HAZARDOUS ASTEROID PREDICTION

[redacted]

Praveena E

Problem Statement

- A Machine Learning Project to Predict the Hazardous Asteroids based on Asteroid Dataset and The Power BI Report for the Same.

Dataset

- **Dataset Name** : Asteroid Dataset - NASA JPL Asteroid Dataset
- **Problem Type** : Classification
- **Dataset Link** : Click Here
- **Shape of Dataset** : 958524 Rows, 45 Columns

Some Features in Dataset

- **NEO** : Near-Earth Object (NEO) flag
- **PHA** : Potentially Hazardous Asteroid (PHA) flag
- **H** : Absolute magnitude parameter
- **Orbit_id** : Orbit solution ID
- **Epoch** : Epoch of osculation in modified Julian day form
- **Equinox** : Equinox of reference frame
- **e** : Eccentricity
- **a** : Semi-major axis au Unit
- **q** : perihelion distance au Unit
- **i** : inclination; angle with respect to x-y ecliptic plane
- **moid_ld** : Earth Minimum Orbit Intersection Distance au Unit

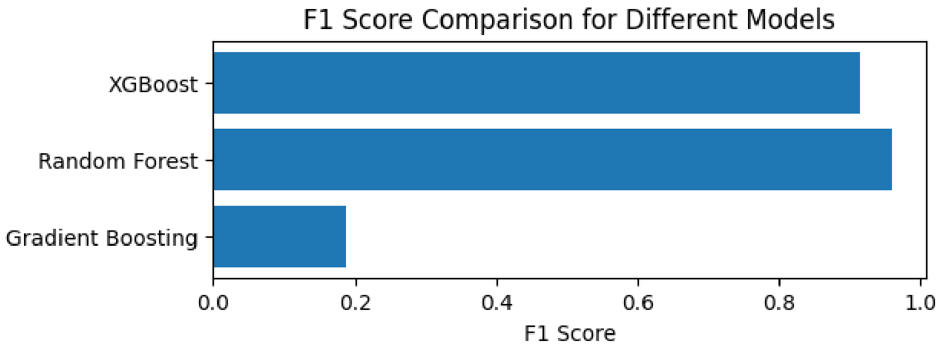
Preprocessing

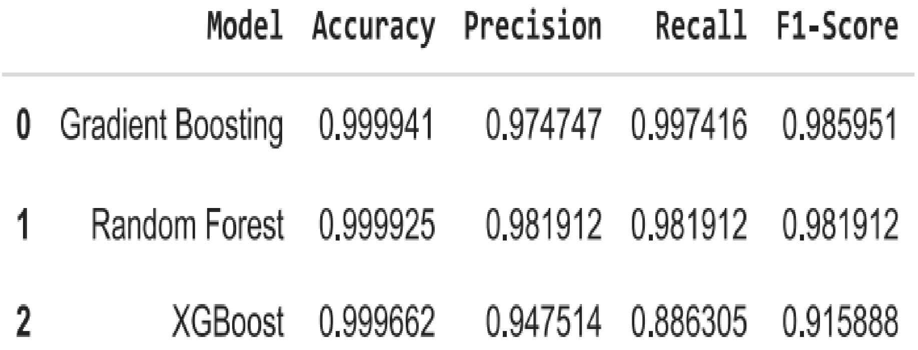
- I conducted preprocessing by removing irrelevant columns like 'diameter', 'albedo', 'diameter_sigma', 'id', 'spkid', 'full_name', 'pdes', 'name', 'prefix', and 'equinox' due to high null values or lack of relevance to the prediction.
- After handling missing values, the dataset's shape was reduced to (932335, 35).
- I then split the data into independent and dependent variables, with 'PHA' as the target.
- To prepare for modeling, I applied one-hot encoding for categorical variables and used MinMax Scaler to normalize the features, ensuring consistency across the dataset.

Model Training

- For training, I used ensemble techniques like Random Forest, Gradient Boosting and Extreme Gradient Boosting to address the highly imbalanced dataset.
- Where 'PHA' had 936,537 'N' values (non-hazardous asteroids) and 2,066 'Y' values (hazardous asteroids).

