Rescue the Earth: Determining Hazard of Asteroids

Issac Chan, Lance Yuan, and Eva Wang

Introduction

For hundreds of years, scientists have been passionate about exploring the mysteries of the universe. Astronomers in the past started to formulate hypotheses and theories to show their curiosity about the universe. Indeed, understanding the nature of the universe is essential to the development of human beings. As technology advances, we are not satisfied with only observing the universe. By utilizing satellites and telescopes, we are able to make analyses and predictions of objects such as meteorites, asteroids, planets, etc. For this project, we are interested in asteroids, which are a common type of "rock" that might be harmful to our planet.

Asteroids are rocky items that orbit the sun (NASA Science, n.d.). One important feature of asteroids is that they are usually smaller than planets. As stated before, asteroids are not always treasures for human beings. For instance, one of the most convincing theories about the extinction of dinosaurs is related to asteroids. To make sure that this tragic history won't repeat itself, scientists endeavors to use the past records of asteroids to predict whether or not they will lead to a crisis on Earth and to what extent it will affect our planet. After all, we shouldn't always rely on the atmosphere to help us destroy asteroids.

Apart from that, the highly predictive features of asteroids are another reason why we chose to apply machine learning to the data. By inputting the size, composition, orbit, and other important characteristics of the asteroids, we are able to create a model that makes inferences about other asteroids and hasn't threatened us in the past (Rumpf, Lewis, & Atkinson, 2017). Compared with highly unpredictable objects such as stock prices, we can make much more accurate predictions for asteroids. This real-world problem is a good starting point for us to master the skills of machine learning.

Let's briefly look at the original dataset we chose and the models that we used to predict the potential outcomes of asteroids. There are 4687 observations and 40 variables for the dataset, which is large enough for us to perform model training and model testing. The data was collected by NASA, and we accessed it from Kaggle. Detailed information related to data preparation is written in the Data description part of our report. There are 7 parts of our report: Introduction, Related Work, Data Description, Methods, Results, Discussion, and Conclusion and Future work. The main purpose of this report is to take you through the entire process of our project, and also act as a reference for the code we wrote. After looking at the variables and other features of the asteroids dataset, we chose to construct our models using Logistic Regression, Neural Network, and XGBoost. Last but not least, we will compare the performance between models and evaluate important aspects, e.g. variable importance.

Related Work

We have already talked about the necessity and feasibility of our project. Not surprisingly, research and studies related to asteroids had already been done by many scientists. The first step is to find a sample of asteroids that could be a threat. Scientists found that the trans-Neptunian belt and the Oort cloud are the main sources of potentially hazardous asteroids to the Earth (Ipatov & Mather, 2004). Now, to our main focus on how to identify the hazard,

researchers found that some key characteristics of the asteroid, such as its size, its orbit, and distance, would impact the decision (O'Callaghan, 2021). Then, we need to eliminate this hazard to the Earth. A common practice today is to "use the gravitational force to change the orbit of asteroids" (O'Callaghan, 2021). It is always vital for governments and scientists to deal with this matter with extra caution. We all want to avoid the incident in 2013 when the Chelyabinsk meteor exploded over Russia and injured hundreds.

Besides the articles published in the magazine, we also found a research article that was published on the Nature, one of the most famous science journals in the world (Chapman & Morrison, 1994). This article reminds people that we are not completely safe under the protection of the atmosphere and the moon. Although most of the asteroids are pulled away by the moon's and other objects' gravity force or disappeared in friction with the atmosphere, there's still a 1/10000 chance that we will be hit by large asteroids or comets . The article further classifies the level of hazard from asteroids into three levels: fireballs and bolides, locally devastating impacts, and globally catastrophic impacts (Chapman & Morrison, 1994). Only by names, we can tell the danger to our home, which highlights the importance of our project. One of the advantages of our project is that we have more advanced technology than before. We can train a machine learning model by using past data from NASA and test the model performance on the test data. We can double-check whether our predictions are reasonable. In general, we are not only trying to do diagnosis analysis but are also trying to make predictions for asteroids. Also, in their article, they came up with possible solutions to deal with those hazardous asteroids.

Last but not least, there are hundreds of similar studies related to asteroids. The essence of this topic is how to interpret most of the useful features of asteroids correctly, what model performs the best in predicting asteroids, and how we can link them to the binary result (hazardous or not). In conclusion, asteroids are a popular topic, and they are worth learning.

