Predictive Modeling of Hazardous Near-Earth Objects (NEOs)

Siddarth Puliyanda

ISYE 7406

Fall 2023

Abstract

The study of Near-Earth Objects (NEOs) represents a vital intersection between planetary science and data science. NEOs are celestial bodies, most commonly asteroids and comets, whose orbits around the Sun occasionally bring them close to Earth's neighborhood. While a vast majority of them pass by harmlessly, their close approaches and chance for impact can pose a potentially serious risk to our planet. Understanding the characteristics, trajectories, frequencies, and other qualities of NEOs is crucial for the development of risk mitigation strategies against future impacts. This research study utilizes the *NASA NEO Dataset* to create data visualizations, perform statistical analysis, and build predictive models that help us identify whether a NEO is potentially hazardous or not. Using Logistic Regression and Random Forest models, we achieved a classification accuracy of 92%, with the Random Forest model showing the best performance.

Introduction

Near-Earth objects (NEOs) are asteroids and comets with orbits within close proximity to the Sun, which can bring them near the Earth's gravitational neighborhood **[1]**. The ones that pose a risk of impact on Earth are known as potentially hazardous NEOs, and are carefully tracked by scientists in order to study their orbital activity **[2]**. It is worth noting that NEOs can be known for tens or hundreds of years before their orbits become dangerous with Earth, but astronomers have estimated it takes 5 to 10 years to prepare for a potential impact, so the earlier it is detected the better **[2]**. The goal of our project is to build a classification model that identifies whether a NEO is potentially hazardous or not. More importantly, the project will investigate the factors impacting the potentially hazardous NEOs.

This real-world problem can be formatted into a research question of classification. Data mining and statistical methods (such as Logistic Regression and Random Forest) will be applied to solve the problem. The following analytical questions may be answered during the EDA and model interpretation parts.

- Hazard analysis: to investigate the properties of NEOs marked as "hazardous" compared to those that are not. Are there certain size or speed thresholds that make a NEO more likely to be categorized as hazardous?
- Velocity analysis: to investigate the distribution of NEO velocities. Are faster NEOs more or less common? How does velocity correlate with other properties, such as size or hazard potential?
- Temporal analysis: to examine the frequency of close approaches over time. Are there certain periods where Earth experiences a higher number of close NEO approaches?

The outline of the project includes sections such as data information/EDA, approach/methodology, results/findings, and conclusion.

Data Information/EDA

We will be using the *NASA NEO (Near-Earth Object) Dataset* hosted on Kaggle for the project **[3]**. The data set contains 904 data points and no missing values, with each data point providing specific information about a unique NEO during its close approach to Earth. The data glance of the first 3 data points is shown in **Appendix Figure 1**.

There are 27 variables or features in the original data set, and a data dictionary is provided in **Appendix Table 1**. The dependent variable is *Is Potentially Hazardous*, a boolean variable with values of True/False, while the remaining 26 independent variables are a mix of quantitative (20) and qualitative (6) data.

Since a lot of repetitive independent variables are included in the data set, as shown in **Appendix Figure 1** and **Appendix Table 1**, variable selection was initially performed in order to eliminate duplicative columns (**Table 1**). For example, the variable of Limited Name was used to represent the NEOs, while the other referential alternatives (including ID, Neo Reference ID, Name, Designation, and NASA JPL URL) were removed. Variables of Min Diameter (km), Max Diameter (km), Relative Velocity (km/s), and Miss Distance (astronomical) were selected, while the alternative features related to the unit conversions (such as m, miles, feet and so on) were ignored. For time series data, Epoch Date Close Approach was used as a continuous variable, while Close Approach Date was used as a categorical variable. In summary, 6 numerical independent variables and 3 categorical variables were chosen (**Table 1**).

Table 1. Data dictionary after variable selection

Variable Name	Data Type	Description
Limited Name	str	Limited version of the NEO's name
Absolute Magnitude (H)	float	Absolute magnitude of the NEO
Min Diameter (km)	float	Min diameter of the NEO in kilometers
Max Diameter (km)	float	Max diameter of the NEO in kilometers
Close Approach Date	str	Date of the close approach
Epoch Date Close Approach	int	Epoch timestamp of the close approach
Relative Velocity (km/s)	float	Relative velocity of the NEO during close approach in kilometers per second
Miss Distance (astronomical)	float	Miss distance of the NEO from Earth during close approach in astronomical units
Orbiting Body	str	Celestial body that the NEO is orbiting

Univariate analysis was performed by histograms of numerical variables (**Appendix Figure 2)**. Numerical independent variables do not follow a normal distribution. Min Diameter, Max Diameter and Miss Distance have potential influential data points/outliers at the higher range. We do not want to delete these abnormal numbers because they may represent the real data and help us understand the big picture. Herein, Logistic Regression or Tree-based models will be created since they are relatively robust to outliers. In a real-world scenario, we would attempt to discover if these abnormal numbers are real or errors.

