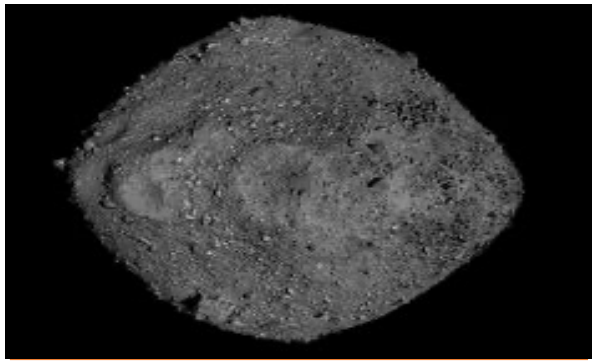
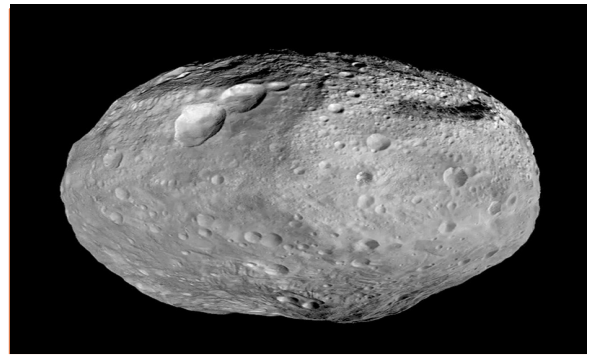
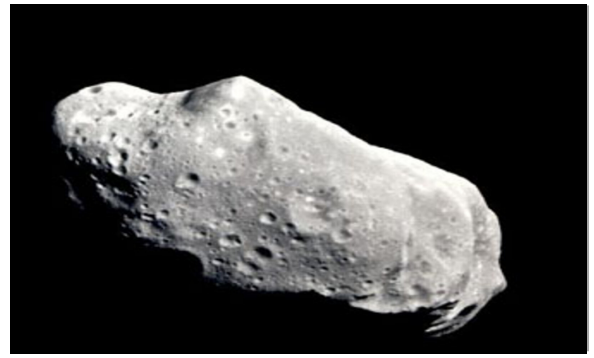
Potentially Hazardous Asteroids

Rachel Goodridge

NEO and Asteroid Facts

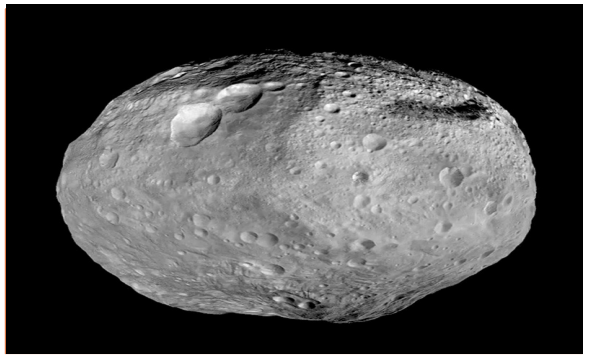



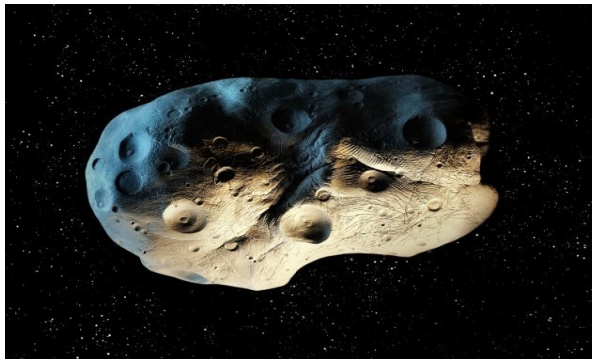




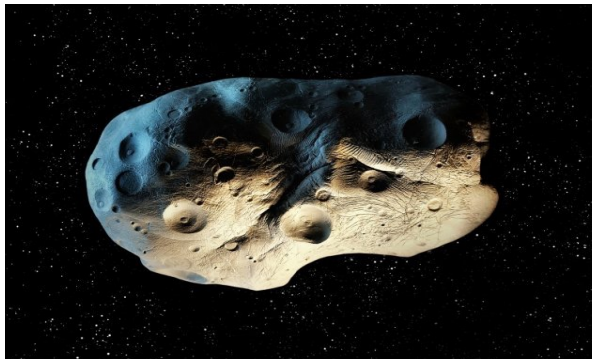
NEO and Asteroid Facts

Near-Earth Object (NEO)





NEO and Asteroid Facts

Near-Earth Object (NEO)	Orbit Path
	
	

NEO and Asteroid Facts

Near-Earth Object (NEO)

Orbit Path

Current Risk

NEO and Asteroid Facts

Near-Earth Object (NEO)

Orbit Path

Current Risk

Documentation

NEO and Asteroid Facts

Near-Earth Object (NEO)