Data Description

For our analysis, we found NASA data from Kaggle which includes the situations of 4687 asteroids from 40 variables. Variables cover the key components of the asteroids, such as their diameter, speed, and distance, which are helpful to determine if the asteroid is dangerous or not. For a detailed description of each variable, please consult the R Markdown file. Among them, "Est.Dia.in.KM.min.", "Est.Dia.in.M.min.", "Est.Dia.in.Miles.min.", and "Est.Dia.in.Feet.min." are the same value in different units; variables on the maximize diameters have the same issue. For the sake of our analysis, we will set our units in kilometers and only use "Est.Dia.in.KM.min." and "Est.Dia.in.KM.max.". To maintain consistency in units, we will only use "Relative.Velocity.km.per.hr" to assess the relative speed of asteroids. For miss distance, we will only use "Miss.Dist..Astronomical.". Apart from these, "Neo.Reference.ID" and "name" are the same reference ID for each asteroid; thus, we consider them not relevant in determining the hazard of asteroids. "Equinox" and "Orbiting.Body" only have one value for all observations, which are J2000 and the Earth correspondingly. We conclude that neither will be solid predictors for our models. "Orbit.Determination.Date" and "Close.Approach.Date" are two date variables. Since we don't know when the data was observed, dates may be less relevant to our analysis. With this, we narrow our focus to 24 variables (including one response variable).

This dataset has no missing or repeated data. The "Hazardous" is determined as the binary categorical response variable which labels whether the asteroid is hazardous or not

(labeled as True or False). There are 755 observations of hazard and 3932 observations of non-hazard, which makes an imbalanced dataset. We will use the stratified function from the splitstackshape package to split the data into 80% training and 20% test sets with an equal proportion of hazardous and non-hazardous observations in each set. Based on the boxplot, we can predetermine that "Absolute.Magnitude", "Orbit.Uncertainity", and "Minimum.Orbit.Intersection" may be strong predictors (see Figure 1).

Methods

To generate reliable predictions, we used three models: Tree, Regression, and Neural Network. For the tree model, we used Random Forest and XGBoost. With the boxplot, we know that certain variables are stronger predictors than others. To avoid highly similar trees, random forests overcome this problem by forcing each split to consider only a subset of predictors. However, the improvement of performance is highly dependent on the number of trees and more trees usually require longer training time. XGBoost works similarly to bagging except the trees are built sequentially. Usually, XGBoost has the highest predictive power among all trees, but they are sensitive to outliers which may be a concern for our case.

Regression is usually an easy model to train and interpret. We used logistic regression, backward selection, and LASSO. However, one vital disadvantage of logistic regression is the assumption of linear relationships between predictors and the response variable, which we are doubtful these relationships would have. Backward selection is overly dependent on p-values to drop variables, but p-values have issues. LASSO is a combination of both shrinkage and selection of variables. However, LASSO selects at most n features and the selected features are highly biased. The Neural Network model has highly flexible for both regression and classification problems and may generate reliable predictions. But it is extremely hard to interpret, computationally very expensive, and time-consuming to train. In the end, we will compare the confusion matrixes and plot ROC to compare which model is the best.

Results

To maintain consistency among analyses, all seeds are set at 123456. Random forests were first modeled and since it generates very high accuracy at 99.47% with 200 trees and 200 nodes, no additional tuning was conducted. All predictors were fit to the XGBoost matrix. XGBoost was set to 100 rounds and resulted in high accuracy at 99.57%, indicating that imbalanced data didn't affect our prediction. SHAP scores were calculated and "Absolute.Magnitude" and "Minimum.Orbit.Intersection" are the top two important predictors. After removing those two variables, XGBoost still generates accuracy at 94.24%. With the high accuracy rate, no additional tuning was implemented. Thus, among all the tree models, XGBoost generates the highest accuracy at 99.57%, sensitivity at 98.68%, and specificity at 99.75%.

Logistic regression and backward selection were applied to the training data. Backward selection kept only 4 predictors: "Absolute.Magnitude", "Est.Dia.in.KM.min.", "Est.Dia.in.KM.max.", and "Minimum.Orbit.Intersection", which half of it coincided with SHAP calculation. And LASSO generated similar results. With the best lambda at 0.1, LASSO cross-validation determines that "Absolute.Magnitude", "Orbit.Uncertainity", and "Minimum.Orbit.Intersection" are key predictors. In the end, logistic regression resulted in a higher accuracy at 96.91%, sensitivity at 88.47%, and specificity at 98.47%.