Bivariate analysis between independent variables was investigated using a correlation matrix (**Figure 1**). Independent variables of Absolute Magnitude, Min Diameter and Max Diameter are highly correlated with a value of abs(corr) more than 0.8. Including highly correlated variables can inflate the model coefficients and provide misleading insights of feature importance. Because of this, highly correlated variables will be removed before modeling or Tree-based modeling will be used.

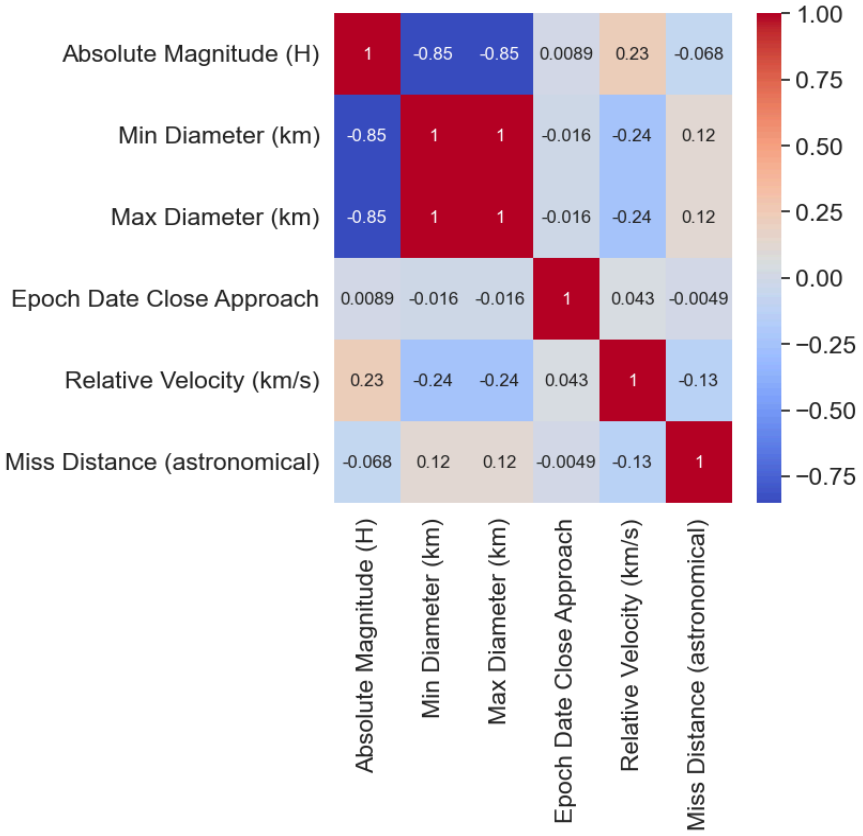

Figure 1. Bivariate analysis by a correlation matrix

Boxplots for size, velocity, and distance effects are displayed below on the hazardous and non-hazardous NEOs. (**Figure 2**). The data set includes imbalanced 70% non-hazardous and

30% hazardous NEOs. Resampling methods may be needed to deal with the imbalanced data during the modeling part.

