Project G6: Asteroid Impact Prediction

Team Members: Aditya Hishobkar, Deepika Sambrekar and Naghma Afreen
Repository: https://github.com/NaghmaA/IDS-Project2023

Business understanding

Identifying your business goals

Background:

- There is an increasing requirement for sophisticated techniques to evaluate and reduce the risk of asteroid impacts on Earth due to the growing interest in space exploration and the potential threats posed by Near-Earth Objects (NEOs). The project's goal is to use Machine Learning (ML) algorithms to analyze data about the size, composition, and orbit of asteroids in order to identify those that may be dangerous.

Business Goals:

- Develop a strong machine learning model that can recognize Near-Earth Objects (NEOs) that could be dangerous to Earth.
- Enhancing our knowledge of asteroidal properties will help us assess impact risk with greater accuracy.
- By offering accurate and current data on potentially dangerous asteroids, you can support international efforts to defend the planet.

Criteria for Business Success:

- Achieve a high degree of precision in recognizing asteroids that may pose a threat.
- Send out early alerts for asteroids that have the potential to cause major impacts and whose orbits cross over Earth's orbit.
- Acquire acknowledgement and cooperation from scientific communities, planetary defense organizations, and space agencies.

Assessing Your Situation:

Inventory of Resources:

- comprehensive dataset (191 MB) with details on the composition, size, and orbit of NEOs.
- ML knowledge available within the group.
- Potential for cooperation with astronomers, researchers, and space agencies.

Requirements, Assumptions, and Constraints:

- Access to up-to-date or real-time asteroid data is a prerequisite.
- Based on past data, machine learning algorithms are assumed to accurately predict potentially dangerous asteroids.
- Limited computer power to handle big datasets is a constraint.

Risks and Contingencies:

- Risk: Predictions that are erroneous due to dataset inaccuracies.
- Implement procedures for data cleaning and validation as a backup plan.
- Risk: Outside variables that have an impact on how accurate orbit predictions are.
- Contingency: Consistently add the most recent orbit data to the model.

Terminology:

- To establish a common vocabulary among team members, define important terms pertaining to impact risk assessment, ML metrics, and asteroid characteristics.

Drawbacks and Advantages:

- Costs: Data acquisition, computational resources, and possible joint venture expenditures.
- Benefits include enhanced public safety, enhanced planetary defense capabilities, and possible joint venture opportunities.

Defining your data-mining goals

Data-Mining Goals:

- Create machine learning algorithms that can reliably identify asteroids that may be dangerous based on factors like size, composition, and orbit.
- Put in place a prediction system that can provide early warnings in real-time or almost real-time.

Data-Mining Success Criteria:

- Attain a high recall and precision rate when recognizing potentially dangerous asteroids.
- Put in place a system that can issue warnings in a timely manner with few false positives.

Understanding data

Gathering data:

Outline data requirements:

- A thorough dataset with details on Near-Earth Objects (NEOs) is our project's main data requirement. Details about the asteroids' composition, size, and orbit will be included in this dataset. Additionally, to guarantee the accuracy of our ML model, real-time or frequently updated data is required.

Verify data availability:

- We have located a publicly accessible, well-documented dataset that is appropriate for our project. The dataset, which weighs 435.4 MB, satisfies our requirements for size, orbit, composition data and contains a wealth of information about NEOs.

Define selection criteria:

Our criteria for choosing the data consist of:

- Adding the following pertinent attributes: composition, size, and orbit.
- Frequent updates: Making sure the dataset is current to enable precise forecasting.
- Credibility of the data: selecting information sources about NEOs that have a history of producing trustworthy data.

Describing data:

The following crucial attributes are provided by the dataset:

- Name/ID: Every Near-Earth Object's special identification number.
- Size: The asteroid's circumference or measurements.
- An asteroid's orbit is defined by its semi-major axis, eccentricity, and orbital period, among other factors.
- Composition: Details regarding the chemical composition of the asteroid.

Exploring data:

We conducted a preliminary analysis during the exploration stage to learn more about the dataset:

- Size Distribution:
  - The distribution of asteroid sizes was visualized.
  - Identified any extreme or out-of-the-ordinary values that might need extra attention.
- Features of the Orbit:
  - Investigated the distribution of orbital parameters.
  - Looked for any trends or groups in the data that would help identify asteroids that might be dangerous.
- Analysis of Composition:
  - Investigated the range of NEO compositions.
  - Evaluated the frequency of specific materials that could affect impact risk.

- Temporal Patterns:
    - Examined the dataset's temporal trends in order to find any patterns or irregularities over time.
    - Looked for patterns or seasonality that might affect our prediction model.

Verifying data quality:

To ensure the quality of our data, we performed the following checks:

- Missing Values:
    - Any missing values in important fields were located and addressed.
    - Considering how missing data might affect how accurate our predictions are.
- Outlier Identification:
    - Made use of statistical techniques to identify anomalies in composition, size, or orbit.
    - Outliers that might distort our ML model were addressed.
- Consistency checks:
    - Made sure that the units and data format were the same for every attribute.
    - Uniform units to make accurate analysis easier.
- Cross-validation:
    - To verify the accuracy of important attributes, the dataset was cross-validated against outside sources.
    - Confirmed that the dataset is consistent with the body of knowledge in science regarding NEOs.

Project Planning

Goal 1: Identifying Potentially Dangerous Asteroids

- ***Task1:*** *Data Exploration and Preprocessing*
    - Explore Dataset to understand its structure and features. Preprocess data, handle missing values, and encode categorical variables.
    - Time Estimate: 12 hours (Aditya: 4 hours, Deepika: 4 hours, Naghma: 4 hours)
- ***Task2:*** *Feature Selection and Scaling*
    - Identify relevant features for predicting asteroid danger. Scale numerical features to bring them to a consistent scale.
    - Time Estimate: 12 hours (Aditya: 4 hours, Deepika: 4 hours, Naghma: 4 hours)
- ***Task3:*** *Train Classification Model*
    - Choose and train a classification model (e.g., Random Forest, SVM) to identify dangerous asteroids. Evaluate the model using appropriate metrics.
    - Time Estimate: 18 hours (Aditya: 6 hours, Deepika: 6 hours, Naghma: 6 hours)

Goal 2: Predicting the Trajectory of Asteroids

- ***Task4:*** *Data Preparation for Trajectory Prediction*
    - Explore and preprocess the dataset for trajectory prediction. Extract relevant features related to current position and velocity.
    - Time Estimate: 12 hours (Aditya: 4 hours, Deepika: 4 hours, Naghma: 4 hours)
- ***Task5:*** *Train Regression Model for Trajectory Prediction*
    - Choose and train a regression model (e.g., Random Forest Regressor, XGBoost) for trajectory prediction. Visualize and validate the predicted trajectories.
    - Time Estimate: 18 hours (Aditya: 6 hours, Deepika: 6 hours, Naghma: 6 hours)

Final Poster

- ***Task6:*** *Preparing Poster*
    - Documenting the process, including methodologies, tools used, and challenges faced. Preparing final poster summarizing findings and model performance.

- Time Estimate: 18 hours (Aditya: 6 hours, Deepika: 6 hours, Naghma: 6 hours)

Methods and Tools:
Data Exploration and Preprocessing: Python, pandas, scikit-learn.
Classification Model: Random Forest Classifier, Support Vector Machine (SVM), etc.
Trajectory Prediction Model: Random Forest Regressor, XGBoost, etc.
Visualization: matplotlib, 3D visualization libraries.