Hazardous Asteroid Prediction

Kayla Thurman

May 2024

Click Here For Overleaf Report

Contents

1 Abstract

Near-Earth Objects

Near-Earth Objects (NEOs) are comets and asteroids whose orbits have been affected by the gravitational pull of nearby planets that bring them close to Earth.[1] The NASA Jet Propulsion Laboratory defines an NEO as an asteroid or comet whose approach brings it within 1.3 astronomical units (au) of the sun. An astronomical unit is the average distance between the Earth and the sun, approximately 150 million kilometers or 93 million miles.[2] These NEOs could serve as sources of raw materials for interplanetary explanation in the future. They can also potentially pose a threat to the Earth.

Physically Hazardous Asteroids

Among these Near-Earth Objects, there are some that could potentially drift into an orbit that puts them at risk of collision with the Earth. Near-Earth Objects that have a Earth Minimum Orbit Intersection Distance of 0.05 au or less with an absolute magnitude of 22.0 or less are considered Physically Hazardous Asteroids (PHAs).[1]

The problem of determining whether an NEO is physically hazardous is a good candidate for an intelligent agent because there are many factors that play into whether or not the object is hazardous. These objects' orbits also change over time and with influence from other celestial objects' gravitational fields. For this reason, it could be useful to have an intelligent agent that is able to determine early on whether or not an asteroid is potentially a PHA.

2 Data Analysis and Preparation

2.1 Dataset

The dataset, "NASA: Asteroids Classification" was retrieved from Kaggle.[3] It pulls data from the Near Earth Object Web Service, which is an API that processes data from NASA's Center for Near-Earth Object Studies' Near-Earth Object database. The dataset contains 4,687 samples, including 3,932 samples that are classified as non-hazardous and 755 samples classified positively as physically hazardous.

Originally, the dataset consisted of 40 columns in total: 39 input features with a single output column positively or negatively identifying the Near-Earth Object as Physically Hazardous. Input features that were non-numeric or names/identifiers that were not necessary for making a determination were removed, after which there were 33 input features remaining. For the sake of more control over the model in later phases of the project, the number of features was reduced to those listed as the determining factors in classifying PHAs according to NASA's Center for Near-Earth Object Studies, for a total of 9 input features with one binary output label:

1. Absolute magnitude
2. Minimum estimated diameter in meters
3. Maximum estimated diameter in meters
4. Miss distance in kilometers
5. Earth Minimum Orbit Intersection Distance
6. Semi-major axis
7. Orbital period
8. Perihelion distance
9. Aphelion distance
10. Hazardous

Below is a heatmap depicting the correlation of the input features and the output label (see Figure 1).

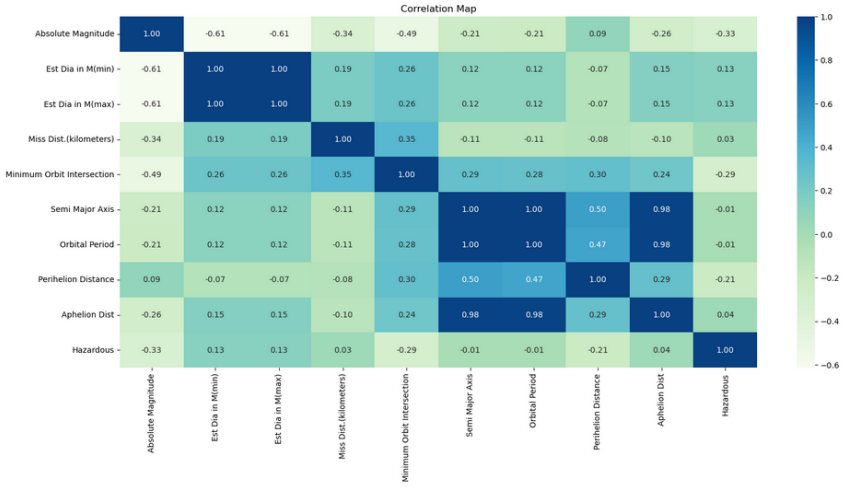

Figure 1: Dataset correlation heatmap

2.2 Input Feature Distributions

Below are histograms detailing the distribution of the input features for the dataset (see **Figure 2**), as well as a table containing key statistics for the input features (see **Figure 3**).