Orbit Path

Current Risk

Documentation

Important Features

NEO and Asteroid Facts

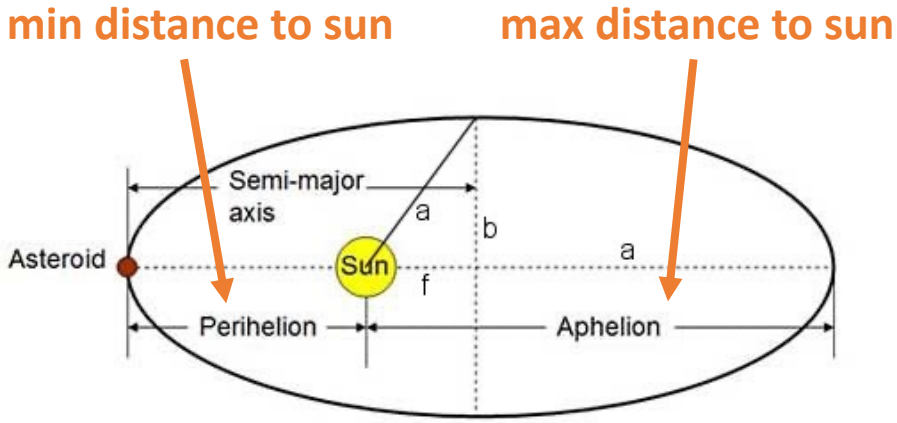
- Near-Earth Object (NEO)
- Orbit Path
- Current Risk
- Documentation
- Important Features
- Defense Strategies

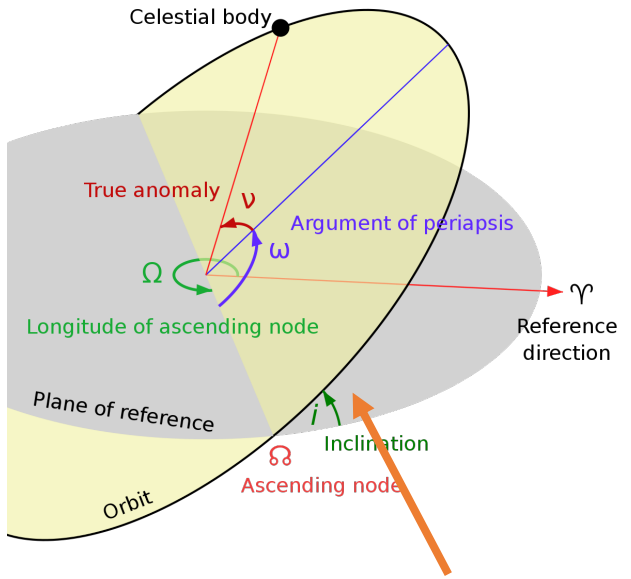
Objectives

- Predict which will be potentially hazardous asteroids (PHAs)
- Limit the number of PHAs on watch list
- Be prepared for possible threats

