Data Science Programming: Group Project Report

This report and presentation are the work of Group 5: Eshaan Arora, Jose Currea, Gayathree Gopi, Aileen Li, and Felipe Zapater

---

Introduction and Background

This group elected to focus on a NASA data set of near-earth objects (NEOs), including objects which might be hazardous to Earth. This was accomplished through the following steps:

1. Exploratory Data Analysis (EDA)
2. Modeling and Cross-Validation using the following models:
    a. K Nearest Neighbors (KNN)
    b. Naïve Bayes
    c. Classification Trees
    d. Random Forest

Project Goals

The primary goal of this analysis was to answer the question–what is the likelihood that a NEO is hazardous (i.e., the likelihood that the NEO will impact planet Earth.)

This is a critical problem to answer, as early detection of potentially hazardous NEOs is critical to warning governments, space experts, and the general public, and would dramatically increase the likelihood of a successful response to these threats. Indeed, the general public has just as much an interest in addressing this issue as the scientific and academic communities, as the hazards posed by NEOs are a threat to all humankind, exemplified by previous NEOs such as the Chelyabinsk meteor in 2013.

Exploratory Data Analysis

The dataset comprised 338,199 observations and 9 variables: *neo_id*, *name*, *absolute_magnitude*, *estimated_diameter_min*, *estimated_diameter_max*, *orbiting_body*, *relative_velocity*, *miss_distance*, and *is_hazardous* (response variable). After inspection, we excluded *orbiting_body* due to its single unique value and noted the dataset's imbalance (87% false, 13% true cases, see **Figure 1**). We compensated for this imbalance using cross-validation. For feature engineering, we normalized the data, one-hot encoded the response variable, and introduced *avg_estim_diameter* and *range_estim_diameter* to simplify asteroid size and

composition metrics. In order to understand the distribution and visualize the relationships for each feature, we generated kernel density plots, pairwise plots, and a Spearman correlation heatmap from a downsized sample (10%) of the data while maintaining the proportion of true and false cases.

Based on the kernel density plots, the *absolute_magnitude* has little skewness while *relative_velocity*, *avg_estim_diameter*, and *range_estim_diameter* have increasing magnitudes of skewness towards the right (**Figure 2**). This means that while most asteroids have smaller diameters with lower velocities, there are some very large asteroids with higher velocities that are pulling the average diameter and velocity upward.

The scatter plots indicated some separation between hazardous and non-hazardous asteroids based on *absolute_magnitude* and *avg_estim_diameter* (**Figure 3**). The plots overall hint at complex relationships between the predictors and the response variable, which might require a more sophisticated modeling to fully understand.

Lastly, the Spearman correlation heatmap showed there is some correlation with the remaining predictors (**Figure 4**). Notably, there is a moderate negative correlation between *absolute_magnitude* and *is_hazardous* (-0.37) suggesting that brighter asteroids are more often less hazardous. Similarly, both *avg_estim_diameter* and *range_estim_diameter* show a moderate positive correlation (0.37) with hazard status, indicating larger asteroids tend to be more hazardous. Interestingly, *miss_distance* shows almost no correlation (-0.01) with hazard status, implying that proximity alone does not determine hazard potential. The heatmap highlights which features are most strongly associated with asteroid hazards. After data exploration and feature engineering, we fit KNN, Naive Bayes, Random Forest, and Classification Trees onto the modified dataset.

Solutions and Insights

I. k-Nearest Neighbors

For our first model, we tested a k-Nearest Neighbors classifier with 17 neighbors, uniform weights, and Euclidean distance and performed 5-fold cross-validation. The model performed well, with accuracies of 0.897 and 0.882 before cross-validation and accuracies of 0.896 and 0.883 after cross-validation. This implies a low possibility of overfitting due to the small differences in train and test accuracies. Our cross-validation showed that the optimal number of nearest neighbors was 15-17 (**Figure 5**). One

consideration, however, is that the low train accuracy in both cases indicates this model probably is not powerful or complex enough to capture the entirety of the signal. In the future, we could create or use new model parameters to improve accuracy.