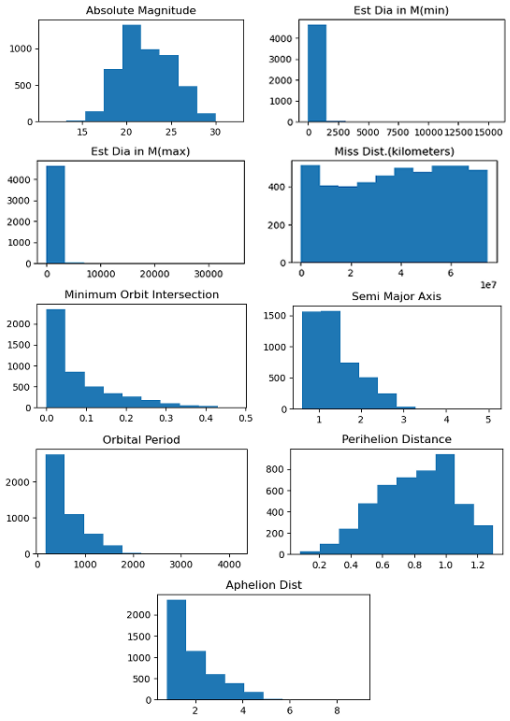

Figure 2: Input feature distribution histograms

	Absolute Magnitude	Est Dia in M(min)	Est Dia in M(max)	Miss Dist. (kilometers)	Minimum Orbit Intersection	Semi Major Axis	Orbital Period	Perihelion Distance	Aphelion Dist	Hazardous
count	4687.000000	4687.000000	4687.000000	4.687000e+03	4687.000000	4687.000000	4687.000000	4687.000000	4687.000000	4687.000000
mean	22.267865	204.604203	457.508906	3.841347e+07	0.082320	1.400264	635.582076	0.813383	1.987144	0.161084
std	2.890972	369.573402	826.391249	2.181110e+07	0.090300	0.524154	370.954727	0.242059	0.951519	0.367647
min	11.160000	1.010543	2.259644	2.660989e+04	0.000002	0.615920	176.557161	0.080744	0.803765	0.000000
25%	20.100000	33.462237	74.823838	1.995928e+07	0.014585	1.000635	365.605031	0.630834	1.266059	0.000000
50%	21.900000	110.803882	247.765013	3.964771e+07	0.047365	1.240981	504.947292	0.833153	1.618195	0.000000
75%	24.500000	253.837029	567.596853	5.746863e+07	0.123593	1.678364	794.195972	0.997227	2.451171	0.000000
max	32.100000	15579.552413	34836.938254	7.478160e+07	0.477891	5.072008	4172.231343	1.299832	8.983852	1.000000

Figure 3: Input feature statistics

2.3 Output Label Distribution

The output label is imbalanced heavily, with more samples labeled non-hazardous (0) than hazardous (1). This can be seen in the visual depicting the output label's distribution (see **Figure 4**).

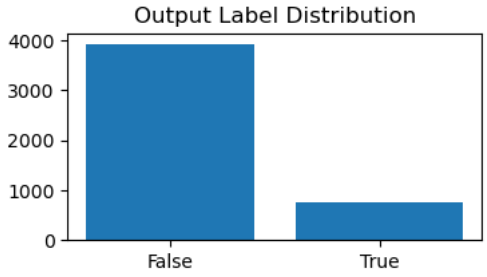

Figure 4: Output label distribution

2.4 Data Processing

Because the input features are not distributed uniformly, the data has been normalized. There are several methods that can be used to accomplish this; mean normalization was used, following this formula.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

The normalized data has the same distribution as the original data, as can be seen below (see Figure 5).