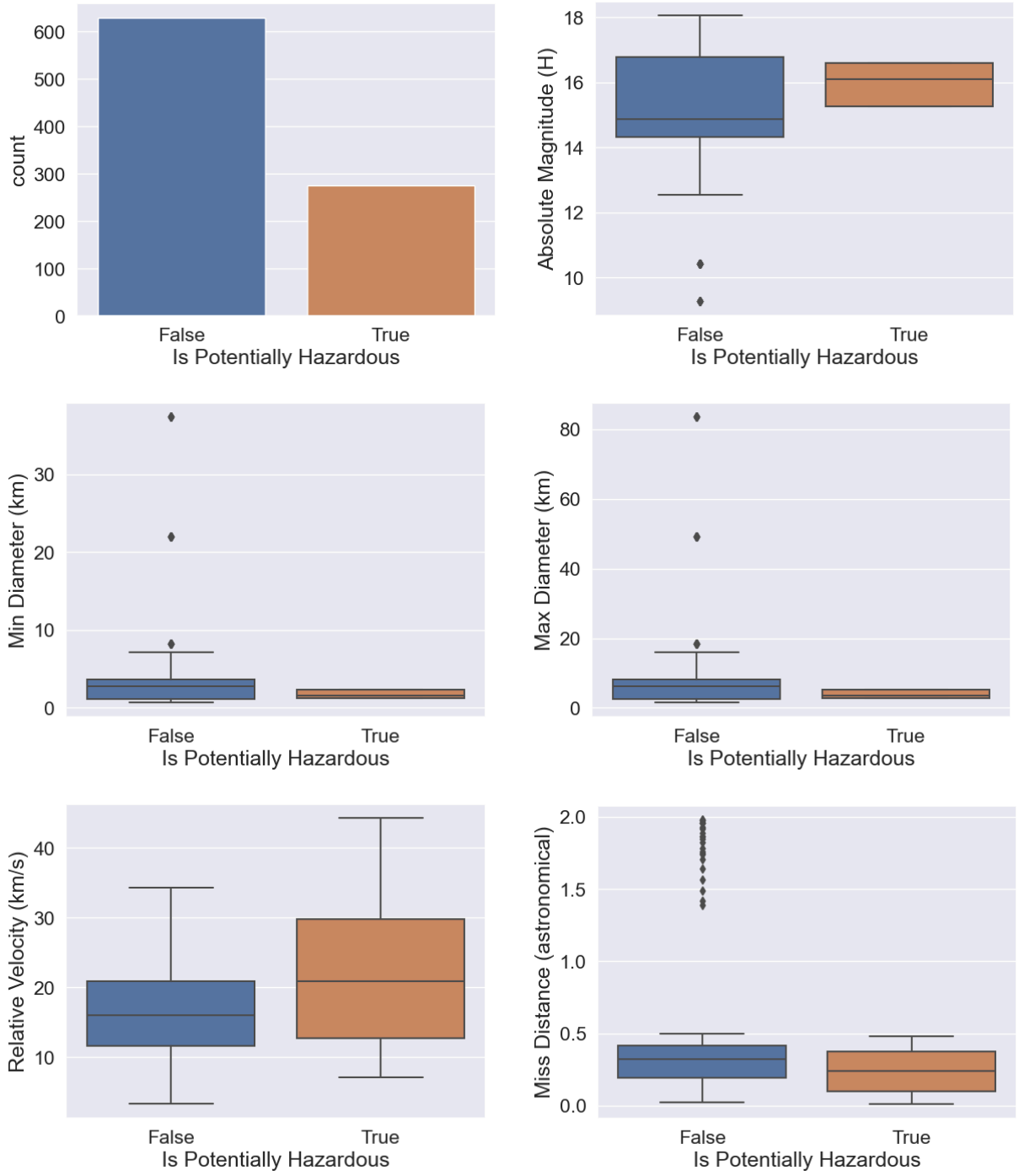

Figure 2. Boxplots of the dependent variable or between independent and dependent variables

The boxplots show that the spread, median, and interquartile range for the Min/Max Diameter of hazardous NEOs is narrower, indicating a smaller range of sizes among them. This

seems a little bit different from our intuition because we would think that larger objects would be more hazardous while smaller objects would be less. It was also found that astronomy "...use the more practical measure of absolute magnitude (H)..."**[4]** to study the size effect. As shown in **Figure 2**, hazardous NEOs have a larger median Absolute Magnitude or size, compared to non-hazardous NEOs. Based on the practical measure of size and previously stated collinearity, Absolute Magnitude instead of Min/Max Diameter will be used in the following modeling part.

As shown in **Figure 2**, the Relative Velocity of both hazardous and non-hazardous NEOs exhibits broad ranges, with overlaps in their distributions. The median velocity for hazardous NEOs is higher than that for non-hazardous NEOs, which is consistent with our intuition that faster objects are more hazardous than slower ones. The Miss Distance of non-hazardous NEOs appears to be slightly larger than that of hazardous ones, while also containing a large amount of outliers with very large miss distances. This makes sense because the further away the NEO is to Earth the less chance they have of impact.