II. Naïve Bayes

The Naïve Bayes model achieved an accuracy of 73.45% on the training set and 73.60% on the testing set, falling short of the prior probability of non-hazardousness at around 87%. The confusion matrix shows 41,733 true negatives and 8,044 true positives, but also highlights 17,279 false positives and 573 false negatives. While the model demonstrated strong recall for hazardous NEOs at 93.3%, it struggled with precision, yielding a low 30% due to many non-hazardous NEOs being misclassified as hazardous.

The model provided important insights, such as identifying that NEOs with higher magnitude, greater diameter, and lower miss distance are more likely to be hazardous. However, its reliance on the Naïve Bayes assumption of feature independence likely limited its performance, particularly by not accounting for the interaction between features like magnitude and diameter. This suggests that a model capable of capturing feature interactions would have been more effective in predicting hazardous NEOs.

III. Decision Trees

The classification tree model we developed to predict hazardous NEOs shows strong overall performance, achieving an accuracy of approximately 89%. This indicates that the model correctly identifies the hazardous nature of NEOs in most cases. The confusion matrix highlights that while the model accurately classifies a large number of non-hazardous NEOs (55,335), it struggles more with the hazardous class, correctly identifying 4,923 hazardous NEOs but misclassifying 3,624 as non-hazardous. This imbalance is further reflected in the precision and recall for the hazardous class, both of which are around 57%, compared to 94% for the non-hazardous class. The model's weighted average f1-score of 89% suggests it performs well overall, but the lower precision and recall for hazardous NEOs indicate areas for improvement. The model tends to produce more false positives and misses some hazardous NEOs, which could be addressed through further tuning or incorporating additional features. Despite these challenges, the

model provides valuable predictions, and is effective in classifying non-hazardous NEOs. However, enhancing its ability to correctly identify hazardous NEOs would be an important next step in refining its performance.

IV. Random Forests

The performance of the Random Forest model before and after cross-validation (CV) shows a few key differences. Before CV, the model achieved a precision of 0.73, recall of 0.14, and an F1-score of 0.24 for class 1.0. After CV, precision for class 1.0 increased to 0.78, recall significantly improved to 0.37, and the F1-score rose to 0.51. This indicates that cross-validation helped the model better identify the minority class (hazardous), improving its ability to correctly detect positive instances. Accuracy remained consistent at 0.88 both before and after CV, indicating that the model's general ability to classify instances correctly did not degrade. In summary, applying cross-validation enhanced the Random Forest model's ability to detect the minority class (class 1.0), improving precision and recall for that class, while maintaining overall accuracy. The trade-off was a slight decrease in the weighted F1-score, but the model became more balanced in handling both classes.

V. Results

The best performing model overall was random forest, with an accuracy of 91%. Though not perfect, this means that the model was able to successfully classify approximately 9 out of 10 near-earth objects as hazardous or not hazardous accurately. Naïve Bayes was the worst performing model, with only 74% accuracy.

VI. Implications

As outlined in the problem statement, the early identification of hazardous NEOs is crucial to mitigating significant damage to Earth. The model's ability to classify a hazardous NEO has several important implications; it improves risk assessment, resource allocation, scientific research, and ultimately, increases public safety. Although the model cannot guarantee the identification of every hazardous NEO, it provides a valuable estimate for driving further evaluation. To enhance the model's predictive accuracy, integrating additional features, such as information about the NEO's composition, could improve its effectiveness in determining hazard potential.