Before training the neural network, data standardization was first conducted for both the train and test data. The challenge of the neural network is to determine the number of hidden layers and hidden neurons. Considering what we have learned from statistical learning, we believed that two hidden layers should be sufficient to generate reliable predictions. We first tested a neural network with 1 neuron on each layer, but the results were not satisfying: an accuracy of 96.26%, a sensitivity of 84.77%, and a specificity of 98.47%. Considering two hidden layers would include too many possibilities to try out manually, a for-loop was run to determine the best combination of hidden layers which can generate the highest accuracy. We limited each layer to have at least two neurons and at most 10 neurons to save some computational power. With this, we found that two hidden layers with 5 neurons on the first layer and 4 neurons on the second layer would generate the best results. Thus, the neural network has an accuracy of 99.68%, a sensitivity of 98.68%, and a specificity of 99.87%.

After comparing the confusion matrixes among three models (i.e., XGBoost, logistic regression, and neural network), we found that the neural network has the highest accuracy at 99.68%; XGBoost and neural network have the highest sensitivity at 98.68%; the neural network has the highest specificity at 99.87%. With the ROC plot, XGBoost has the highest AUC at 0.998 (see Figure 2). In general, our dataset generates very strong predictions for all three methods. Since all confusion matrixes were made based on test data, there is no issue of overfitting. And the imbalanced response variable didn't impact our results. Considering the interpretation matter, we select XGBoost as the best model for predicting the hazard of asteroids. Among all 23 predictors, we concluded that "Absolute.Magnitude" and "Minimum.Orbit.Intersection" are the most important predictors across all models.

Discussion

XGBoost is the most suitable machine learning model for predicting the hazard of asteroids with the dataset that we have now. One of the main reasons why XGBoost fits this data set the best is because we can determine the most significant variable with a high-accuracy model. By finding out the importance of variables, we can amend the model by using those significant variables or taking them out to form a new model. These two ways can help us find out more insights about the dataset and models created. After taking out those variables that are nearly perfect, we can determine if the model is solely dominated by the two variables. We can eliminate the chance of failing to predict the outcome accurately when there are missing variables. Also, we may simplify the model to fewer variables to create a model with a parsimonious structure. Nevertheless, logistic regression also does a great job of finding significant variables and simplifying the model. Therefore, XGBoost is the best model for our dataset because it has a parsimonious structure with high accuracy.

Logistic regression does not perform as well as the other two models with this data set. It still has high accuracy, but it is about 3% off from XGBoost and Neural Network. However, it is still useful for creating a simple model for this dataset. Backward selection can easily help us simplify the model to have significant variables only. With LASSO cross-validation, we find that there are multiple important variables that we should focus on. Logistic regression is a simple model that allows us to observe important insights from the dataset. For that reason, we should use it as the first step in our project to get a direction before exploring deep into the data.

The neural network performs really well with this dataset. It has extremely high accuracy and specificity with this data set. However, it is not an explainable model with a simple neural network, and the layer units are generated in a random state that we cannot control. Therefore, a simple neural network may not be a suitable model for predicting if the asteroid is hazardous, as it is not easy to explain. Some higher-level neural networks may be able to predict even better results in an explainable way. It may also spot some other important features from our dataset that can benefit our project, but we do not have that knowledge right now. So, we ended up using the simple neural network, which gives us an unexplainable model that has high accuracy.

As we have mentioned above, variable importance is another focus of this project. By identifying significant and insignificant variables, we can create better models with a simpler structure. According to our testing result generated by the SHAP value (see Figure 3), we found that the two most significant variables are the absolute magnitude and minimum orbit intersection. Besides using the SHAP value, we also ran LASSO to find out important variables. We discovered that LASSO's outcome matches with the outcome in SHAP value. Absolute magnitude, minimum orbit intersection, and orbit uncertainty are the three variables with the most significance in the model.

These three variables are also part of the variables that we identified as important variables when we overviewed the variables at the start of the project. After identifying these three variables, we will alert any new discovery of asteroids with extreme values in these variables. We may create alerting systems that help us to quickly identify hazardous asteroids and take appropriate actions.

Last but not least, spotting false negatives should be another focus if this project goes on because false negatives would cost the most to us. We may face a really bad situation if we fail to spot all dangerous asteroids. Therefore, multiple types of models and different model settings should be created to eliminate the chance of leaving out any false negatives.

After spotting all the dangerous asteroids, monitoring and planning are the next phases of this project. We may use other machine learning models to predict the route or movement of these dangerous asteroids, as well as predict the best timing and methods to protect the Earth from asteroids.

