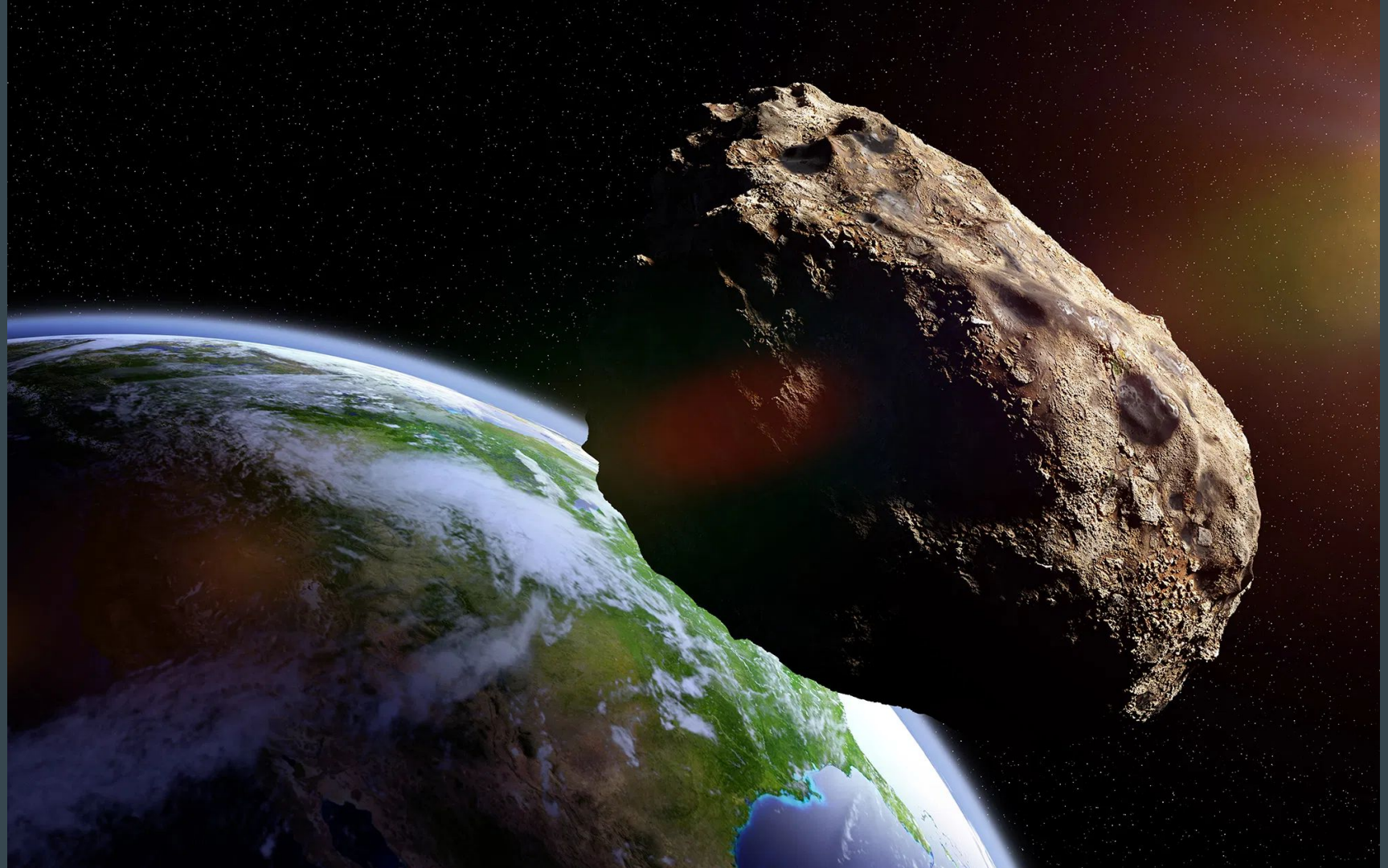
NASA: Asteroids Classification

by DATABUSTERS



Why did we chose NASA, Asteroids Classification dataset?

