Asteroid Close Approach Data Analysis and Prediction

Team name: Rotten Idli

Dadi Sasank Kumar
24-798329

Muppalla Komal
24-394625

Chandravadhan Raj Bethala
24-430989

October 13, 2024

Project Overview

The goal of this project was to analyze a dataset of asteroid close approaches to Earth. The dataset consisted of information on the asteroids' orbital characteristics, velocity, approach dates, and distances from Earth. Some fields were incomplete, requiring imputation and preprocessing before building a machine learning model to predict whether an asteroid is hazardous to Earth.

Dataset Summary

- **Total Rows:** 4534
- **Total Columns:** 24

Key Columns and Data Types:

- **Numerical Columns (20):**
  Name, Epoch Date Close Approach, Relative Velocity km per hr, Miles per hour, Miss Dist. (Astronomical), Miss Dist. (lunar), Miss Dist. (kilometers), Miss Dist. (miles), Jupiter Tisserand Invariant, Epoch Osculation, Semi Major Axis, Asc Node Longitude, Perihelion Arg, Aphelion Dist, Perihelion Time, Mean Anomaly, Mean Motion, approach_year, approach_month, approach_day.
- **Categorical Columns (4):**
  - Relative Velocity km per sec: Categorical descriptor for asteroid velocity in qualitative terms (e.g., 'Very Slow', 'Fast').
  - Orbital Period: Categorical descriptor indicating the orbital period (e.g., 'Low', 'Medium', 'High').
  - Orbit Uncertainty: Categorical descriptor indicating the uncertainty in the orbit (e.g., 'Low', 'Medium', 'High').
  - Hazardous: A boolean column indicating whether the asteroid is hazardous (True) or not hazardous (False).

Handling Missing Data

The dataset contained missing values in several columns. A detailed summary of missing values by column is shown below:

Column Name	Missing Values
Epoch Date Close Approach	1254
Relative Velocity km per sec	1350
Relative Velocity km per hr	1501
Miles per hour	866
Miss Dist. (Astronomical)	601
Miss Dist. (lunar)	1117
Miss Dist. (kilometers)	1368
Miss Dist. (miles)	652
Jupiter Tisserand Invariant	1732
Epoch Osculation	1527
Semi Major Axis	1188
Asc Node Longitude	1096
Perihelion Arg	1134
Aphelion Dist	815
Perihelion Time	1564
Mean Anomaly	918
Mean Motion	1508
approach_year	819
approach_month	1528
approach_day	543
Orbital Period	530
Orbit Uncertainty	1767

Table 1: Summary of Missing Data

Imputation of Missing Data

K-Nearest Neighbors (KNN) Imputation:
KNN imputation was used to fill in missing numerical values. KNN finds the K nearest records with complete data and imputes the missing values based on the average of the K nearest neighbors. This method works well for datasets where relationships between variables are not strictly linear.

For categorical columns like 'Relative Velocity km per sec', mode imputation (filling with the most frequent value) was applied where applicable.

Feature Engineering

Binning of Continuous Variables:
To improve interpretability, some continuous variables were binned. For example, the 'Relative Velocity km per hr' column was divided into categorical bins:

- **Very Slow:** [1207.81, 26064.64)
- **Slow:** [26064.64, 49649.14)
- **Medium:** [49649.14, 73233.64)
- **Fast:** [73233.64, 160681.49)

Correlation Analysis:
A correlation matrix was generated to identify relationships between numerical columns. Below are a few significant findings:

- **Epoch Date Close Approach** showed a moderate positive correlation with 'Name' (0.19). This could imply that asteroids with higher 'Name' values tend to have more recent or future close approaches.
- **Miss Distances** (in various units) had weak negative correlations with 'Relative Velocity km per hr' (-0.17). This suggests that faster-moving asteroids may tend to pass closer to Earth.

Unique Values in Categorical Columns:

- 'Relative Velocity km per sec': {'Very Slow', 'Slow', 'Fast', 'Very Fast'}
- 'Orbital Period': {'Low', 'Medium', 'High'}
- 'Orbit Uncertainty': {'Low', 'Medium', 'High'}

Model Building and Evaluation

After data preprocessing and imputation, the dataset was split into training and test sets. The target variable for classification was whether the asteroid is **Hazardous** or not.

Model Selection

Several models were considered for the task, including logistic regression, decision trees, and random forests. After experimentation, an **ensemble method** combining multiple models was selected for final training due to its superior performance.

Training

The model was trained on the preprocessed data with features including asteroid velocities, distances, and orbital parameters. Categorical features were one-hot encoded to convert them into numerical format for training.

Performance

The final ensemble model achieved an accuracy of **85%** on the test set, indicating good performance in predicting the hazardous nature of asteroids.

Data Visualization

Correlation Heatmap:
The correlation matrix was visualized using a heatmap to understand relationships between key features. Some important observations:

- Positive correlations between 'approach_year' and 'Name' values indicate a trend where certain asteroid IDs may be more frequent in later years.
- Weak correlations between velocity and miss distances suggest that while velocity might play a role in close approaches, it is not a strong standalone predictor.

Histograms:
Several histograms were plotted to show the distribution of key features such as 'Miss Distances', 'Relative Velocities', and 'Orbital Periods'.

Conclusion

In this project, I successfully handled missing values using **KNN imputation**, performed feature engineering (including binning and encoding), and built an ensemble classification model to predict the hazardous nature of asteroids. The model performed well with an accuracy of **85%**, making it a reliable tool for predicting potentially hazardous asteroids.

Final Datasets

- **Imputed Dataset:** final_imputed_dataset.csv
- **Preprocessed Dataset:** Preprocessed_dataset.csv

This project demonstrates the importance of preprocessing and imputation in dealing with incomplete datasets, as well as the effectiveness of ensemble models in classification tasks with mixed data types.