Conclusion and Future work

In conclusion, we recommend the XGBoost model for the future prediction of hazardous asteroids. Scientists should pay extra attention to variables on the magnitude of an asteroid, the minimum orbit intersection risk, and the uncertainty of the asteroid's orbit. Above all, there are more things that we can do better in this project, such as finding the best number of layers and units on average for neural network models by running the model randomly many times. On the other hand, XGBoost has room for improvement also, including increasing the number of rounds run by the model. More than that, we can also fit a better model by tuning a better eta in XGBoost. By going through these steps, we can create a model that is more suitable for predicting if the asteroid is hazardous or not.

Contribution

- Lance: trained and tested regression models, drafted the Introduction and Related work section, edited the entire report
- Issac: trained and tested neural networks, drafted the Discussion and Conclusion sections, edited the entire report
- Eva: trained and tested tree models, drafted the Data description, Methods, and Results sections, edited the entire report

Bibliography

Chapman, C., & Morrison, D. (1994). Impacts on the Earth by asteroids and comets: assessing the hazard. *Nature 367,* 33–40. doi:10.1038/367033a0.

Ipatov, S.I., & Mather, J. C. (2004). Comet and asteroid hazard to the terrestrial planets. Advances in Space Research, 33(9), pp. 1524-1533. doi: 10.1016/S0273-1177(03)00451-4.

O'Callaghan, J. (2021). How do we know if an asteroid headed our way is dangerous? *Horizon, The EU Research & Innovation Magazine*. Retrieved from https://ec.europa.eu/research-and-innovation/en/horizon-magazine/how-do-we-know-if-asteroid-headed-our-way-dangerous.

Rumpf, C. M., Lewis, H. G., & Atkinson, P. A., (2017). Asteroid impact effects and their immediate hazards for human populations. *Geophysical Research Letters, 44*(8). doi: 10.1002/2017GL073191.

What Is an Asteroid? (n.d.). *NASA Science.* Retrieved November 21, 2022 from https://spaceplace.nasa.gov/asteroid/en/.