Normalized Data Visualizations

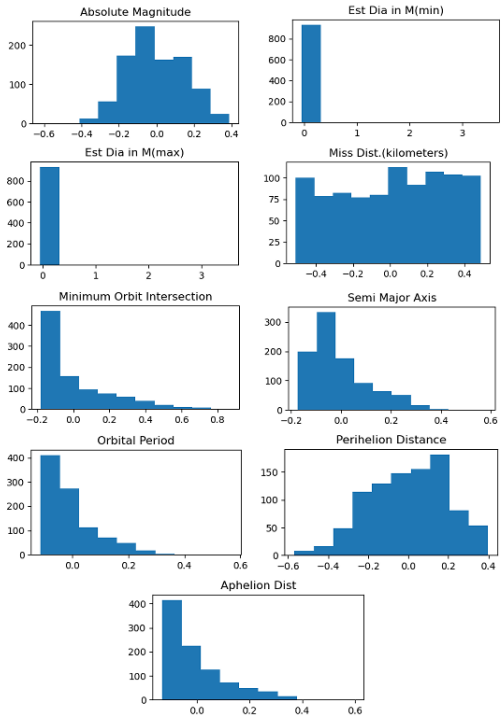

Figure 5: Normalized input distribution

2.5 Data Splitting

The data will be randomly shuffled and split such that 80% of the data will be used as the training set and 20% of the data will be used for the validation set. Depending on the performance of the model, this may be altered in later steps.

3 Model Selection and Evaluation

3.1 Modeling

The architecture used for the model is a feed forward neural network. Because the data is shuffled before models are trained and evaluated, results vary each time. Models for this step were compiled on May 5, 2024.

3.1.1 Baseline Model

The model used as a control starts with one input layer and an output layer and no hidden layers. Three iterations in total were tested, including a model

with one hidden layer and a model with two hidden layers. The performance of each of these models can be seen in the table below.

Of the three baseline models, the one with one hidden layer performed best overall. Below are the learning curves measuring accuracy and loss for this model.

Baseline Model Learning Curves

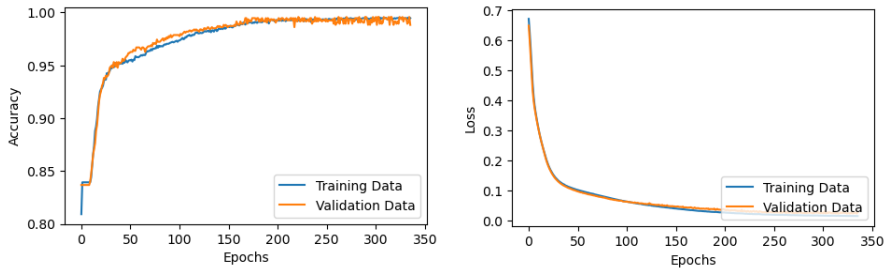

Figure 6: Baseline 10-8-1 NN Learning Curves

3.1.2 Sigmoid Activation

The baseline model uses relu activation on the input layer and all hidden layers, with a sigmoid activation function on the output layer. Also tested was a model that uses sigmoid activation on all layers; its learning curves are depicted below.

Sigmoid Activation Learning Curves

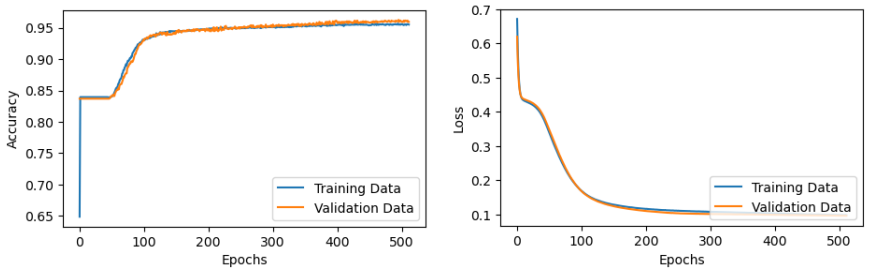

Figure 7: 10-8-1 Sigmoid Activation (all layers) Learning Curves

3.1.3 Other Architectures

Other neural network architectures were experimented with; their learning curves can be seen below.

