Optimizing feature processing and selection on Nasa Asteroid Dataset

Shankul Shukla
Machine Learning Intern
AI Technology and Systems
shuklashankul@gmail.com
www.ai-techsystems.com

***Abstract* – This analysis deals with the application of different feature processing and feature selection techniques to take out more than 99% accuracy on the Nasa Asteroid dataset based on various machine learning techniques. The best approach to deliver better accuracy on the dataset is being mentioned in the paper with comparative analysis with other used approaches.**

I. Introduction

In recent days the technology advancements have ordered as to deliver long term and effective solution to the more complex problems such as medical, physics and astronomy. Everyone is fascinated by looking into the dark sky in the night and look onto stars and asteroids passing by and we usually think if one of the asteroids hit earth what would happen as long back similar theory has been given in the context of the dinosaur extinction.

So, in this research paper, we are dealing with the NASA asteroid dataset which is created using NeoWs (Near-Earth Object Web Service) which is a RESTful web service for near-earth Asteroid information. The dataset has features responsible for claiming an asteroid to be hazardous or non-hazardous. Without any processing on features, there are 39 features given which include features like Absolute Magnitude of the asteroid, Relative Velocity km per sec, Epoch Osculation, Neo Reference ID, name, Close Approach Date, etc. Some of the features are interesting to observe and others carry no significant information to classify an asteroid as hazardous which is discussed in the latter part of the paper.

In this paper, this dataset is being used to study the data pre-processing techniques, feature selection, and model generation methods. In the process, we look into various feature selection algorithm and technologies like regularization [1], Sequential feature selection algorithms [2], some ensemble-based feature selection algorithms [3]. Each of the algorithms has been implemented and studied to find the best suited for the use case. Several machine learning models to learn meaningful insights from the dataset is also being implemented and tested to get the accuracy on the test set as high it could be.

II. Feature Pre-processing

Multiple features are available in the dataset which is of different types like categorical, continuous and date-time features. Creating a good machine learning model, we want each selected features to work in accordance to provide a classifier which can perform best given those features. As our goal is always to understand the distribution of the data in our model and avoiding the model to understand any random noise in the data.

As the value of features Orbiting Body and Equinox is singleton so reflect no information so I removed these features. I plotted scatter plots to understand to whether does the name or Neo Reference ID are somehow related to classify an asteroid as hazardous that there is a difference in the data distribution *(Figure 1)* so these features are also being removed. We can see there is no relation among the date discovered values as we cannot feed simple date values into the model we have to process with some reference but no specific reference feature can be observed in the analysis so feature showing date-time attributes are also being removed.

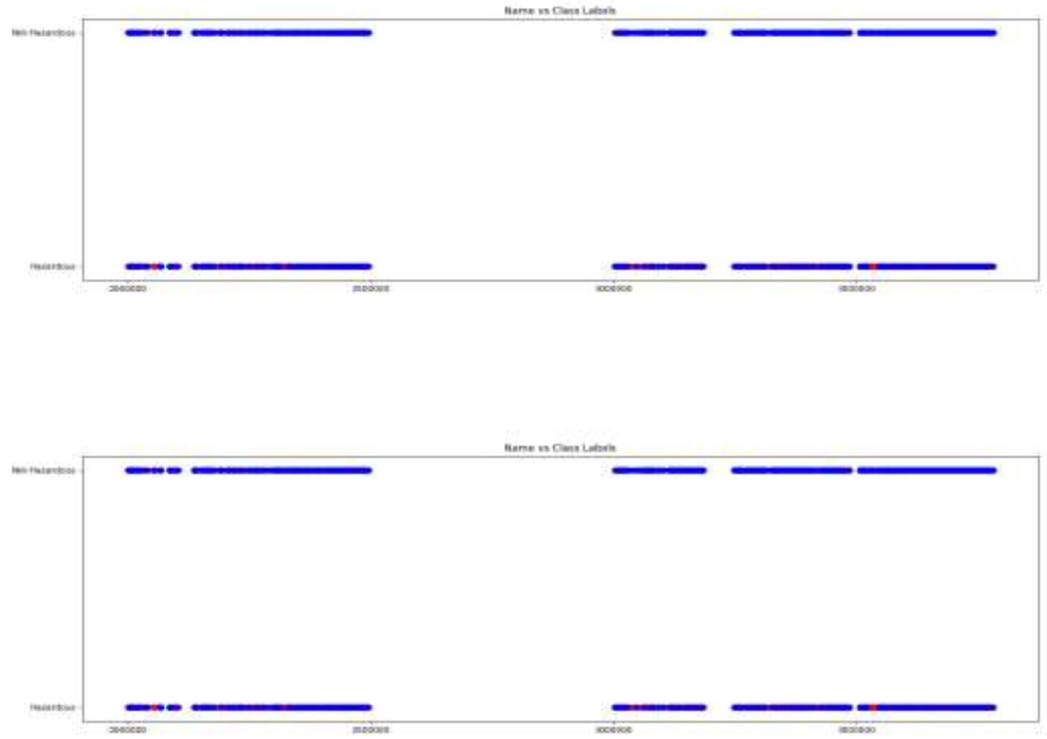
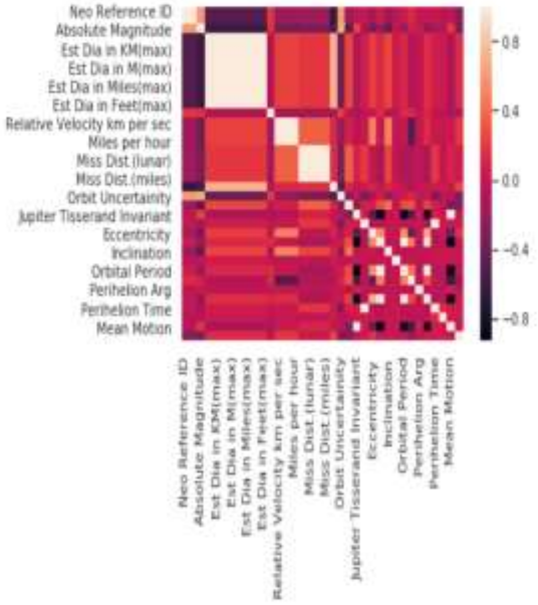


