Near-Earth Objects Predictions For Being Hazardous

Name - Pooja Mule, CWID - 20016077
Name - Anirudh Chintha, CWID - 20016080
Name - Teshwani Gogineni, CWID - 20012016

Goal

The aim of this project is to predict whether an object passing in the vicinity of the earth should be treated as hazardous based on the dataset provided.

Accuracy, Precision, Recall, F1 score

Model	# Features	Precision		Recall		F1 Score		Support		C	Gamma
		TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE		
GridSearchCV	4	0.92	0.85	1	0.14	0.96	0.24	16329	1668	NA	NA
SVC	4	0.92	0.85	1	0.14	0.96	0.24	16329	1668	100	1
GridSearchCV with PCA	2	0.92	0.85	1	0.14	0.96	0.24	16329	1668	NA	NA
SVC with PCA	2	0.92	0.85	1	0.14	0.96	0.24	16329	1668	100	1
SVC with PCA and downsampling	2	0.99	0.83	0.75	1	0.86	0.9	1567	1879	100	1

Approach/ Design Choices

1. Approach -
    a. Process data -
        i. After analyzing the pairplots, we have **dropped** the features which are either constant or do not add any significance to predictions namely - id, name, sentry object, and orbiting body, relative velocity
        ii. **Features utilized** - est_diameter_min, est_diameter_max, miss_distance, and absolute_magnitude
    b. Data transformation and handling outliers -
        i. The data is normalized using the z-score table.
        ii. We tried different methods to handle **outliers** like **replacing** with percentiles and medians before removing it completely.
        iii. We realized **removing outliers** provides **better performance** than replacing them. Total of **2.94**% data was classified as outliers.
    c. Model selection -
        i. After analyzing the data it was understood that the structure of data is **non-linear**.
        ii. The **Support Vector Machine** model would **perform well** in this case since it **uses kernel tricks** to perform more efficiently. Hence, the SVM model is picked for the purpose of these predictions.

iii. After doing grid search and providing the C and gamma values, it was found that predictions did not change after tuning these parameters with the model.

d. Principal Component Analysis -
    i. The top 2 components returned by PCA are **miss distance** and **absolute magnitude.**
    ii. The model predictions **remained the same** after performing PCA.
e. Optimizations -
    i. When we downsampled, the prediction of the case **False Positive** (Model predicted object is hazardous but it actually is not) increased from 0.2% to 11%. But the case of **True Negative** (Model predicted object is not hazardous but it actually is) reduced dramatically from 7% to 0.1%.
    ii. So the downsampling actually is better since we do not want to falsely predict that object is **not** hazardous when it **actually is**.