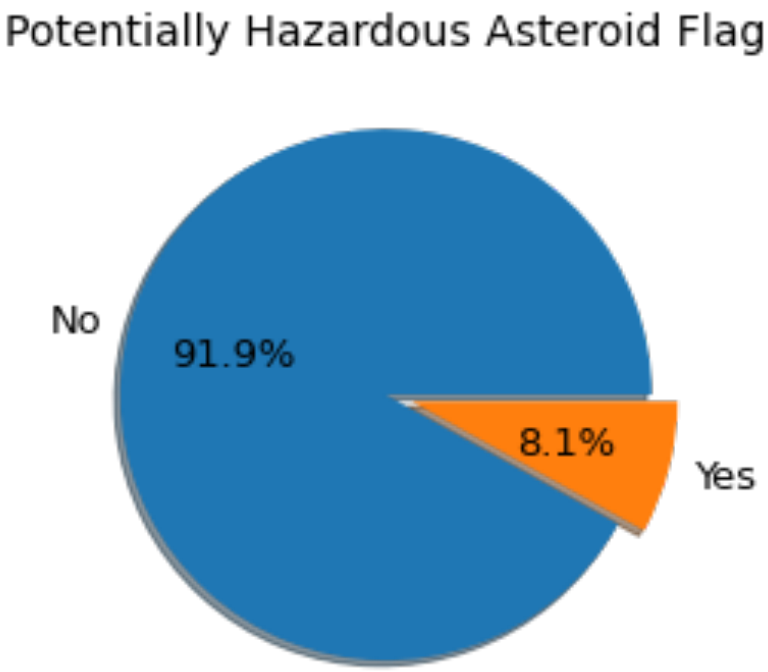
Data and Tools

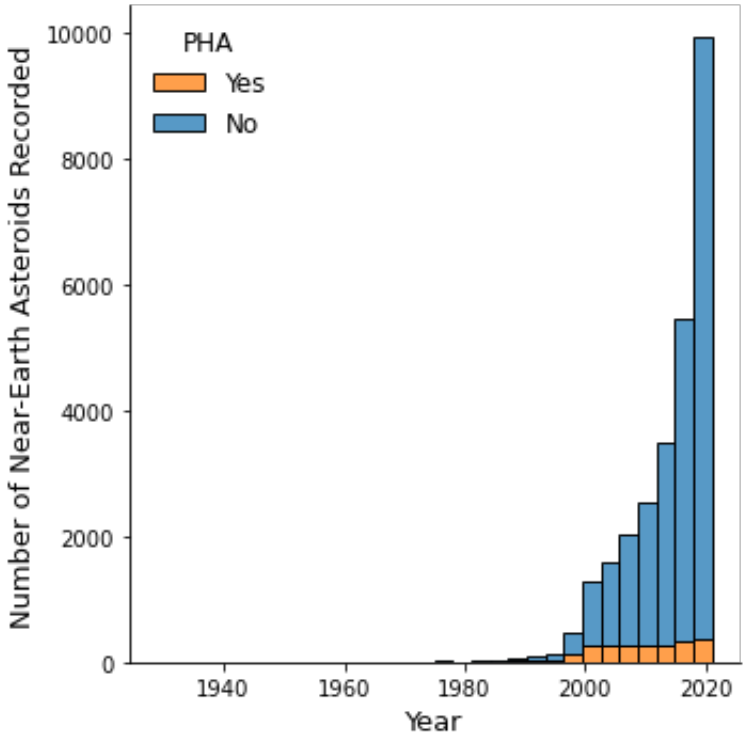
- Small-Body Database from NASA
- 27,000 rows and 14 features
- imbalanced data
- classification models from sklearn