Model Evaluation





	Model	Accuracy	Precision	Recall	F1-Score
0	Gradient Boosting	0.999941	0.974747	0.997416	0.985951
1	Random Forest	0.999925	0.981912	0.981912	0.981912
2	XGBoost	0.999662	0.947514	0.886305	0.915888

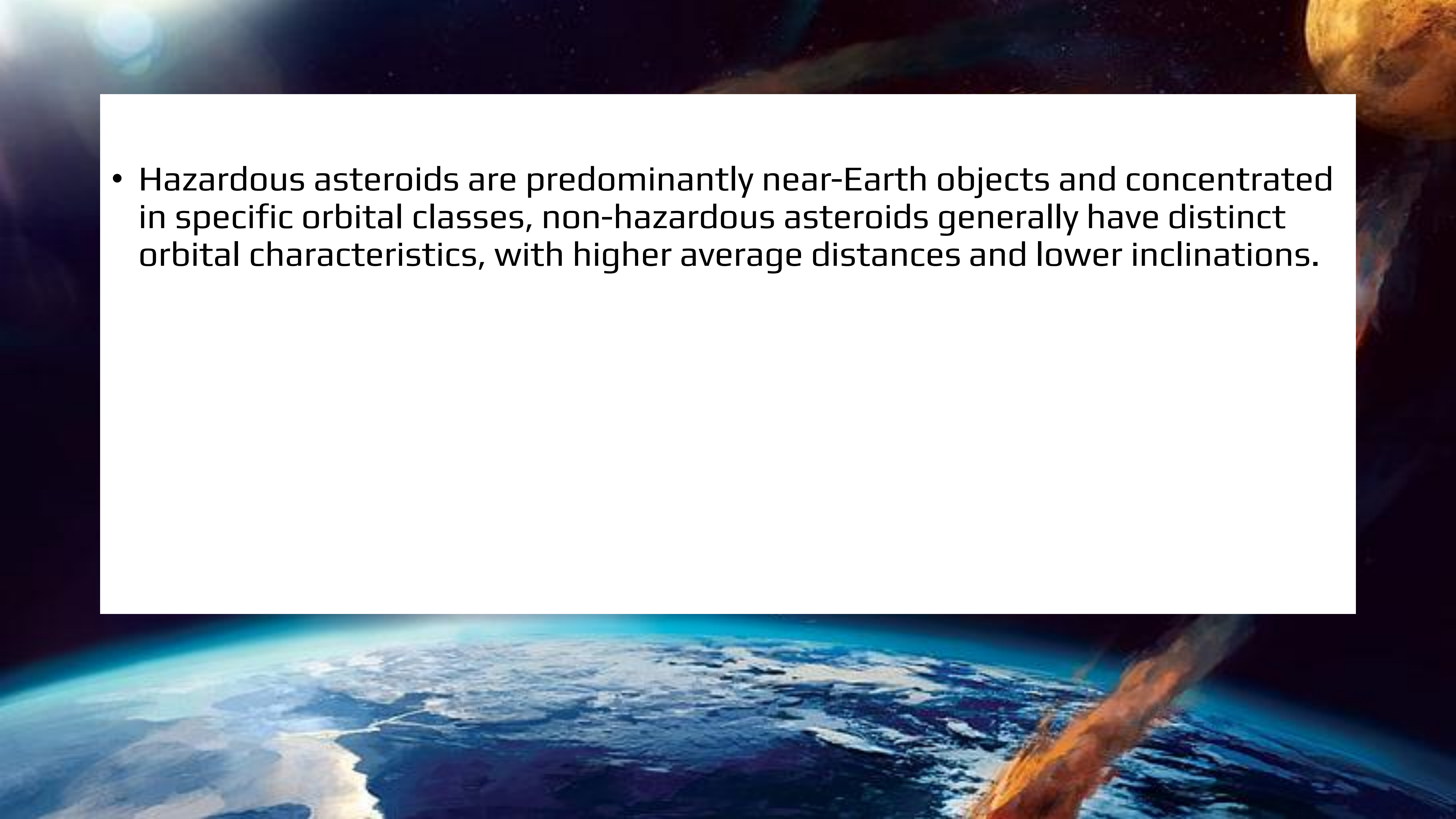
The results clearly show that ***Random Forest*** outperforms other models, delivering the best precision and recall for this prediction.

key Insights into the Asteroid Dataset

- The Power BI report revealed key insights into the asteroid dataset.
- Out of 932,000 total asteroids, 2,066 are hazardous, and 23,000 are classified as near-Earth objects (NEO).
- All non-near-Earth objects are non-hazardous, while most of the hazardous asteroids are near-Earth object.
- Most asteroids belong to the MBA classification, while hazardous asteroids are found in APO, AMO, ATE, and IEO classes, with 27% of IEO asteroids being hazardous.
- Hazardous asteroids also exhibit higher average eccentricity, mean motion, and inclinations, whereas non-hazardous asteroids have higher average semi-major axis, minimum orbit intersection distance, and perihelion distance.

Conclusion

- The preprocessing and modeling steps effectively prepared the data, allowing for accurate predictions of hazardous asteroids.
- Among the models tested, Random Forest emerged as the most reliable, offering the highest precision and recall.
- This approach demonstrates the effectiveness of ensemble techniques in handling imbalanced datasets and accurately identifying potential threats.



- Hazardous asteroids are predominantly near-Earth objects and concentrated in specific orbital classes, non-hazardous asteroids generally have distinct orbital characteristics, with higher average distances and lower inclinations.