- Asteroids help astronomers trace solar system evolution
- Asteroids help astronomers understand processes in an evolving solar system
- Some asteroids may be hazards to Earth
- Essentially, asteroids were the building blocks of planets

DATASET

4687 rows / 40 columns / no missing values / no null values

```
1 df_raw = pd.read_csv("nasa.csv")
2 df_raw
```

	Neo Reference ID	Name	Absolute Magnitude	Est Dia in KM(min)	Est Dia in KM(max)	Est Dia in M(min)	Est Dia in M(max)	Est Dia in Miles(min)	Est Dia in Miles(max)	Est Dia in Feet(min)	...	Asc Node Longitude	Orbital Period	Perihelion Distance
0	3703080	3703080	21.600	0.127220	0.284472	127.219879	284.472297	0.079051	0.176763	417.388066	...	314.373913	609.599786	0.808259
1	3723955	3723955	21.300	0.146068	0.326618	146.067964	326.617897	0.090762	0.202951	479.225620	...	136.717242	425.869294	0.718200
2	2446862	2446862	20.300	0.231502	0.517654	231.502122	517.654482	0.143849	0.321655	759.521423	...	259.475979	643.580228	0.950791
3	3092506	3092506	27.400	0.008801	0.019681	8.801465	19.680675	0.005469	0.012229	28.876199	...	57.173266	514.082140	0.983902
4	3514799	3514799	21.600	0.127220	0.284472	127.219879	284.472297	0.079051	0.176763	417.388066	...	84.629307	495.597821	0.967687
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4682	3759007	3759007	23.900	0.044112	0.098637	44.111820	98.637028	0.027410	0.061290	144.723824	...	164.183305	457.179984	0.741558
4683	3759295	3759295	28.200	0.006089	0.013616	6.089126	13.615700	0.003784	0.008460	19.977449	...	345.225230	407.185767	0.996434
4684	3759714	3759714	22.700	0.076658	0.171412	76.657557	171.411509	0.047633	0.106510	251.501180	...	37.026468	690.054279	0.965760
4685	3759720	3759720	21.800	0.116026	0.259442	116.025908	259.441818	0.072095	0.161210	380.662441	...	163.802910	662.048343	1.185467
4686	3772978	3772978	19.109	0.400641	0.895860	400.640618	895.859655	0.248946	0.556661	1314.437764	...	187.642183	653.679098	0.876110