Class Imbalance

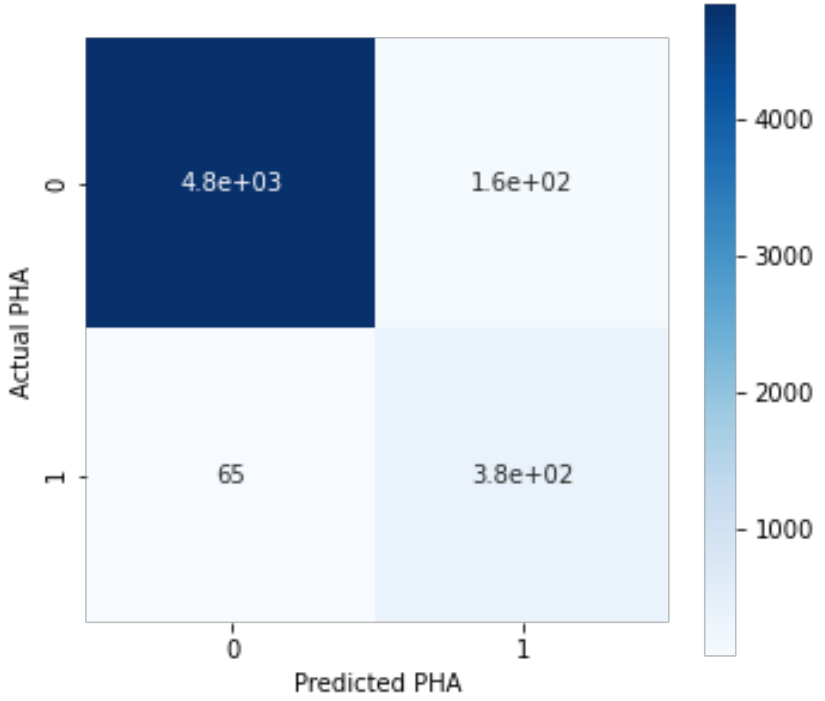


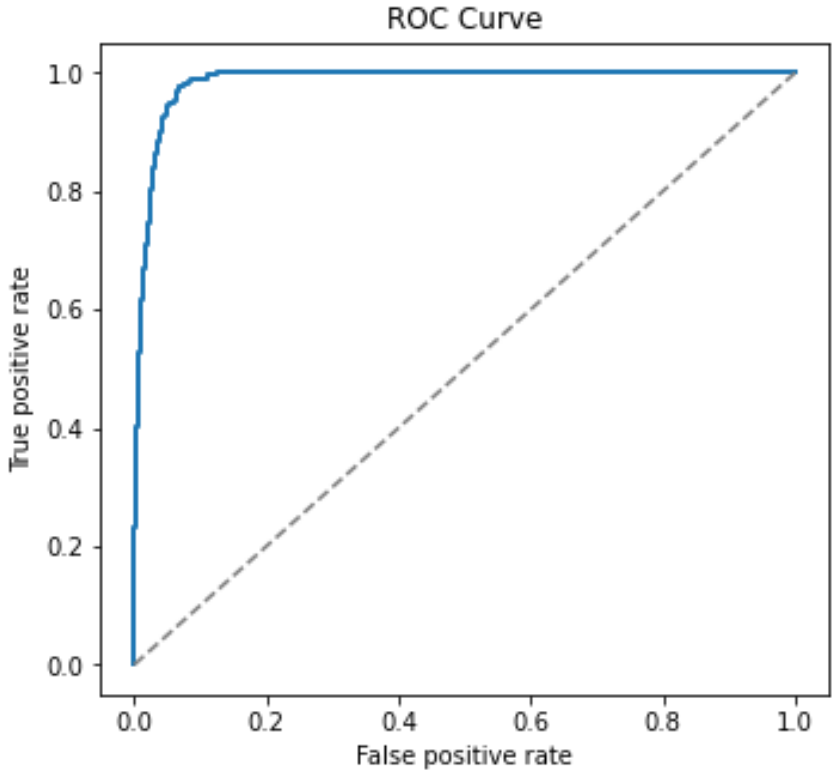


XGBoost Classifier

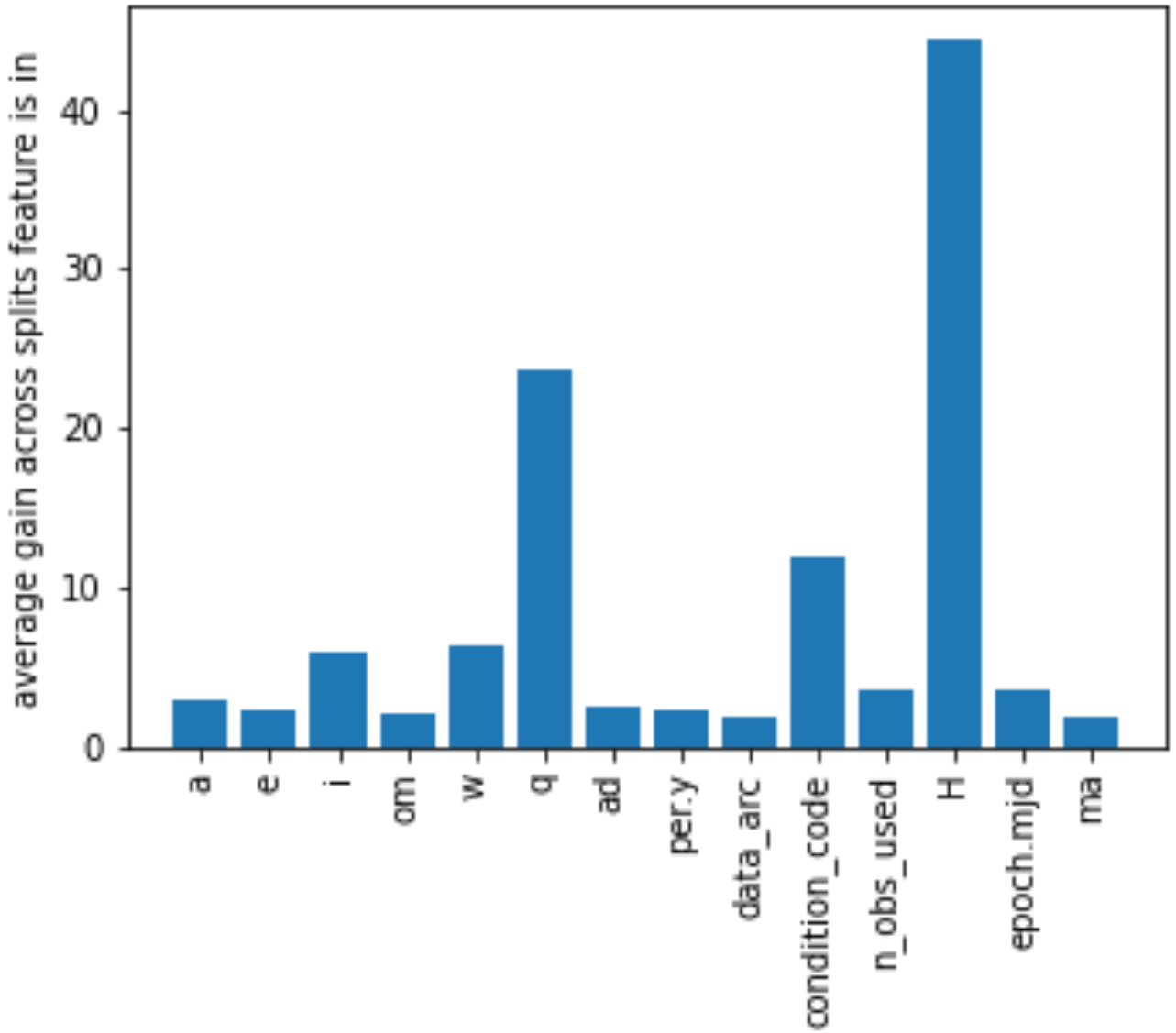
- Over sample with SMOTE
- Optimize many parameters

```
Training Scores
Accuracy : 0.9999074588191745
Precision : 0.9996963716411112
Recall : 1.0
F1 : 0.999848162769511
ROC AUC : 1.0

Validation Scores
Accuracy : 0.9594346549192364
Precision : 0.7078651685393258
Recall : 0.8532731376975169
F1 : 0.773797338792221
ROC AUC : 0.9848517171316268

Test Scores
Accuracy : 0.9625550660792952
Precision : 0.7361111111111112
Recall : 0.8393665158371041
F1 : 0.7843551797040168
ROC AUC : 0.9838750061012758
```

