modelling methodology. Section 4 presents experimental results and model comparisons. Section 5 discusses implications for planetary defence and potential extensions, including real-time data integration and multi-sensor fusion.

Data

The dataset used in this study was sourced from NASA's Near-Earth Object (NEO) API (https://api.nasa.gov/), which provides detailed information on asteroids and other celestial objects approaching Earth. For this workflow, a **10-year historical window (2015–2025)** was considered to capture sufficient temporal variability for training robust hazard prediction models.
The extracted dataset includes the following columns:

- id: Unique identifier for each NEO
- name: Name of the NEO
- absolute_magnitude_h: Absolute magnitude (brightness) of the NEO
- min_diameter_km & max_diameter_km: Minimum and maximum estimated diameters (km)
- close_approach_date & close_approach_date_full: Date and full timestamp of closest approach to Earth
- epoch_date_close_approach: Epoch time of closest approach

- relative_velocity_kps: Relative velocity with respect to Earth (km/s)
- miss_distance_km: Closest approach distance to Earth (km)
- orbiting_body: Celestial body the NEO is orbiting
- is_potentially_hazardous: Boolean flag indicating potential hazard
- is_sentry_object: Boolean flag indicating whether the NEO is on the Sentry risk table
- nasa_jpl_url: Reference URL from NASA's JPL for the NEO

Related Work

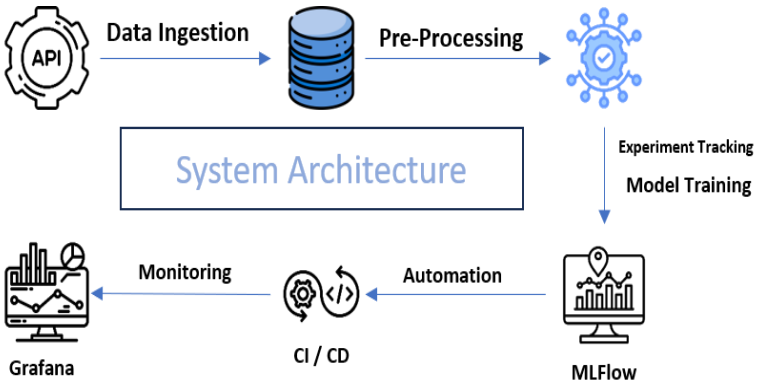
Prior research on Near-Earth Object (NEO) hazard classification has predominantly focused on applying machine learning models to static datasets for offline analysis. Traditional approaches often rely on pre-collected observational data, limiting their ability to adapt to newly discovered NEOs in real time.

Several studies have addressed class imbalance, a common challenge in NEO datasets where hazardous objects are significantly underrepresented. Techniques such as **SMOTE (Synthetic Minority Over-sampling Technique)** and its variants have been shown to improve the detection of minority classes, enhancing the predictive performance of classifiers.

While these studies demonstrate the potential of machine learning for hazard classification, **few works have implemented fully automated pipelines** encompassing continuous data ingestion, model retraining, performance monitoring, and data drift detection. The integration of such end-to-end pipelines is essential for maintaining operational reliability and ensuring timely identification of potentially hazardous objects.

System Architecture

The proposed pipeline consists of the following components:



- **Data Ingestion:** NEO observational and derived data are periodically fetched via NASA's NeoWs API.
- **Database:** PostgreSQL stores raw and processed data, enabling version control and query efficiency.
- **Preprocessing:** Automated scripts handle missing values, feature scaling, and outlier detection. SMOTE is applied to training data to address class imbalance.
- **Modelling and Experiment Tracking:** Supervised learning models are trained and logged with MLFlow; DAGsHub tracks dataset versions and experiments.
- **Automation & CI/CD:** GitHub Actions orchestrate pipeline execution, triggering data updates, retraining, and deployment without manual intervention.
- **Monitoring:** Grafana dashboards visualize model predictions, performance metrics, and data drift detection alerts

Methodology

- **Data Handling:** SMOTE is applied only on training data to avoid data leakage.
- **Modelling:** Random Forest and Gradient Boosting classifiers were evaluated for hazard prediction.
- **Automation:** GitHub Actions schedules periodic data fetching, model retraining, and evaluation.
- **Drift Detection:** Continuous monitoring identifies significant changes in feature distributions, triggering retraining when needed.

Results

The automated pipeline demonstrated:

- **High Recall for Hazardous NEOs:** 0.98
- **Robust F1-score:** 0.51 for minority class
- **Operational Reliability:** Continuous updates with zero manual intervention
- **Data Drift Detection:** Alerts for significant deviations in key features (e.g., diameter, relative velocity)

The system successfully maintained model performance over successive retraining cycles while tracking experiment metadata and dataset versions.

Discussion

By combining SMOTE with an automated pipeline, the system reliably identifies hazardous NEOs, minimizing false negatives, which is crucial for planetary defence. The integration of MLOps tools ensures reproducibility, transparency, and operational stability. Future improvements could include ensemble modelling, real-time API endpoints, and expanded visualization metrics.

Conclusion

This study presents a fully automated, MLOps-driven pipeline for NEO hazard classification. By integrating data ingestion, preprocessing,

modelling, monitoring, and retraining, the framework demonstrates that machine learning can be applied at scale in operational environments for planetary defence applications.

References:

- Chawla, N. V., et al. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.
- NASA NeoWs API Documentation.
- MLflow Documentation.
- Grafana Documentation.
- DagsHub Documentation.