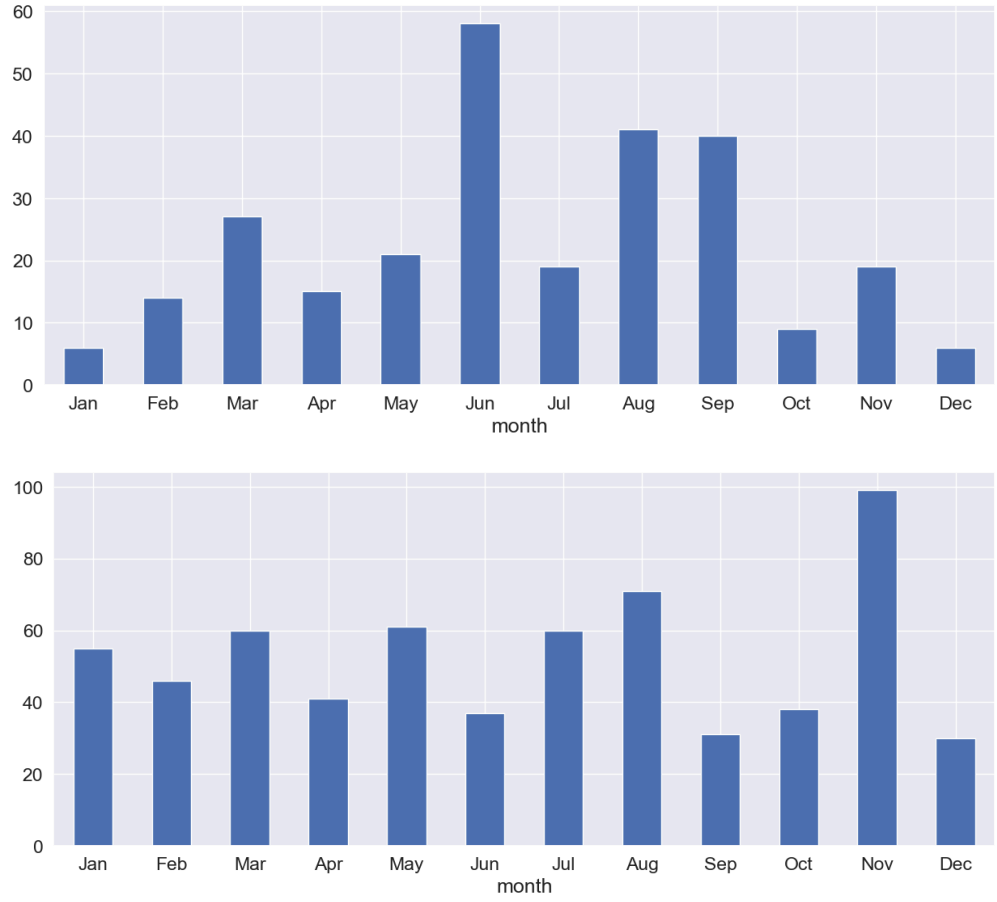

Figure 3. Monthly close approaches for hazardous (top) and non-hazardous (down) NEOs

The frequency of close approaches of NEOs over time was examined with the variables Close Approach Date and Epoch Data Close Approach. The peak for hazardous NEOs happens in June and appears lower in the Winter months, while the peak for non-hazardous NEOs happens in November with a fairly consistent range throughout the year (**Figure 3**). Since the boxplot for Epoch Date Close Approach between hazardous and nonhazardous NEOs appears identical and will not add any value, Epoch Date Close Approach will be excluded from the modeling (**Appendix Figure 3)**.

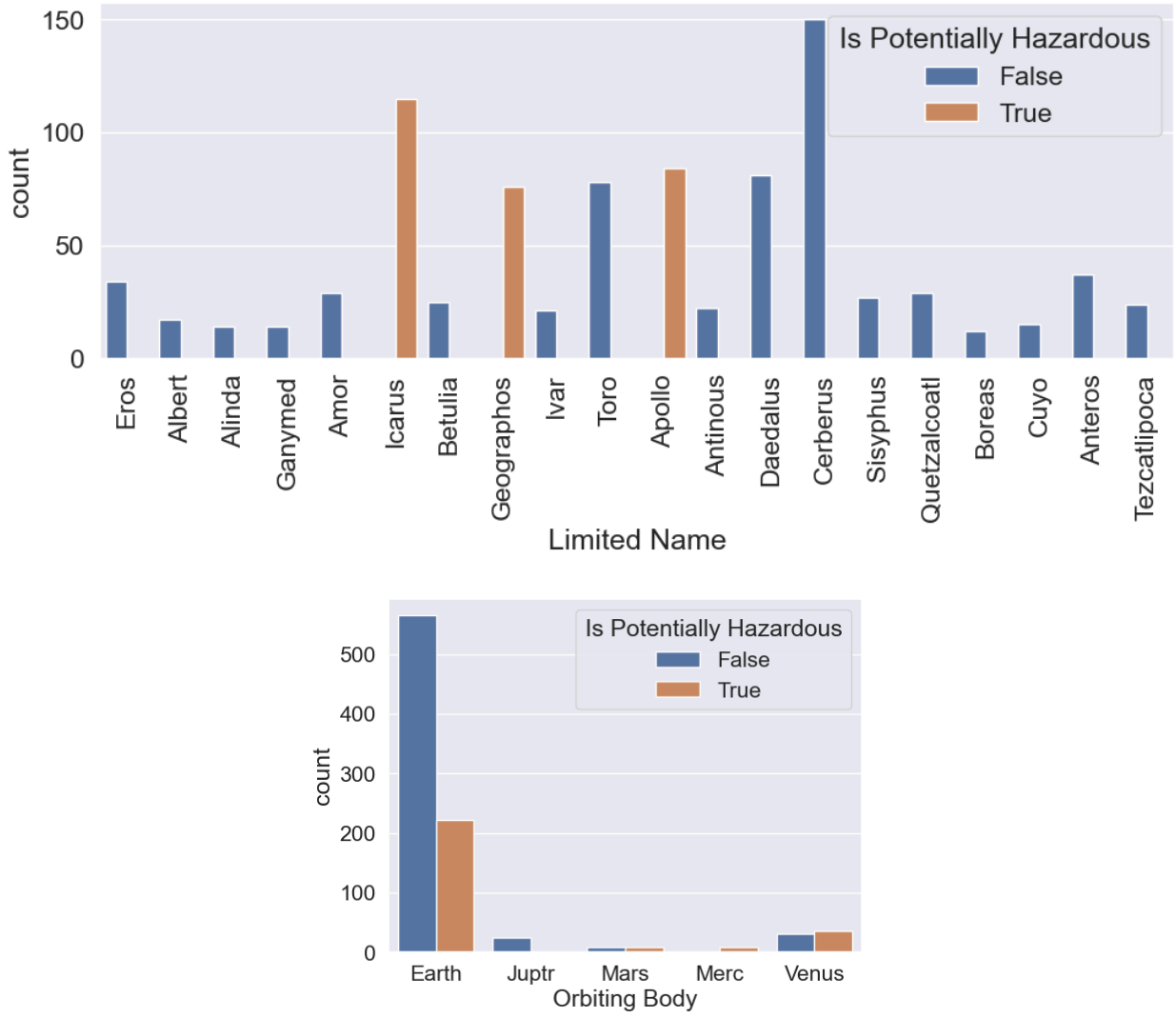

Figure 4. Data visualization between categorical independent variables and dependent variable