Appendix

Figure 1

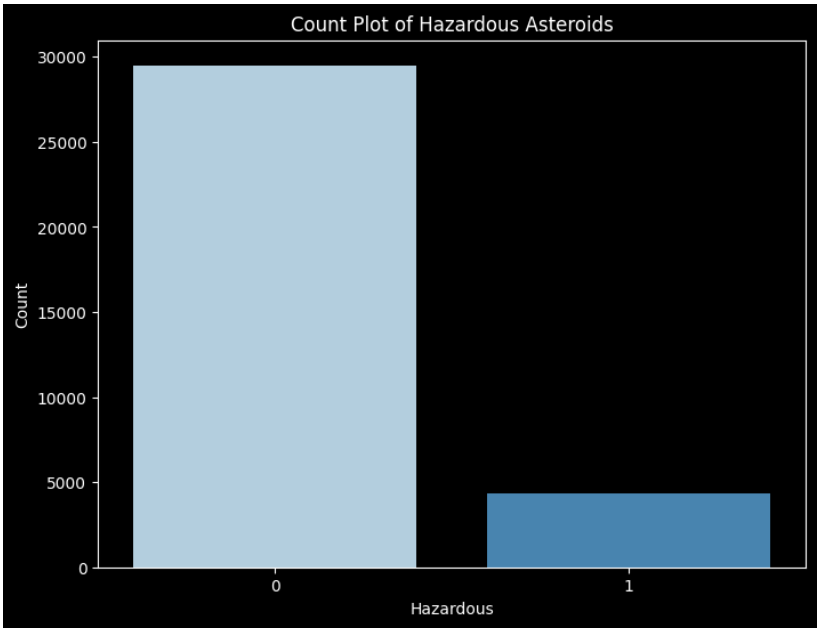

Figure 2

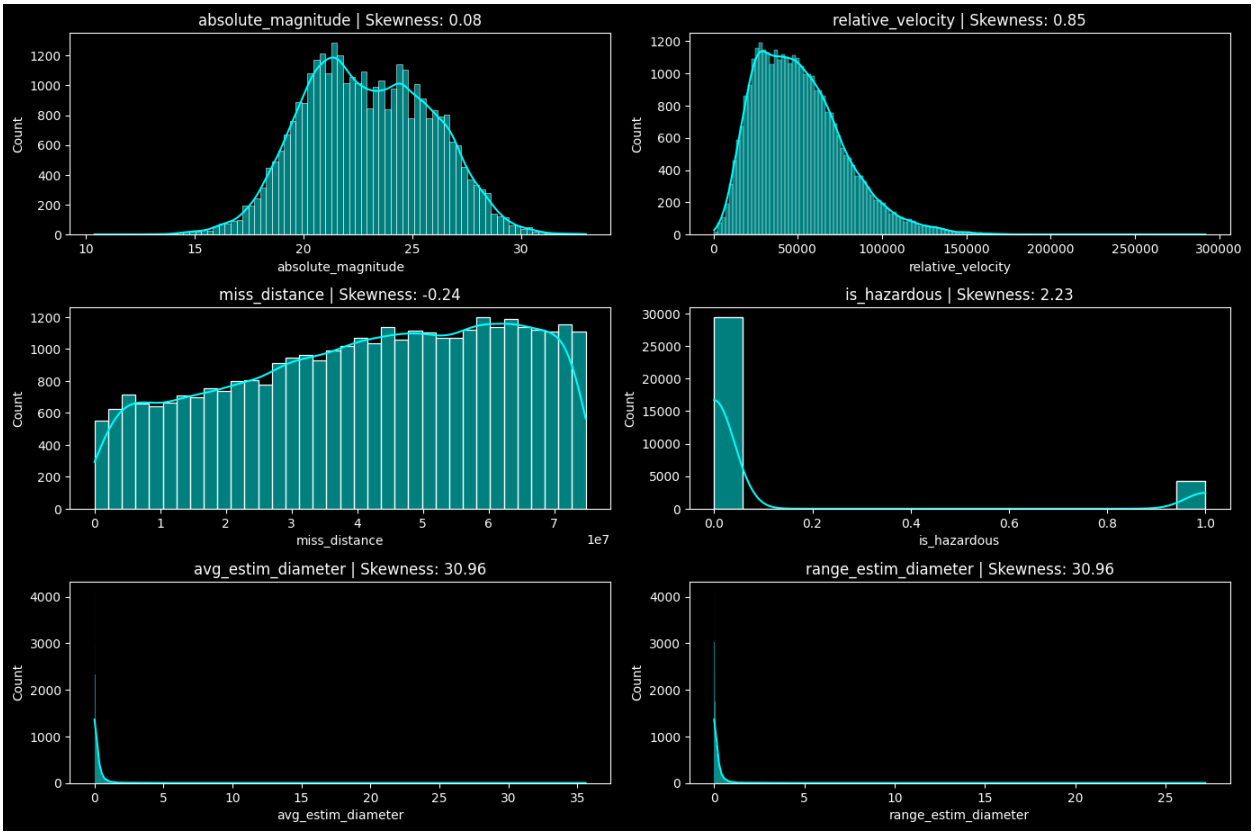

Figure 3

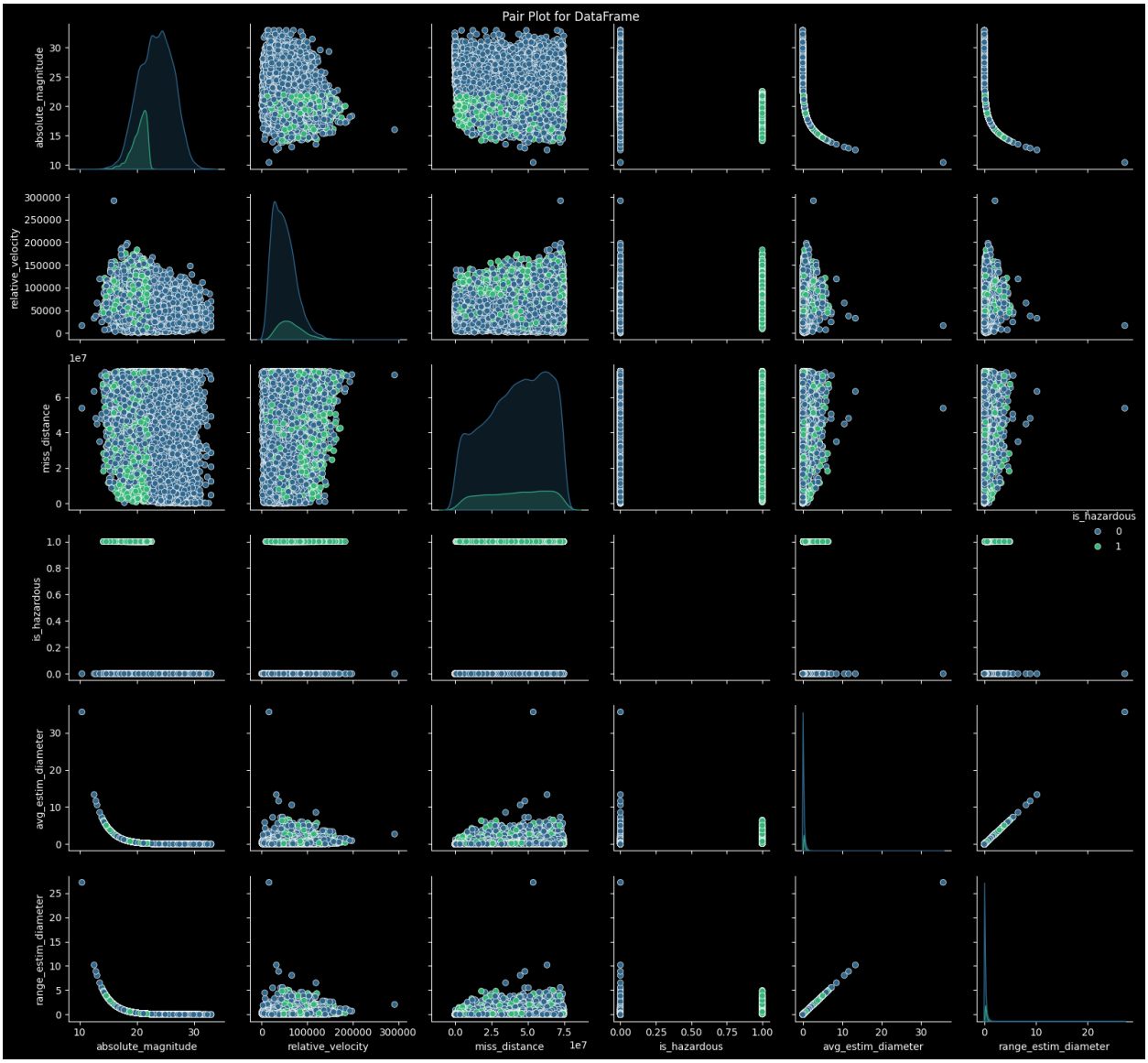
Pair Plot for DataFrame
absolute_magnitude
relative_velocity
miss_distance
1e7
is_hazardous
avg_estim_diameter
range_estim_diameter
is_hazardous
0
1

Figure 4

Figure 5