2-1 Neural Network Learning Curves

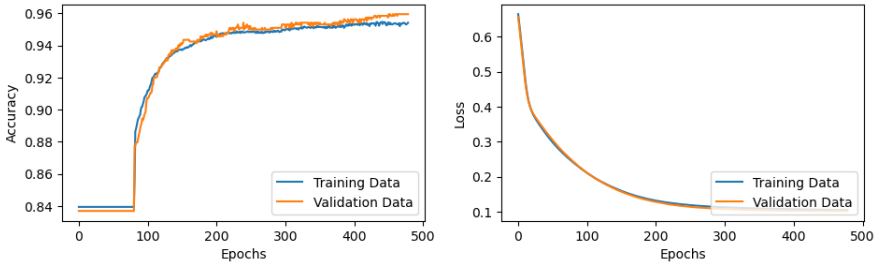

Figure 8: 2-1 NN Learning Curves

18-8-4-1 Neural Network Learning Curves

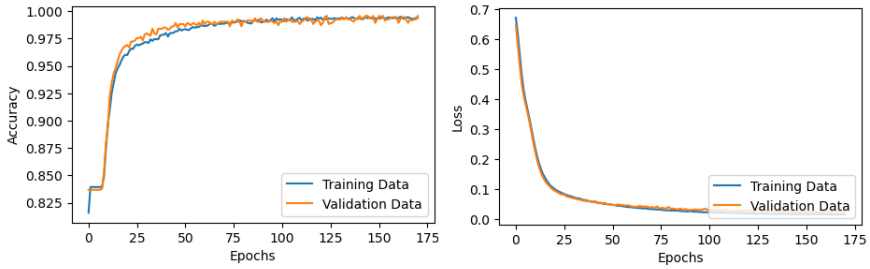

Figure 9: 18-8-4-1 NN Learning Curves

3.1.4 Overfitting Model and Learning Curves

A model with architecture 256-256-256-256-256-1 was tested in order to overfit the model; its learning curves can be seen below.

Overfitting Model Learning Curves

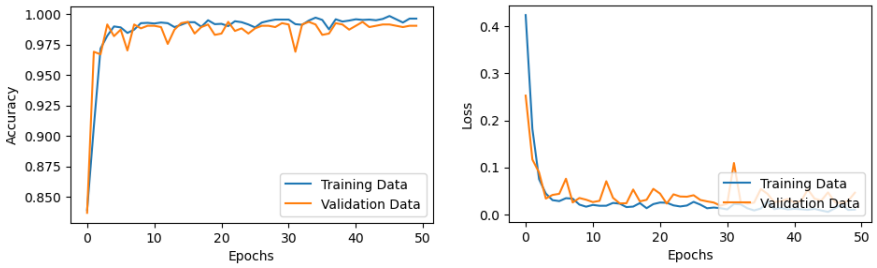

Figure 10: Overfitting NN Learning Curves

3.2 Model Evaluation

The performance of all the models tested was evaluated on accuracy, precision, recall, and F1 scores. These metrics are compared in the figure below.

Model	Accuracy	Precision	Recall	F1 Score
Baseline Model (10-1)	99.15	100.00	94.77	0.97
Baseline Model (10-8-1)	99.57	99.34	98.04	0.99
Baseline Model (10-8-4-1)	99.25	98.03	97.39	0.98
Sigmoid (all neurons)	95.95	88.08	86.93	0.88
Neural Network (2-1)	95.84	88.00	86.27	0.87
Neural Network (18-8-4-1)	99.57	98.69	98.69	0.99
Overfit Model (64-16-8-4-1)	99.25	98.03	97.39	0.98

Figure 11: NN Performance Comparison

Overall, the second baseline model with structure 10-8-1 performed the best, with consistently high scores on every metric.

4 Feature Importance and Reduction

4.1 Feature Importance

After testing different architectures and comparing performance, individual input features were analyzed to determine how much impact they had on the performance of the model.

4.2 Performance of Individual Features

The importance of each input feature was determined by running a model that took each individual feature as its only input using the architecture determined to be highest performing in the previous step. The accuracy was intended to indicate which feature impacted the model the most. However, each input feature performed similarly through several iterations, so in the feature reduction step, input features were removed progressively according to loss from highest to lowest (see Figure 12).

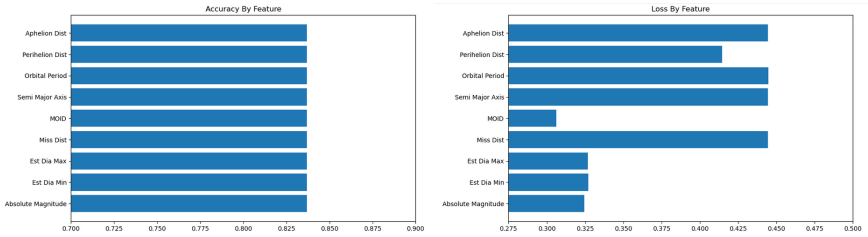

Figure 12: Input feature accuracy and loss comparison

4.3 Iterative Feature Reduction

After determining the order in which to remove features from the model, features were removed one at a time from the model and the performance of the model was checked at each step, using accuracy as the measurement metric.