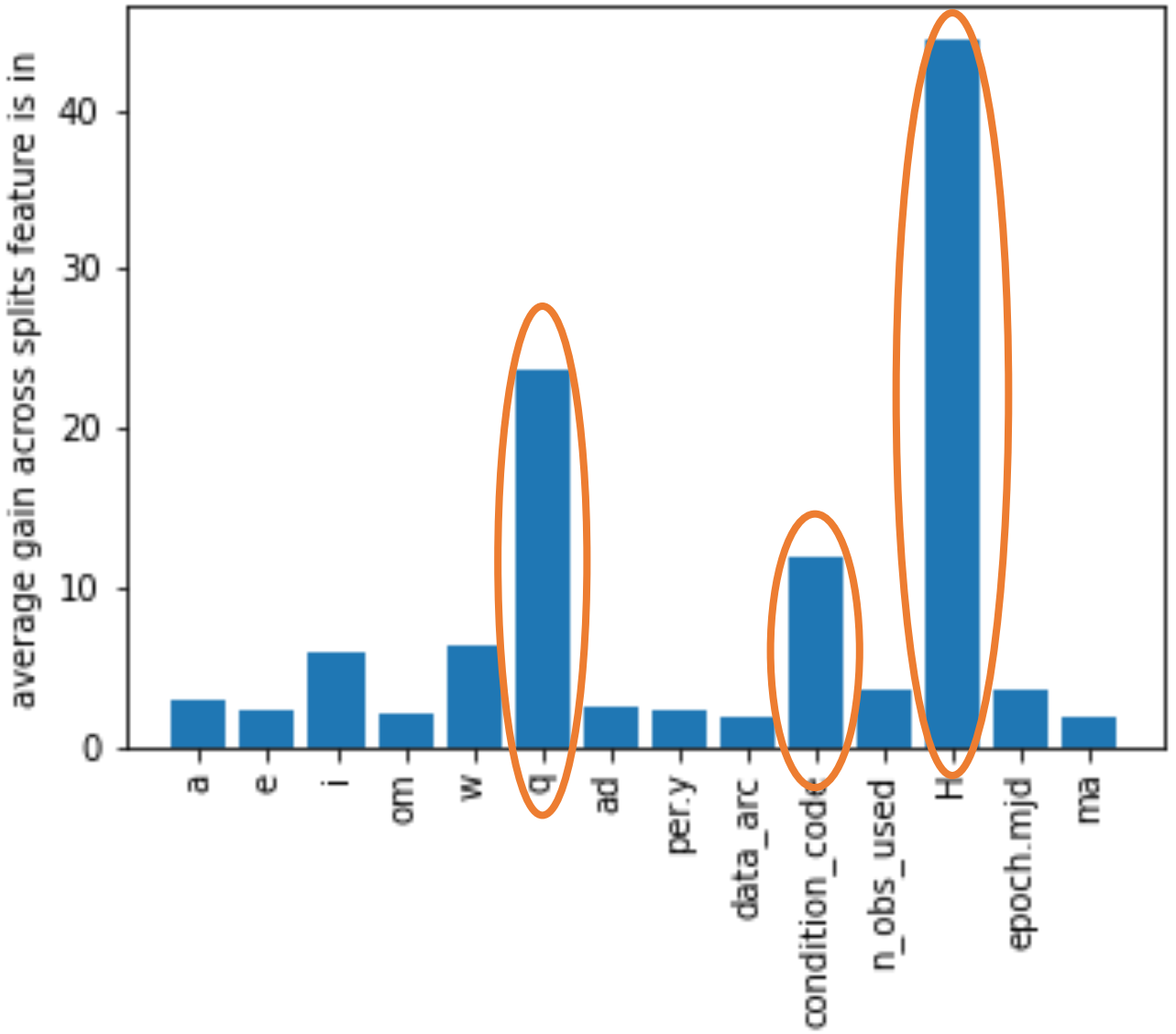




Feature Importance



Feature Importance



Conclusion

- XGBoost was the best-performing model with F1 test score of 0.784
- Features with highest importance include absolute magnitude (H), closest distance to the sun (q), and orbit uncertainty (condition_code).

