Executive Summary: Classification of Near-Earth Objects (NEOs)

Team NASA: Munawar Ali, Cagatay Ayhan, Ece Karacam, Mostofa Hisham, Waleed Ahmed,
Mentor: Kashif Bari
GitHub: https://github.com/munawarali93/NASA-Near-Earth-Objects

Overview

Near-Earth Objects (NEOs) are comets and asteroids that have been nudged by the gravitational attraction of nearby planets into orbits that allow them to enter the Earth's neighborhood. As they orbit the Sun, Near-Earth Objects occasionally approach close to Earth. We use the dataset provided in Kaggle. We explore how different models including logistic regression and random forest perform in detecting hazardous objects.

Stakeholders: This project has relevance for space agencies, scientists, and the public at large. For our project, we focus on delivering results for space agencies such as NASA and CSA.

KPIs: We decided that the key metrics for this project are F1-score, Precision and Recall. This follows from the expectation that the dataset is imbalanced and very few objects are hazardous.

Approach & Results

We will use four different models to classify whether a NEO is hazardous using the dataset available here where we have 90836 observations with 9 features.

1. Random Forest: This model is an ensemble learning method which relies on bagging: trees are fit independently, and results are averaged for all models. Model parameters used were n_estimators = 50, max_depth = 4. This gave a very good result with an F1-score of 95%.
2. Decision Trees: Decision trees with varying maximum depths have been trained. We implement the algorithm up to 20 layers, beyond which would have overfitted the data. Nonetheless, predictions using decision trees are not entirely accurate due to the skewed nature of the data in favor of the non-hazardous materials. More specifically, they exhibit an average accuracy of 90%, their F1 score is only 38%. In conclusion, these trees do not perform well in detecting hazardous objects.
3. Logistic Regression: Logistic regression for binary classification was trained. As logistic regression assumes independence of features so features depending upon each other were removed. The model gave better results i.e., F1 score for the train and test set was 95 and 94 respectively.
4. XGBoost: XGBoost is a versatile algorithm for binary and multiclass classification. It uses an ensemble of decision trees to model the relationship between features and the target variable. It employs logistic function for binary classification. Sigmoid transformation ensures predictions between 0 and 1. After doing some careful selection of hyperparameters we get better results here as compared to other models we trained. F1 score for the train and test set was 96 and 95 respectively.

Summary:

We presented several models for identifying Near Earth Objects using the Kaggle dataset. We compared Random Forest, XGBoost, Decision Trees, and Logistic regression and we found that the XGboost is the best performing model with an F1-score of 96%.