A histogram for the variable Limited Name was created to see which categories the hazardous and non-hazardous NEOs fell into. **Figure 4** shows that hazardous NEOs belong to groups of Icarus (eccentric orbits), Apollo (Earth-crossing orbits) and Geographos (elongated

orbits with rapid rotation). Limited Name is used for reference by the NASA database, and will be excluded from the following modeling. Finally, using the column Orbiting Body it is shown that NEOs orbiting the Earth have more chances to be hazardous.

Approach/Methodology

The project objective is to build a predictive model for identifying whether a NEO is potentially hazardous or not. The approach includes pre-modeling, modeling and post-modeling parts (**Figure 5**), using Python as the programming language and several of its libraries such as Pandas for data manipulation, Matplotlib and Seaborn for visualization, and Scikit-Learn for predictive modeling.

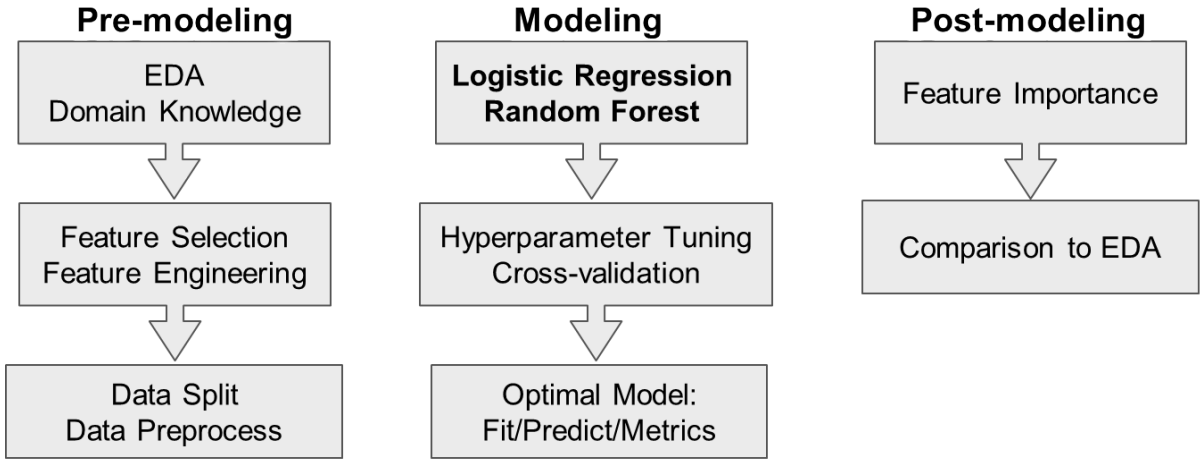

Figure 5. Approach and methodology

As previously mentioned, EDA was performed to study the factors determining a hazardous NEO or not. Feature selections are based on the EDA results and domain knowledge of astronomy. The data was first prepared by combining non-significant NEO names into a single 'others' category and simplifying the 'Orbiting Body' feature by consolidating less frequent categories into 'others' as well. The target variable 'Is Potentially Hazardous' was also transformed into a binary format for easier use with the Python libraries.

For modeling, the features selected are Absolute Magnitude (H), Relative Velocity (km/s), Miss Distance (astronomical) and Orbiting Body and the target is Is Potentially Hazardous. The original data set was split into train/test (80/20) data sets using a sampling technique to maintain the proportion of the classes in both datasets. The data preprocessing technique of standardization was applied to the numerical features of Absolute Magnitude (H), Relative Velocity (km/s), and Miss Distance (astronomical). One-hot-encoding was applied to the

categorical feature of Orbiting Body to generate dummy variables of Orbiting Body_Earth, Orbiting Body_Venus and Orbiting Body_others.

Based on these inputs, two models were built: Logistic Regression and Random Forest Classifier. Logistic regression was first tried due to its simplicity and easy interpretability with classification problems. Tree-based models with ensemble methods such as random forest were considered since they were robust to outliers, abnormal distributions, and collinearity. For the Random Forest Classifier, GridSearchCV was used to conduct hyperparameter tuning and cross-validation to choose the optimal model and prevent over-fitting. Model selection focuses on the ROC-AUC value due to the lacking information of costs of Type I and Type II errors. The selected optimal model was applied to predict the test data set and evaluated by accuracy, precision, recall and F1-score. After modeling, feature importance was plotted and the results were compared to the previous EDA results.