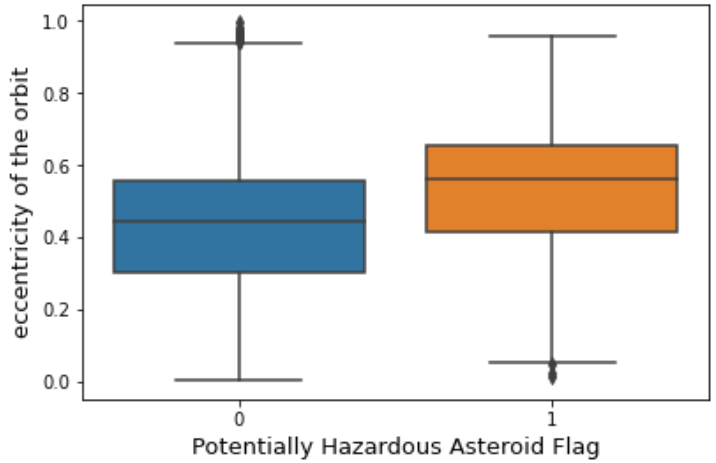
Thank you!

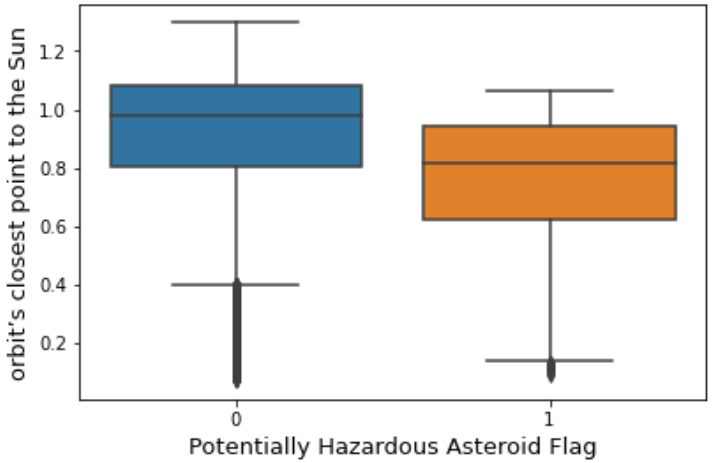
Appendix

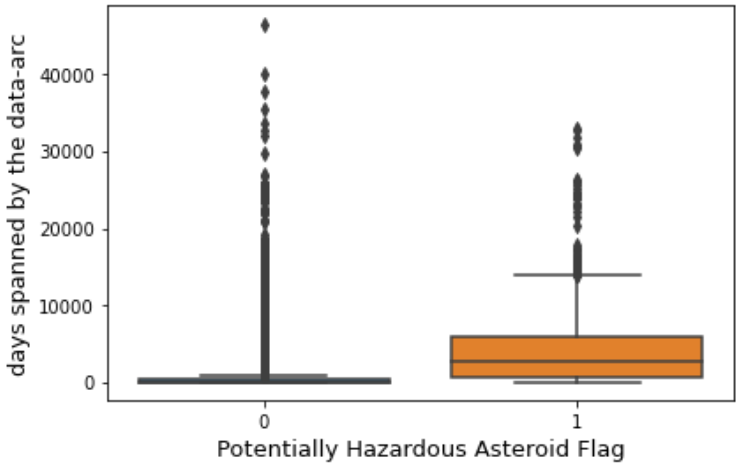
Data Info

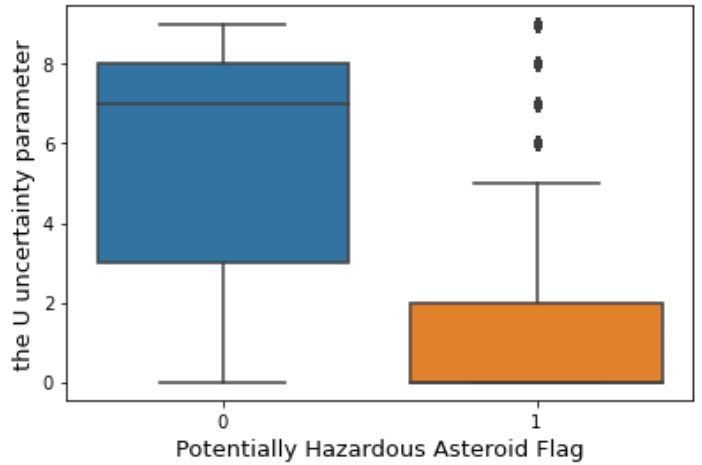
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27240 entries, 0 to 27239
Data columns (total 17 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   full_name       27240 non-null  object 
 1   a               27240 non-null  float64
 2   e               27240 non-null  float64
 3   i               27240 non-null  float64
 4   om              27240 non-null  float64
 5   w               27240 non-null  float64
 6   q               27240 non-null  float64
 7   ad              27240 non-null  float64
 8   per.y           27240 non-null  float64
 9   data_arc        27240 non-null  float64
 10  condition_code  27240 non-null  float64