Figure 1

We can observe that the Est diameter in KM, M Miles, feets as well as Relative velocity in KM/hr, km/sec and miles/hr, these features are completely correlated to each other, similar observation can be found out in the miss distance in AU, Lu, KM, Miles, so we are going to remove the redundant features and take only the Est dia in km., relative velocity in km/hr and estimated distance in km as a representative for the combination of the features. We will test if there any data values having null values so we can apply some interpolation or remove data values but there are no found null values in the dataset. As the Est Dia in KM(min) and (max) are two redundant feature so we are combining them to create a single feature that is avg_dia which will hold an average of two and remove other two features. We have observed the variance of each of the feature no feature has found to have low variance so that we can remove them.

III. Feature Selection

In this paper we started the feature selection by finding correlation among the features, it tells us the relationship between variables. We found out that several features are not at all related to each other, some are too much related like Est Dia In miles and other distances are highly correlated so we try to combine them and remove the redundant ones. Plot showing correlation matrix among features-



In the process of finding the best features for the prediction model, we tried several sequential feature selection algorithms [2] which transform d dimensional features to desired k dimensional features. We tried out the

Sequential backward selection approach [4], features recommend by the approach does not provide good accuracies over the dataset and were quite computationally expensive to implement.

The best approach according to analysis which gives best results was finding feature importance with random forests [3]. The ensemble method used Gini importance [5] values to extract out most prominent features, it is defined as a total decrease in node impurity averaged over all trees of the ensemble. The feature importance values are being shown in the table based on the approach -

1) Minimum Orbit Intersection	0.457077
2) avg_dia	0.139554
3) Absolute Magnitude	0.118470
4) Orbit ID	0.052274
5) Perihelion Distance	0.036931
6) Orbit Uncertainity	0.036709
7) Inclination	0.034942
8) Miss Dist.(kilometers)	0.020437
9) Eccentricity	0.013865
10) Relative Velocity km per hr	0.012724
11) Aphelion Dist	0.011906
12) Perihelion Arg	0.010393
13) Orbital Period	0.008560
14) Jupiter Tisserand Invariant	0.008353
15) Mean Motion	0.008296
16) Mean Anomaly	0.007673
17) Semi Major Axis	0.007634
18) Asc Node Longitude	0.006776
19) Perihelion Time	0.006256
20) Epoch Osculation	0.001172

Bringing all features to the same scale is very important to have a consistent model so a standard scaler is being used to scale out the training and test set.

IV. *Training*

We divided our dataset into train and test set with a test set having 25% of the data and an equal proportion of class as that in training set. We tried out with Support vector classifier, logistic regression classifier, decision tree classifier, and random forest classifier, the results are being shown onto the table in the result.

In the support vector classifier, I have used the RBF kernel with gamma value 0.2 and penalty parameter C as 1.0. For the logistic regression classifier, chosen values of C (inverse of regularization strength) is 100.0 and solver being lbfgs as it handles L2 or no penalty. For the random forest classifier, we have used 5 estimators with gini as impurity criterion. For the decision tree classifier, we have restricted its depth to 8 and used the gini impurity criterion.

V. *Result*

In the analysis, we found out that most prominent features among all the features that best results in any of the ML algorithm used these features are -Minimum Orbit Intersection, avg_dia, Absolute Magnitude, Orbit ID, Perihelion Distance, Orbit Uncertainty, Inclination, and Miss Dist. (kilometers). These features lead out to result shown in in the table below-

ML algorithm	Accuracy (%) in test set
Support Vector Classifier	93
Logistic Regression Classifier	94.7
Radom forest classifer	99.65
Decision tree classifier	99.65

Random forest is an ensemble of decision tree but showing similar accuracies depicting that we do not want an ensemble of decision tree estimators. So, in the analysis Decision tree classifier provides the best result for the Nasa Asteroid Dataset.

VI. *Reference*

[1] A. Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In ICML, 2004.

[2] Miao, Jianyu & Niu, Lingfeng. (2016). A Survey on Feature Selection. Procedia Computer Science.

[3] Akhiat, Yassine & Chahhou, Mohamed & Zinedine, Ahmed. (2019). Ensemble Feature Selection Algorithm. International Journal of Intelligent Systems and Applications.

[4] Čehovin, Luka & Bosnic, Zoran. (2010). Empirical evaluation of feature selection methods in classification. Intell. Data Anal.

[5] Gupta, B., P. Uttarakhand, I. A. Rawat. Analysis of Various Decision Tree Algorithms for Classification in Data Mining. – International Journal of Computer Applications, Vol. 163, 2017, No 8, pp. 975-987.