Results/Findings

GridSearchCV combines the hyperparameter tuning and 5-fold cross-validation to choose the optimal model and prevent over-fitting. For Logistic Regression models, penalty of "elasticnet" was added to combine the L1 and L2 regularization. Hyperparameters of C and L1_ratio were tuned. Smaller C value specifies a stronger regularization and L1_ratio specifies the percentage of L1 regularization in the Elastic-Net mixing parameter. Model selection focuses on the ROC-AUC value due to the lacking information of costs of Type I and Type II errors. Among 18 models, the optimal Logistic regression achieves a ROC-AUC value of 0.77 with a C value of 1 and L1_ratio of 0.33 (**Table 2**).

Table 2. Hyperparameter tuning for Logistic Regression models

Hyperparameter		**Test Metrics**	
C	**L1_ratio**	**Mean ROC-AUC**	**Rank**
1	**0.33**	**0.765481**	**1**
1	0.1	0.765391	2
1	0.033	0.765390	3
1	0.01	0.765300	4
1	0	0.765255	5
1	1	0.765254	6
10	0.1	0.764938	7
10	0.033	0.764938	7

For Random Forest Models, hyperparameters of max_depth, min_samples_leaf and max_features were considered. Max_depth and max_features specify the maximum depth of the tree and maximum features at each split. Min_samples_leaf is the minimum number of samples required to be at a leaf node. Model selection focuses on the ROC-AUC value and computational complexity. Among 27 models, the optimal Random Forest model achieves a ROC-AUC value of 1.00 with a max_depth of 5, min_samples_leaf of 5 and max_features of 0.8 (**Table 3**).

Table 3. Hyperparameter tuning for Random Forest models

Hyperparameter			**Test Metrics**	
Max_depth	**Min_samples_leaf**	**Max_features**	**Mean ROC-AUC**	**Rank**
5	**5**	**0.8**	**1.000000**	**1**
5	10	0.8	1.000000	1
8	5	0.8	1.000000	1
8	10	0.8	1.000000	1
8	15	0.8	1.000000	1
5	15	0.8	0.999955	6

The metrics of the optimal Logistic Regression and Random Forest models are listed in **Table 4**. The optimal Random Forest model has a higher value of accuracy, precision, recall and F1-score at greater than 0.90 for all values, compared to the optimal Logistic Regression with values around 0.60. Thus, the optimal Random Forest model performs the best.

Table 4. Metrics for optimal models

Optimal Model	**Train**	**Test**			
	Accuracy	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.68	0.60	0.66	0.60	0.62
Random Forest	1.00	0.92	0.93	0.92	0.91

Both the optimal Logistic Regression and Random Forest models show that the top 3 important features are Absolute Magnitude (H), Relative Velocity (km/s), and Miss Distance (astronomical) (**Figure 6**). However, the order of the top 3 features is different in the models. Although the performance of the optimal Logistic Regression model is worse than the Random Forest model, the sign and value of coefficient of Logistic Regression indicates that large and fast NEOs closer to the Earth during the close approach are more likely to be the hazardous NEOs. The finding is consistent with the previous EDA results. Thus, we highly recommend

space scientists to be cautious of the large and fast NEOs closer to the Earth during the close approach.

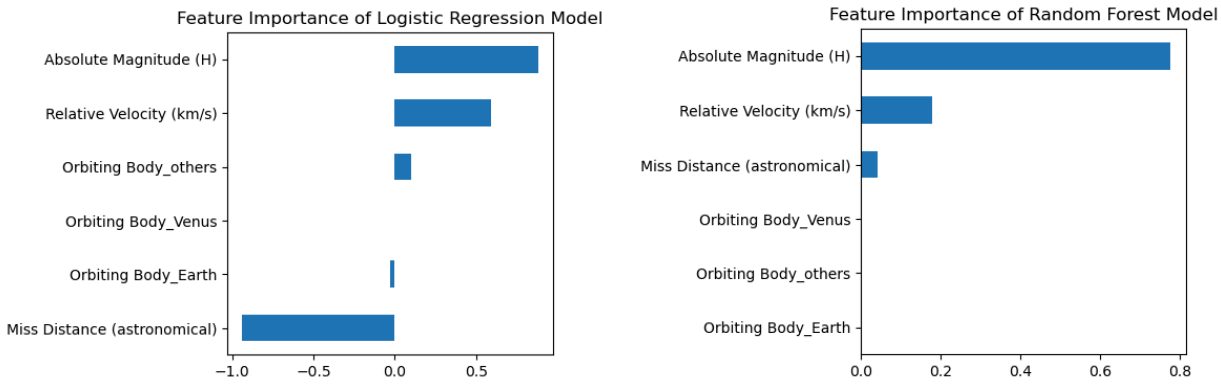

Figure 6. Feature importance plots

Conclusion

Scientists have documented historical NEOs and their hazard levels to Earth in the *NASA NEO (Near-Earth Object) Dataset*. By building a Random Forest model with hyperparameter tuning and cross-validation, the resulting predictive model achieves an accuracy of 92% for identifying whetherNEOs are hazardous, which is better than alternative models (e.g. Logistic Regression). When new asteroids are discovered, this model can be used to more accurately identify its hazard level, helping scientists to focus on tracking the right NEOs more efficiently. Future work of this analysis may include data mining for larger training sets and modeling by more advanced methods (such as boosting or neural networks) for performance improvements if required in the real-world applications.