 11  n_obs_used      27240 non-null  int64  
 12  H               27240 non-null  float64
 13  epoch.mjd       27240 non-null  int64  
 14  ma              27240 non-null  float64
 15  pha             27240 non-null  int64  
 16  year            27232 non-null  float64
dtypes: float64(13), int64(3), object(1)
memory usage: 3.5+ MB
```

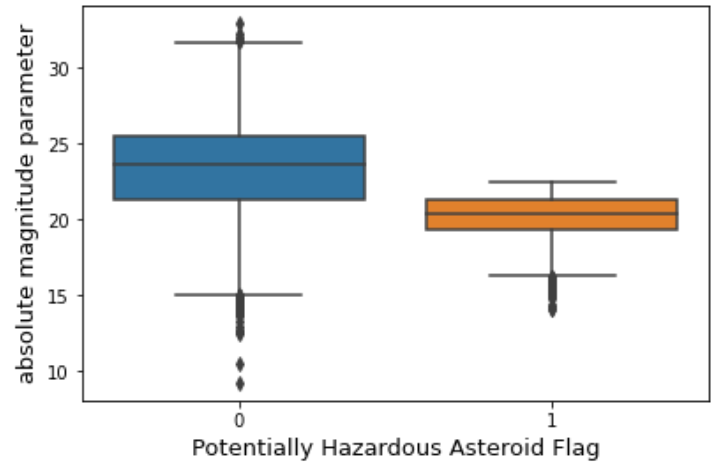
Feature Box Plots



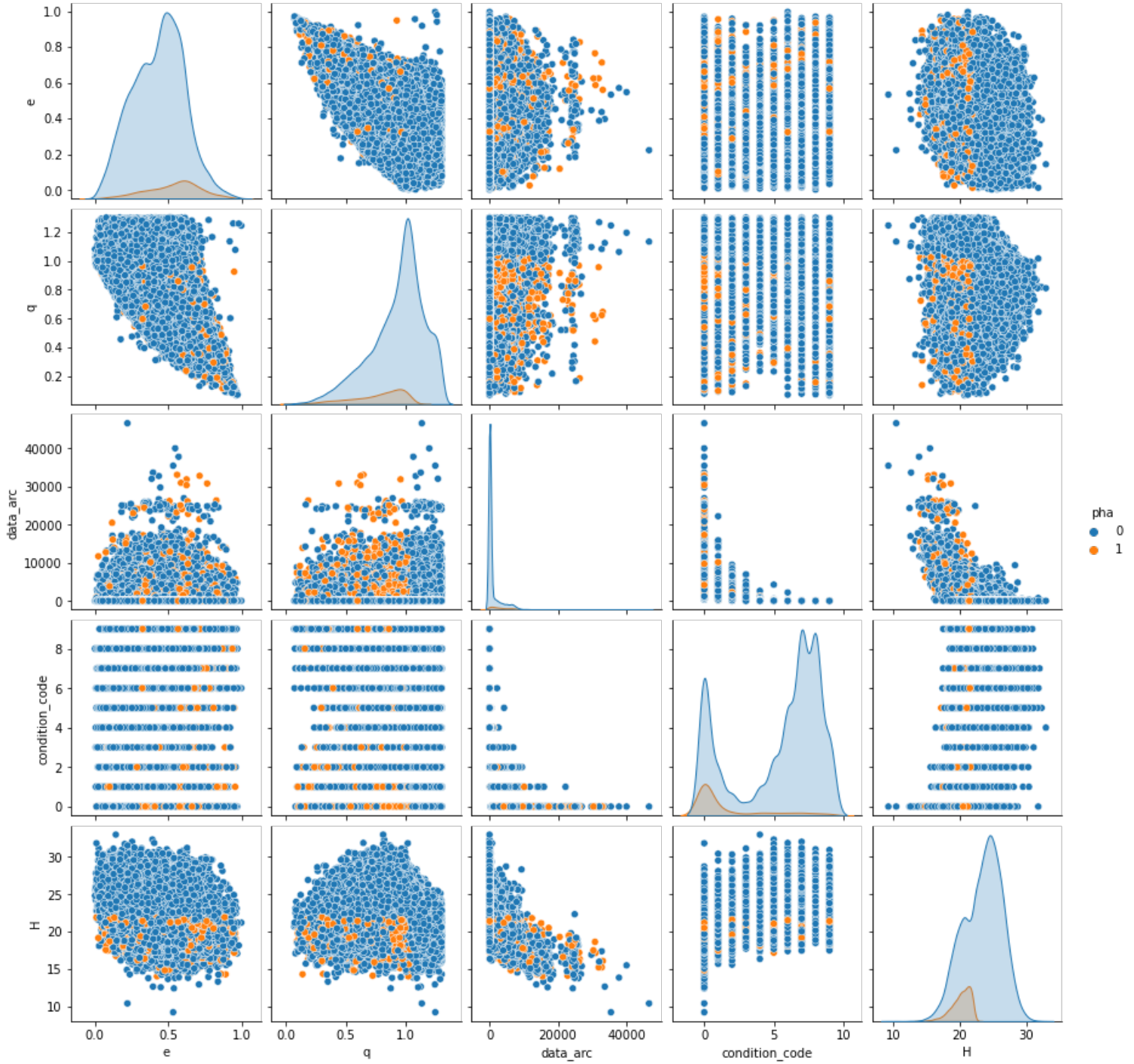








Pairplot



List of Models and Methods Tried

Methods

- Over sampling
- Over sampling with SMOTE
- Optimizing many parameters
- Hard and soft predictions
- Soft cutoff threshold optimization

Models

- K-Nearest Neighbors
- Logistic Regression
- Decision Tree
- Bagging Classifier
- Random Forest
- Extra Trees
- AdaBoost Classifier
- Gradient Boosting
- XGBoost Classifier
- Voting Classifier