4687 rows × 40 columns

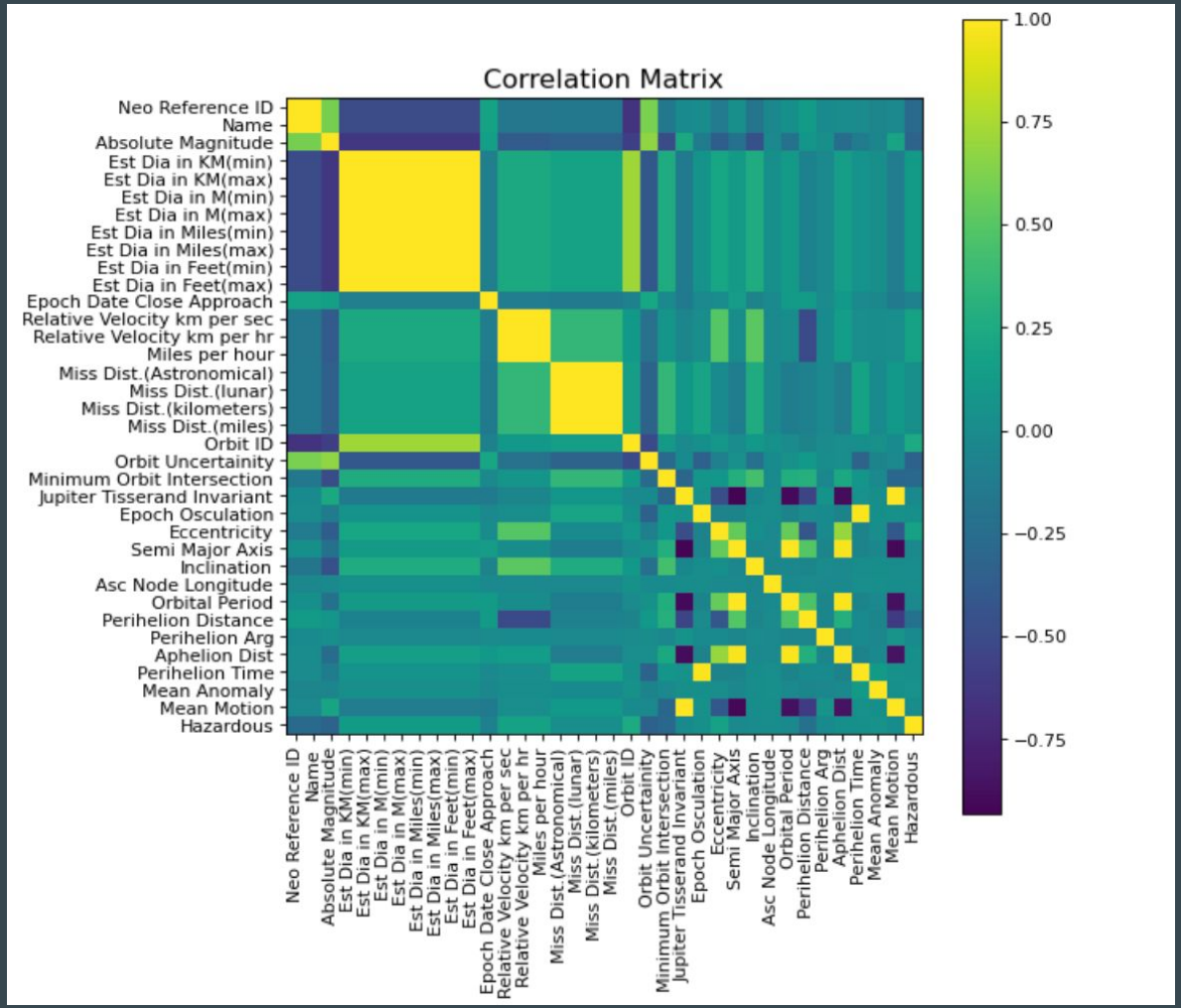
Columns

- Neo Reference ID
- Name
- Absolute Magnitude
- Estimated Diameter of Asteroid
- Close Approach Date
- Epoch Date Close Approach
- Relative Velocity
- Miss Distance
- Orbit ID
- Orbit Determination Date
- Orbit Uncertainty
- Minimum Orbit Intersection
- Equinox
- Orbiting Body
- Jupiter Tisserand Invariant
- Eccentricity
- Inclination
- Asc Node Longitude
- Semi Major Axis
- Orbital Period
- Perihelion Distance & Time
- Aphelion Distance
- Mean Anomaly
- Mean Motion
- **Hazardous**