- **Lessons learned from this project or course**

The biggest lesson I have learned is that EDA prior to modeling is crucial, especially with classification problems. Thorough and comprehensive EDA provides meaningful insights of the factors that contribute to determining hazardous or non-hazardous NEOs. EDA also helps the feature selection, allowing for the final model to be simpler and less prone to overfitting. However, if the situation was more complex, for example, if the data set contained thousands of independent variables, we could not do EDA on each variable one by one. We could instead take advantage of the data mining and statistical methods learned from the course (such as forward selection, backward elimination, or PCA) for variable/feature selection.

References

1. Asteroid Watch (nasa.gov)

2. Potentially hazardous object - Wikipedia
3. https://www.kaggle.com/datasets/shivd24coder/nasa-neo-near-earth-object-dataset/data
4. https://en.wikipedia.org/wiki/Potentially_hazardous_object#Size

Appendix

	0	1	2
ID	2000433	2000433	2000433
Neo Reference ID	2000433	2000433	2000433
Name	433 Eros (A898 PA)	433 Eros (A898 PA)	433 Eros (A898 PA)
Limited Name	Eros	Eros	Eros
Designation	433	433	433
NASA JPL URL	http://ssd.jpl.nasa.gov/sbdb.cgi?sstr=2000433	http://ssd.jpl.nasa.gov/sbdb.cgi?sstr=2000433	http://ssd.jpl.nasa.gov/sbdb.cgi?sstr=2000433
Absolute Magnitude (H)	10.41	10.41	10.41
Min Diameter (km)	22.006703	22.006703	22.006703
Max Diameter (km)	49.208483	49.208483	49.208483
Min Diameter (m)	22006.702711	22006.702711	22006.702711
Max Diameter (m)	49208.483223	49208.483223	49208.483223
Min Diameter (miles)	13.674327	13.674327	13.674327
Max Diameter (miles)	30.576724	30.576724	30.576724
Min Diameter (feet)	72200.470524	72200.470524	72200.470524
Max Diameter (feet)	161445.160099	161445.160099	161445.160099
Is Potentially Hazardous	False	False	False
Close Approach Date	1900-12-27	1907-11-05	1917-04-20
Close Approach Date (Full)	1900-Dec-27 01:30	1907-Nov-05 03:31	1917-Apr-20 21:19
Epoch Date Close Approach	-2177879400000	-1961526540000	-1663036860000
Relative Velocity (km/s)	5.578619	4.394491	4.816784
Relative Velocity (km/h)	20083.029075	15820.167199	17340.422466
Relative Velocity (miles/h)	12478.81326	9830.036668	10774.664171
Miss Distance (astronomical)	0.314929	0.471486	0.499257
Miss Distance (lunar)	122.507447	183.407876	194.211053
Miss Distance (km)	47112732.928149	70533232.893794	74687814.599751
Miss Distance (miles)	29274494.765192	43827318.620435	46408855.985038
Orbiting Body	Earth	Earth	Earth

Figure 1. Data glance of the first 3 data points

Table 1. Data dictionary

Name	Data Type	Description
ID	int	Unique identifier for the NEO
Neo Reference ID	int	Another reference identifier for the NEO
Name	str	Name of the NEO
Limited Name	str	Limited version of the NEO's name
Designation	int	Designation of the NEO
NASA JPL URL	str	URL to the NEO's information on NASA's Jet Propulsion Laboratory (JPL) website
Absolute Magnitude (H)	float	Absolute magnitude of the NEO
Min Diameter (km)	float	Min diameter of the NEO in kilometers
Max Diameter (km)	float	Max diameter of the NEO in kilometers
Min Diameter (m)	float	Min diameter of the NEO in meters
Max Diameter (m)	float	Max diameter of the NEO in meters
Min Diameter (miles)	float	Min diameter of the NEO in miles
Max Diameter (miles)	float	Max diameter of the NEO in miles
Min Diameter (feet)	float	Min diameter of the NEO in feet
Max Diameter (feet)	float	Max diameter of the NEO in feet
Is Potentially Hazardous	boolean	Indicate whether the NEO is potentially hazardous
Close Approach Date	str	Date of the close approach
Close Approach Date	str	Full date and time of the close approach

(Full)		
Epoch Date Close Approach	int	Epoch timestamp of the close approach
Relative Velocity (km/s)	float	Relative velocity of the NEO during close approach in kilometers per second
Relative Velocity (km/h)	float	Relative velocity of the NEO during close approach in kilometers per hour
Relative Velocity (miles/h)	float	Relative velocity of the NEO during close approach in miles per hour
Miss Distance (astronomical)	float	Miss distance of the NEO from Earth during close approach in astronomical units
Miss Distance (lunar)	float	Miss distance of the NEO from Earth during close approach in lunar distances
Miss Distance (km)	float	Miss distance of the NEO from Earth during close approach in kilometers
Miss Distance (miles)	float	Miss distance of the NEO from Earth during close approach in miles
Orbiting Body	str	Celestial body that the NEO is orbiting