- Stacking Classifier
- Gaussian Naïve Bayes

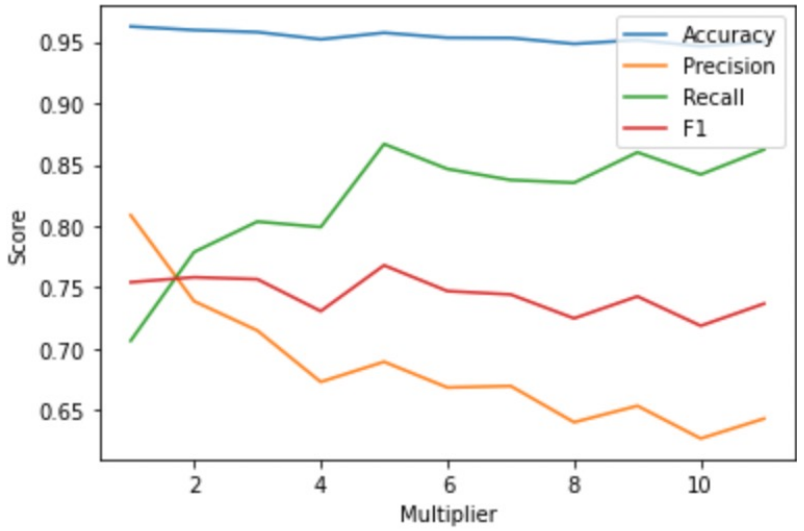
Model Scores

Model Name	F1 Validation Score
K-Nearest Neighbors	0.5047701647875108
Logistic Regression	0.4351407000686342
Decision Tree	0.7004048582995952
Bagging Classifier	0.728382502543235
Random Forest	0.635477582846004
Extra Trees	0.5936329588014981
AdaBoost Classifier	0.6473429951690821
Gradient Boosting	0.7417519908987485
XGBoost Classifier	0.773797338792221
Voting Classifier	0.7227615965480042
Stacking Classifier	0.6675461741424803
Gaussian Naïve Bayes	0.3392052437525604

XGBoost Parameter Optimization

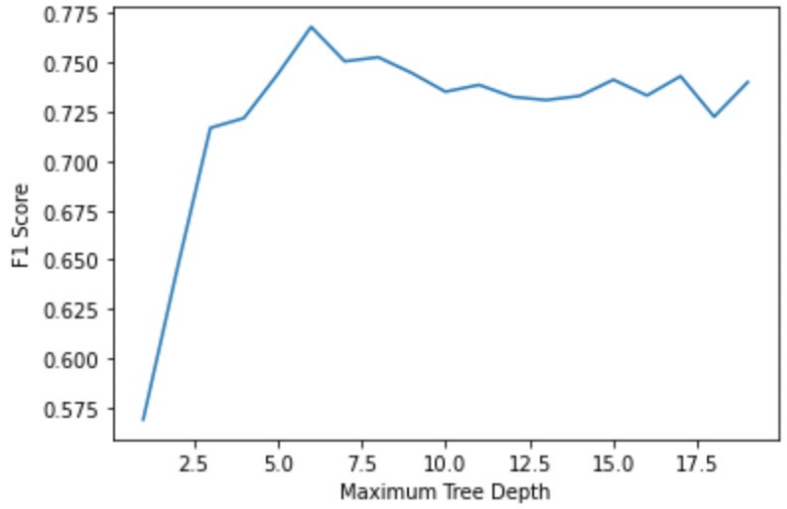
Oversampling with SMOTE

```
Max F1 : 0.768 at Multiplier 5
```



Maximum Tree Depth

```
Max F1 : 0.768 at Tree Depth 6
```



Soft Cutoff Threshold

```
ax F1 : 0.773797338792221 at Threshold 0.5
```