Figure 1. The boxplot among all predictors and the response variable

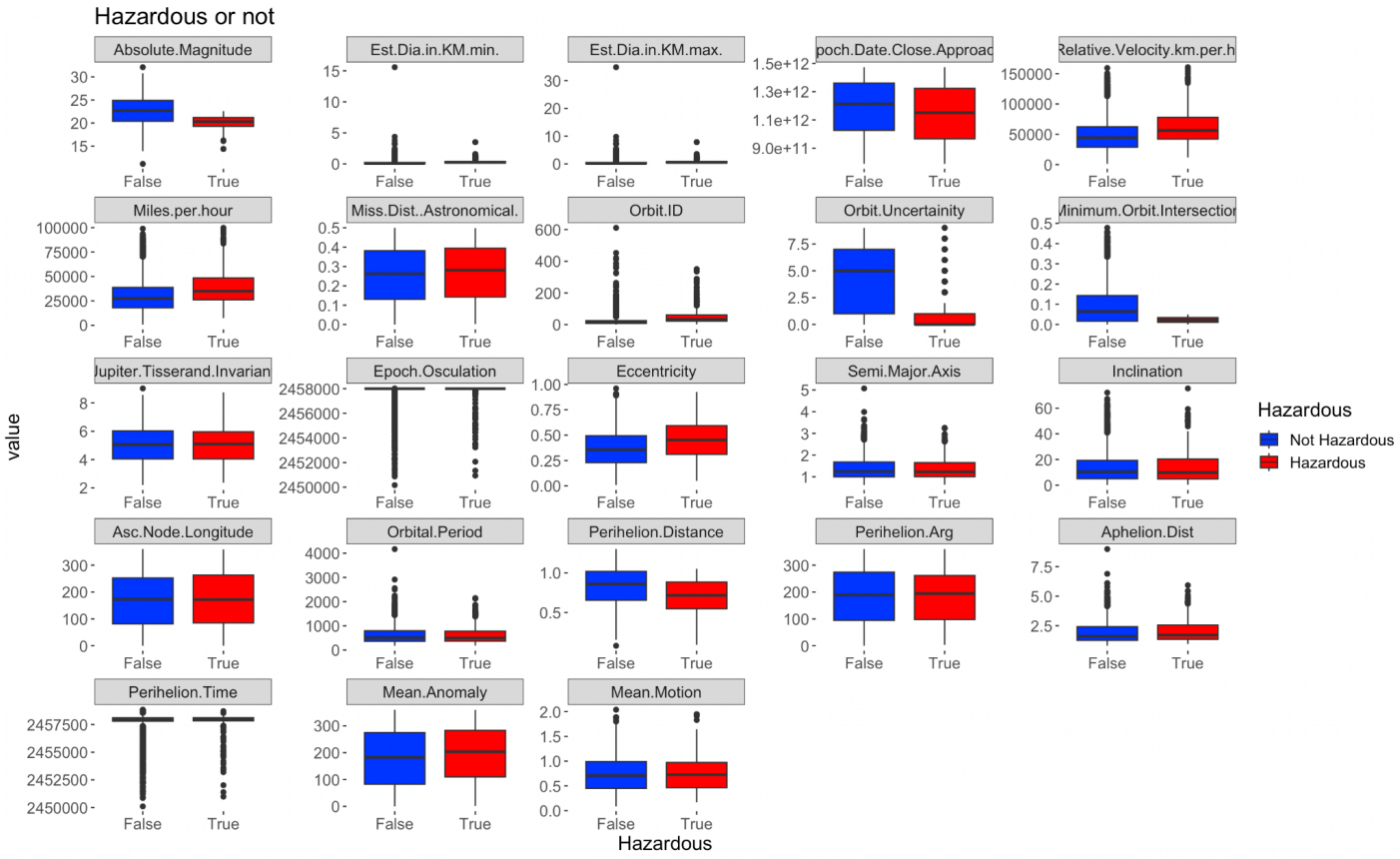

Figure 2. The ROC plot among XGBoost, logistic regression, and neural network.

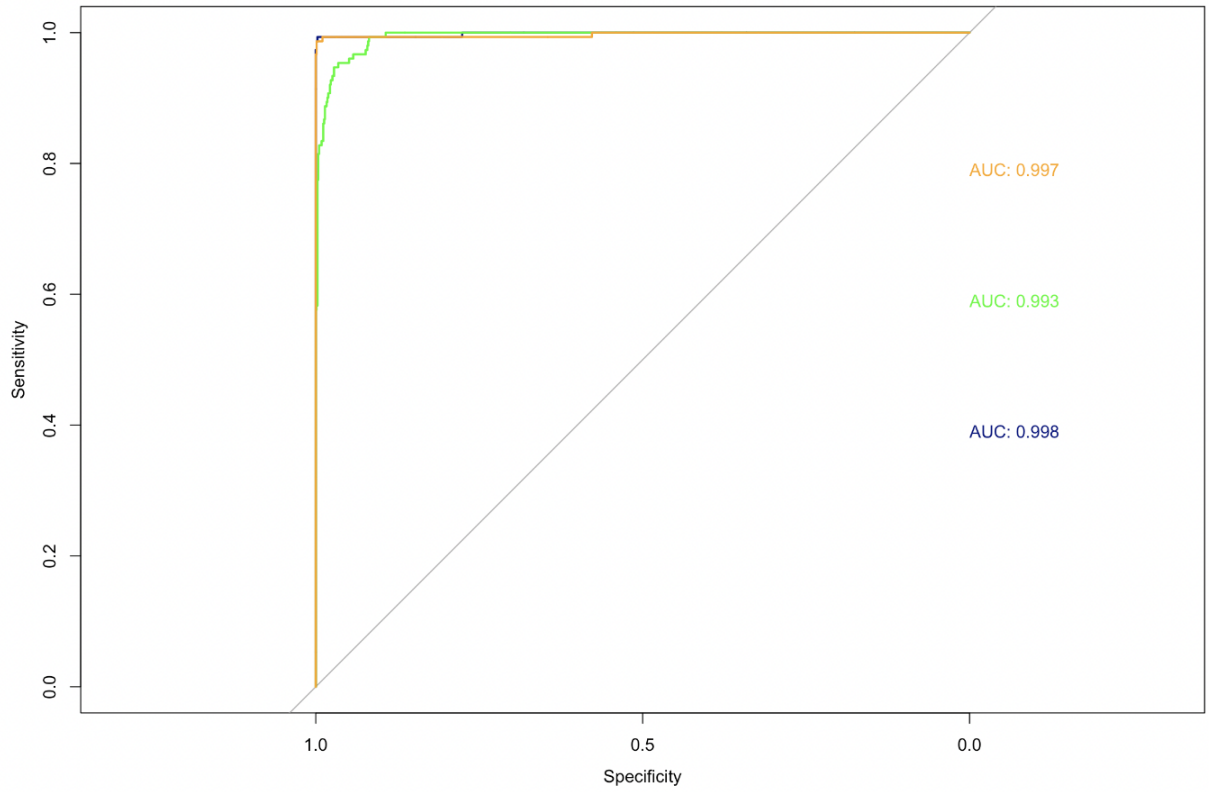

Figure 3. SHAP values on variable importance.