EDA

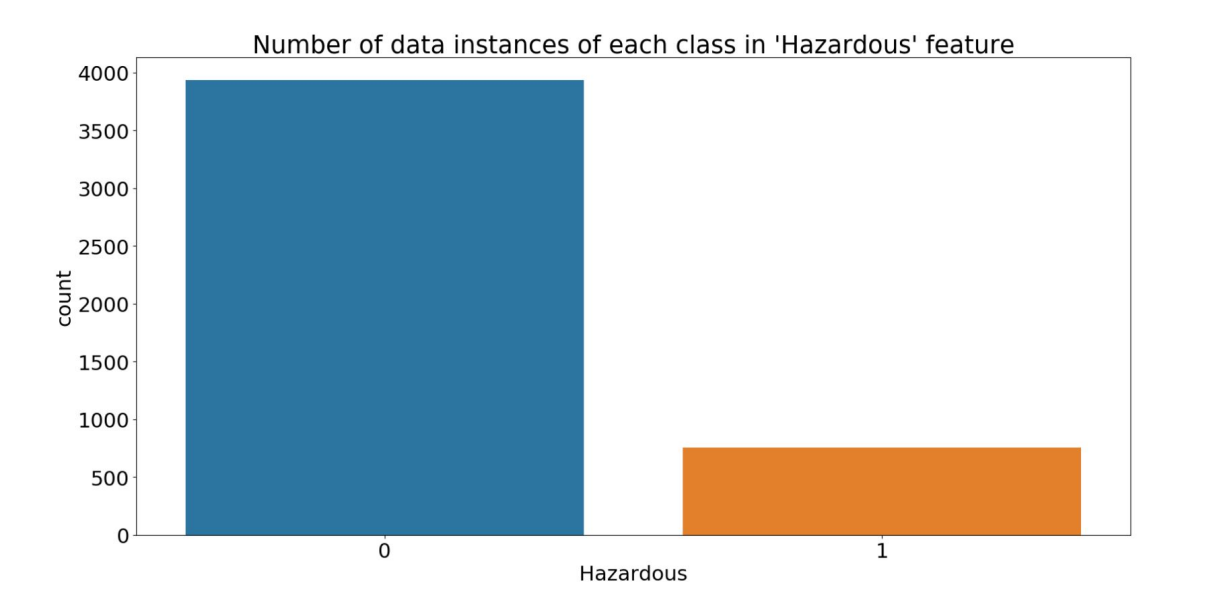
- Checking null values
- Checking histograms
- Correlation matrix
- Checking the columns with data type “object”
- Encoding/ Standardization

Correlation Matrix



Imbalanced dataset

- Random Under - Sampling
- Random Over - Sampling



Source: https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/

Preprocessing

- **Removing columns:** / duplicated / different metrics / dates

['Orbiting Body', 'Equinox', 'Est Dia in M(min)', 'Est Dia in M(max)', 'Close Approach Date', 'Est Dia in Miles(min)','Est Dia in Miles(max)', 'Est Dia in Feet(min)', 'Est Dia in Feet(max)', 'Miss Dist.(Astronomical)', 'Miss Dist.(lunar)', 'Miss Dist.(miles)', 'Orbit Determination Date', 'Neo Reference ID', 'Name', 'Orbit Uncertainty'])

- **Normalization:** encoding(df, 'Hazardous', LabelEncoder)

- **Unbalanced dataset:**

  rus = RandomUnderSampler()

  ros = RandomOverSampler()

Models

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. XGBoost
5. Bayes
6. SVM
7. KNN