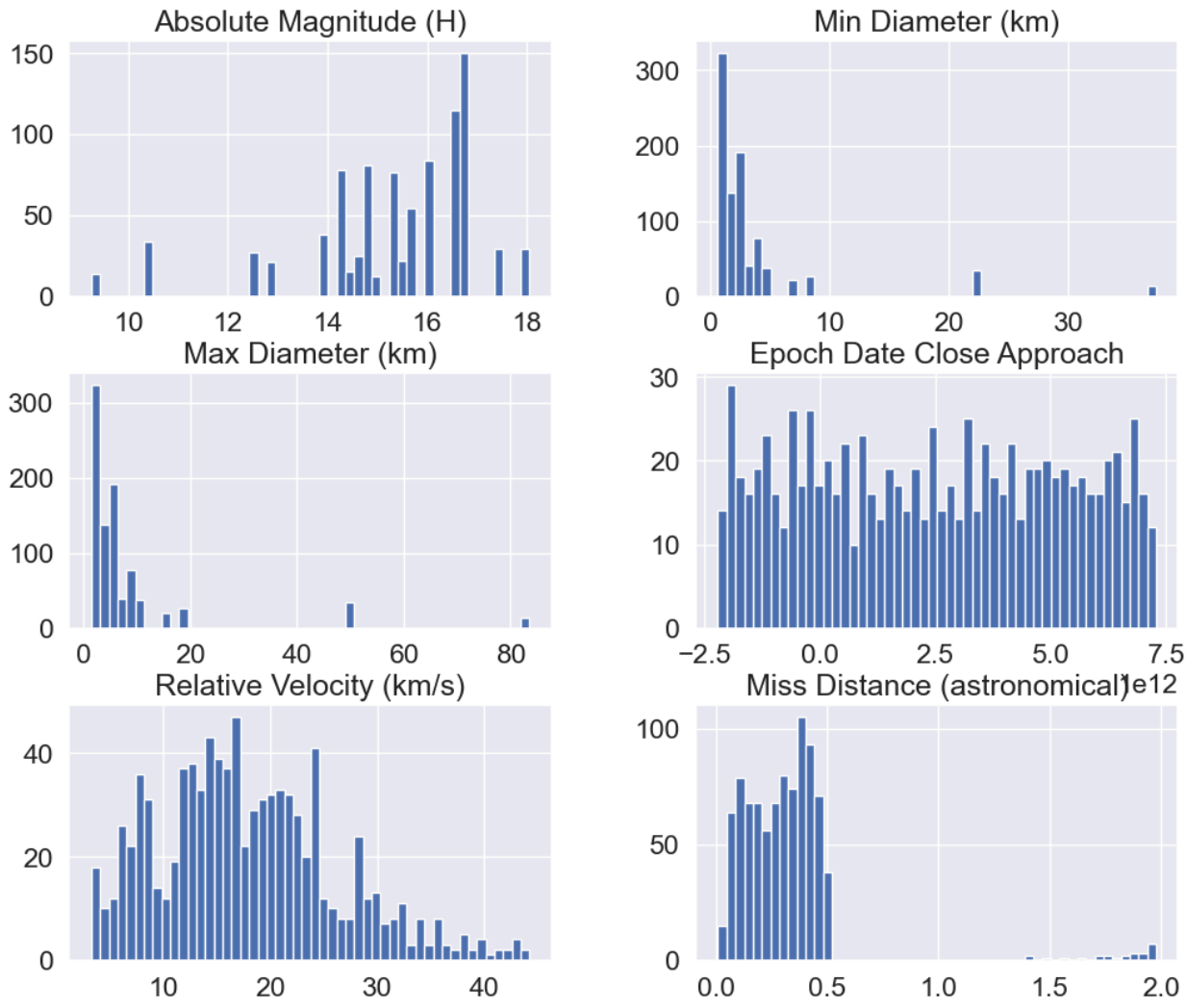

Figure 2. Univariate analysis by histograms

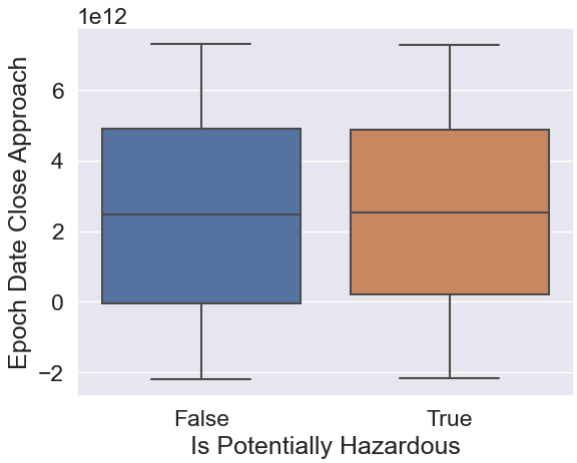

Figure 3. Boxplots between independent variable of Epoch Date Close Approach and dependent variable of Is Potentially Hazardous