The steepest drop in accuracy occurred when the third feature, Perihelion Distance, was removed from the model. However, this result was inconsistent across several trials. The model still performs relatively well when the two features with the lowest individual loss statistics, Absolute Magnitude and Earth Minimum Orbit Intersection Distance, are the only input features in the model (see Figure 13). This is somewhat expected, as these two factors are the main determining factor in whether or not a Near-Earth Object is classified as a Physically Hazardous Asteroid,[4] though it was expected that the other features would still have a notable impact on the performance of the model.

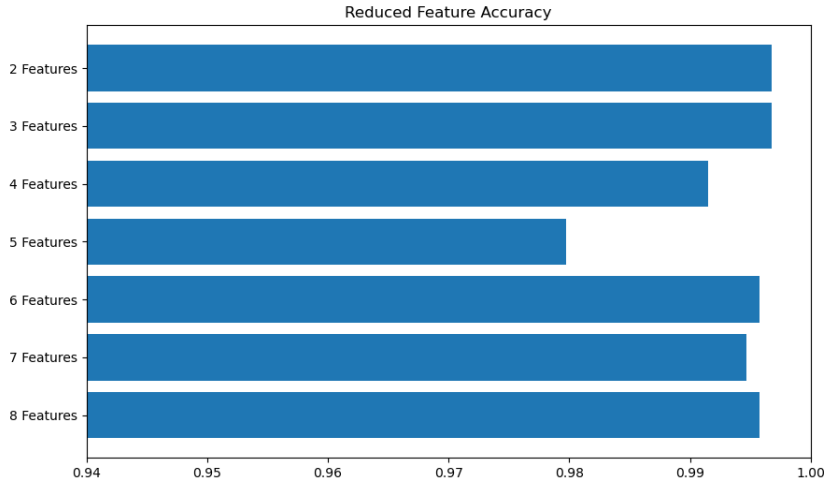

Figure 13: Performance of reduced feature models

5 Conclusions

5.1 Best Model Performance

The best model tested was the baseline neural network with architecture 10-8-1. In order to further analyze the performance of the model, a Receiver Operating Curve (ROC) was visualized and the area under the curve (AUC) was calculated.[5] The ideal value for the AUC is 1;[6] as can be seen below, the model achieved an AUC very close to 1, at 0.9993.

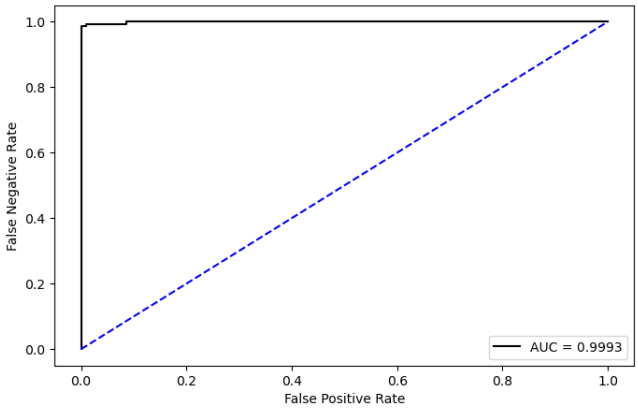

Figure 14: ROC/AUC of NN 10-8-1 model

5.2 Future Considerations

One of the more unexpected occurrences during this process was the performance of the reduced model with only two features, Absolute Magnitude and Earth Minimum Orbit Intersection Distance. The original dataset consisted of 39 input features with one output label, and the database that the original dataset was pulled from contains even more potential input features. It could be beneficial to experiment with neural networks testing more of these individual input features to determine how much bearing they have on the final classification.

References

[1] NASA Center for Near Earth Object Studies. NEO Basics, 2019. Last accessed 1 May 2024.

[2] NASA Jet Propulsion Laboratory. Keeping an Eye on Space Rocks. Last accessed 1 May 2024.

[3] Lovish Bansal. NASA: Asteroids Classification. Last accessed 27 April 2024.

[4] NASA Center for Near Earth Object Studies. Glossary-PHA. Last accessed 5 May 2024.

[5] Normalized Nerd. Machine Learning with Scikit-Learn Python — ROC & AUC. Last accessed 5 May 2024.

[6] Stat Quest with Josh Starmer. ROC and AUC, Clearly Explained! Last accessed 5 May 2024.