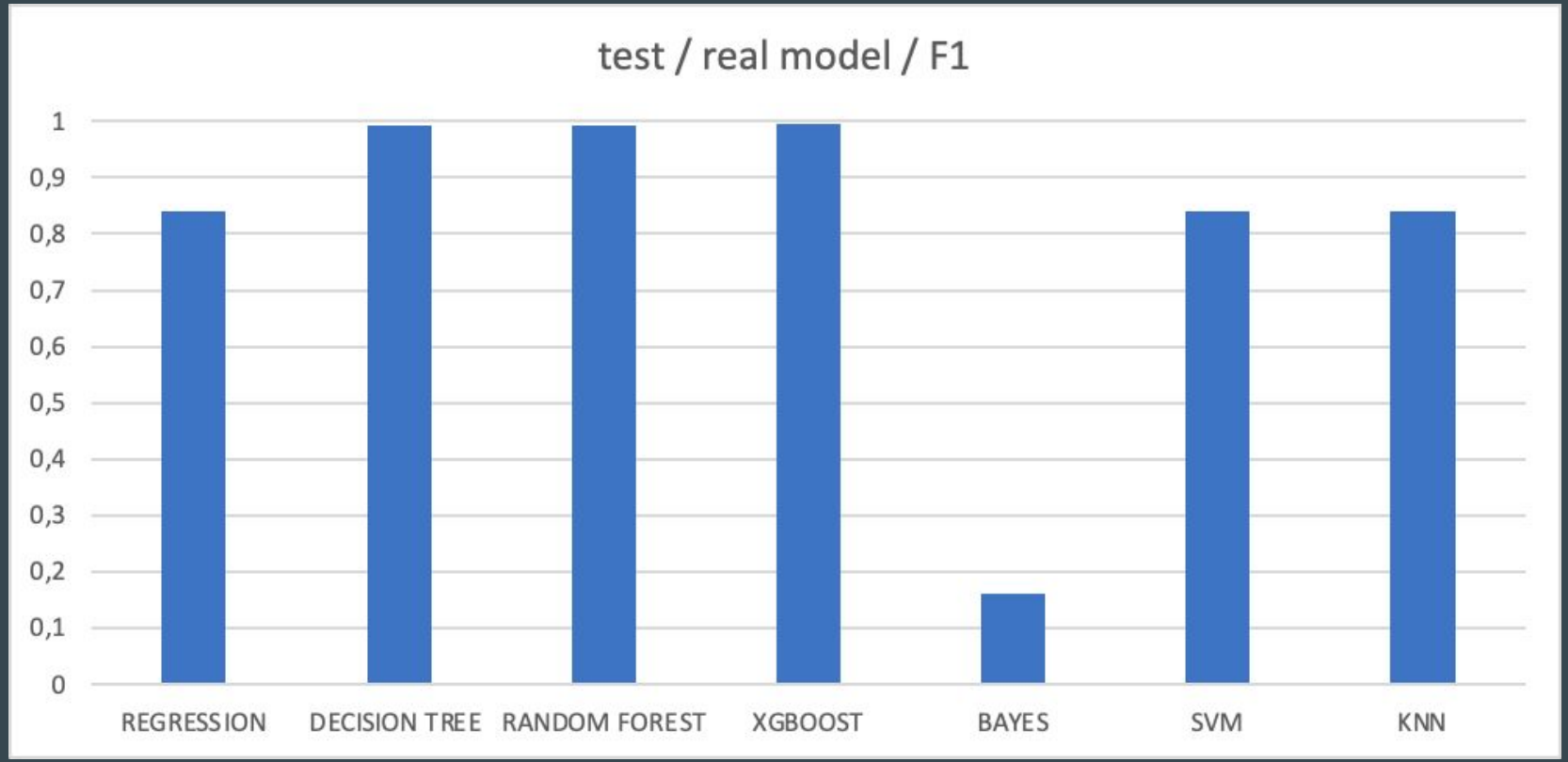
Metrics

1. Accuracy is not the best metric to use when evaluating imbalanced datasets, as it can be misleading

   / Metrics that can provide better insights are: /

2. Precision: the number of true positives divided by all positive predictions. Low precision indicates a high number of false positives
3. Recall: the number of true positives divided by the number of positives values in the test data. Low recall indicates a high number of false negatives

