SPACECODE_PS3_BEETLEJUICE2.0

The Sentinel Shield

Near-Earth Comet (NEC) Classification

A Machine Learning Approach to Classifying Potentially Hazardous Objects for Planetary Defense

Date: January 11, 2026

1. Problem Statement & Objective

Our solar system is a cosmic shooting gallery. Among the millions of objects orbiting the Sun, Near-Earth Objects (NEOs) - specifically comets and asteroids - pose a unique challenge to Earth's safety. While most bypass our planet at safe distances, a select few are designated as Potentially Hazardous Objects (PHOs).

According to NASA/JPL CNEOS standards, a "Hazardous" classification is triggered by two critical physical thresholds:

- Proximity: A Minimum Orbit Intersection Distance (MOID) of 0.05 AU or less
- Size/Magnitude: An Absolute Magnitude of 22.0 or brighter, suggesting the object is large enough (approx. 140 meters) to cause significant regional damage upon impact

Our objective was to develop a supervised machine learning model to automate this Hazardous Classification. The model must accurately predict whether a celestial body is a threat based on its orbital elements and physical properties. We prioritized RECALL over precision - missing a hazardous object (False Negative) is far more catastrophic than a false alarm (False Positive).

2. Data Description & Preprocessing

We worked with NASA's Near-Earth Object Wide-field Survey dataset containing orbital parameters from 1950 to 2025. The dataset includes over 1.3 million records with the following feature categories:

- Physical Properties: Absolute Magnitude and Estimated Diameter
- Orbital Shape & Tilt: Eccentricity, Inclination, Semi-Major Axis, Perihelion/Aphelion
- Proximity Metrics: Minimum Orbit Intersection Distance (MOID) and Miss Distance
- Dynamics: Relative Velocity, Orbital Period, Jupiter Tisserand Invariant
- Reliability: Orbit Uncertainty and Orbit Determination Date

Key preprocessing steps performed:

- Dropped high-cardinality columns (Neo Reference ID, Name, Date, Close Approach Date, Orbit Determination Date, Equinox, Orbiting Body) - these add no predictive value and would cause the model to memorize instead of generalize
- Imputed missing values with median for numeric columns - median is robust to outliers which is important for orbital data with extreme values
- Converted Hazardous target to binary (0/1) and removed records with missing labels
- Applied StandardScaler to normalize all features to zero mean and unit variance - critical for algorithms like Logistic Regression that are sensitive to feature scales

3. Handling Class Imbalance

Only ~10% of objects are hazardous. A naive model could achieve 90% accuracy by always predicting "Safe" - useless for planetary defense. We addressed this using:

- SMOTE: Created synthetic hazardous samples via interpolation
- Class Weights: Penalized misclassifying hazardous objects more heavily
- Undersampling: Reduced majority class to balance the dataset

Class weighting proved most effective - it maximizes recall without increasing training time.

4. Model Training

We trained 5 classifiers with class_weight="balanced" to handle imbalance:

- Logistic Regression - interpretable baseline
- Random Forest - ensemble of decision trees
- Gradient Boosting - sequential error correction
- XGBoost - optimized gradient boosting with regularization
- LightGBM - fast gradient boosting, leaf-wise growth

5. Results

Model performance on test set (sorted by Recall):

Model	Recall	Precision	F1	AUC-ROC
Random Forest	99.94%	99.95%	99.94%	99.99%
Gradient Boosting	99.93%	99.96%	99.94%	99.99%
XGBoost	99.92%	99.90%	99.91%	99.99%
LightGBM	99.91%	99.92%	99.91%	99.99%
Logistic Regression	95.69%	67.33%	79.04%	98.80%

Random Forest achieved the best recall (99.94%), correctly identifying virtually all hazardous objects. Tree-based ensembles significantly outperformed Logistic Regression.

6. Most Important Features

XGBoost feature importance revealed the most predictive factors:

- Minimum Orbit Intersection Distance (MOID) - closest approach to Earth's orbit
- Absolute Magnitude - indicates object size (lower = larger)
- Estimated Diameter - physical size in kilometers
- Miss Distance - actual closest approach during observation

These align with NASA's physical criteria, validating that our model learned the correct physics.

7. Conclusion

This project successfully developed a machine learning system for automated hazardous classification of Near-Earth Objects. Our key achievements include:

- Developed a robust preprocessing pipeline for handling high-dimensional orbital data with 1.3M+ records
- Addressed extreme class imbalance (~10% hazardous) using multiple techniques, with class weighting proving most effective
- Trained and evaluated 5 classification algorithms - Random Forest and Gradient Boosting achieved the highest recall (>99.9%)
- Identified the most important predictive features (MOID, Absolute Magnitude, Diameter), validating alignment with NASA's physical hazard criteria

The best performing model (Random Forest) correctly identifies 99.94% of all hazardous objects while maintaining 99.95% precision. This means we miss fewer than 1 in 1000 dangerous asteroids while keeping false alarms minimal - exactly what's needed for a reliable planetary defense system.

Recommendations for production deployment:

- Use XGBoost or LightGBM for best balance of performance and inference speed
- Consider ensemble voting of multiple models for maximum safety in critical applications
- Implement threshold tuning to further optimize the recall-precision trade-off based on operational requirements

Key Takeaway: For planetary defense and similar rare-event detection problems, class weighting and ensemble methods are essential. High accuracy alone is meaningless - what matters is catching every hazardous object, because missing even one could be catastrophic.