Metrics - comparison / real model



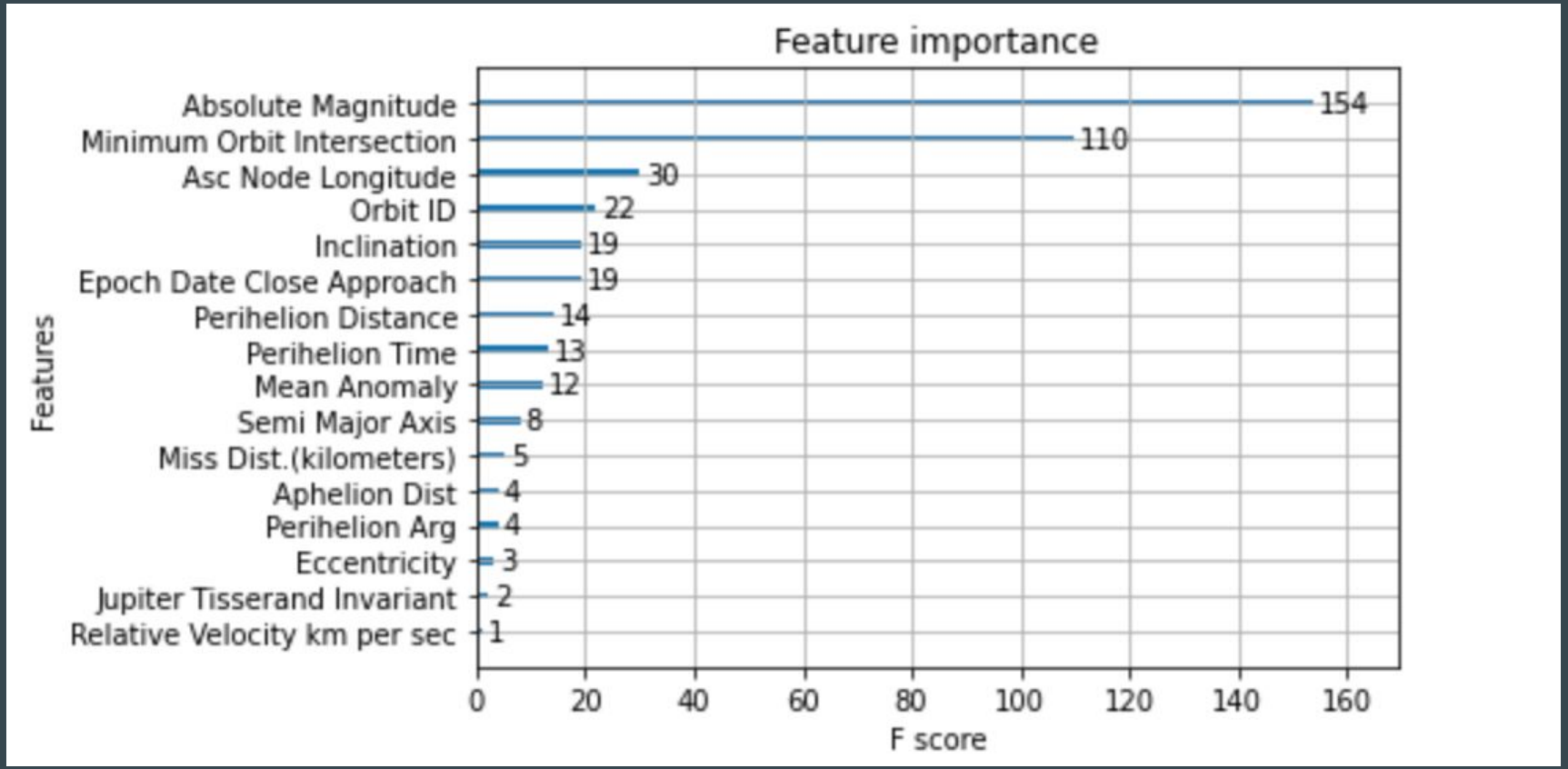
Basic model - no hyperparameters

MODEL:	TEST		TRAIN	
/ metric /	accuracy	F1	accuracy	F1
REGRESSION	0.8389	0.8389	0.8389	0.8389
DECISION TREE	0.9938	0.9938	0.9991	0.9991
RANDOM FOREST	0.9949	0.9949	0.9989	0.9989
XGBOOST	0.9951	0.9951	0.9989	0.9989
BAYES	0.8389	0.8389	0.8389	0.8389
SVM	0.8389	0.8389	0.8389	0.8389
KNN	0.8201	0.8201	0.8430	0.8430

Real model - /w hyperparameters

MODEL:	TEST		TRAIN	
/ metric /	accuracy	F1	accuracy	F1
REGRESSION	0.8389	0,8389	0.8390	0.8389
DECISION TREE	0.9933	0,9934	0.9953	0.9953
RANDOM FOREST	0.9938	0,9938	0.9993	0.9993
XGBOOST	0.9951	0,9951	0.9989	0.9989
BAYES	0.1611	0,1611	0.1611	0.1611
SVM	0.8389	0,8389	0.8389	0.8389
KNN	0.8389	0,8389	0.8387	0.8387

Feature importance



Thank you! Please try our APP!