Project 8 Report

Michael Hanks & Linhao Wu

Github: https://github.com/MichaelPHanks/CS5830Project8
Slides: CS 5830 Project 8 Presentation

1. Introduction

Since the birth of the planet we live on, about 4.6 billion years ago, it has had a large bombardment of asteroids that somehow make their way to impact Earth. This even includes an asteroid impact that caused the extinction of dinosaurs. Because of the threat that asteroids cause to humans and the planet, we have started gathering more information about the ones in our solar system. We are trying to be able to, given the asteroids information, know whether it is potentially hazardous or not. We use a dataset that contains many attributes related to positioning of asteroids close to Earth. The models for predicting that we use are neural networks and decision trees. We believe this information and prediction study is relevant to all scientists and also the public, given this affects everyone on the planet.

2. Dataset

The dataset we are working with is from NASA. This contains information of about 90,000 asteroids (smaller bodies of rock that orbit the sun). Keep in mind that only about 8,000 pieces of the data are considered ‘hazardous’. Each row is an asteroid that has been located. The attributes given in the dataset are as follows: id, name, minimum estimated diameter, maximum estimated diameter, velocity relative to Earth, miss distance, intrinsic luminosity, and hazardous. The classification we are aiming for is the hazardous attribute. There are a few more attributes, but they are not needed because they all have the same value. For the model, id and name are not going to be used because they are irrelevant for predicting if an asteroid is hazardous or not. Due to its rich features and accurate labeling, this dataset is well suited for exploratory data analysis and machine learning model training, especially when performing classification tasks. In the data preprocessing phase, first confirm that there are no missing values in the data set. Next, the Boolean features sentry_object and hazardous are converted to integers (0 for False and 1 for True) so that the model can process the data correctly. We removed about 74,000 pieces of data to show a balanced analysis. We just removed the first 74,000 that had false for the hazardous attribute.

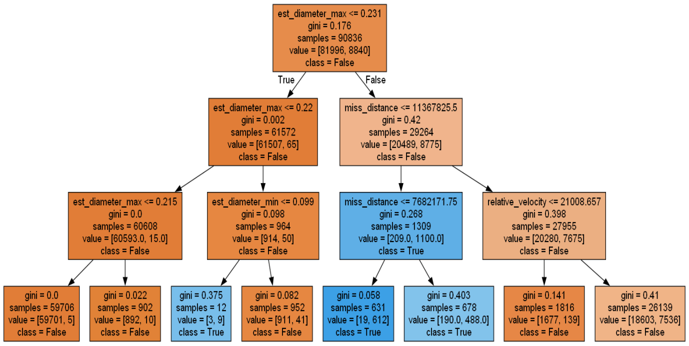
3. Analysis Techniques

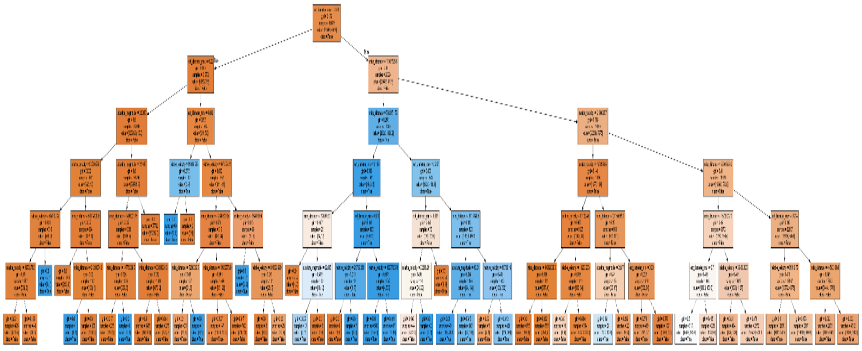
The analysis techniques that were used, as briefly described above, are decision trees and neural networks. These models will give direction in terms of predictability of the dataset. Decision trees are more like human-based decisions, which can give us some insight for how the asteroids are chosen for being hazardous. We also used the f-scores, which are values considering both precision and recall, for our evaluations of the models. With the decision trees, we used two different max depths for the tree to check the differences in predictability. Simultaneously, we utilized neural network technology based on the TensorFlow and Keras frameworks. Through experimental design, we constructed three different architectures of neural networks, ranging from simple single-hidden-layer structures to more complex multi-hidden-layer structures, aiming to explore the impact of network depth and width on prediction performance. These networks are not only suitable for handling complex relationships between multidimensional features and target variables but also capable of effectively solving binary classification problems through meticulous adjustment and optimization. In terms of code implementation, we focused not only on the construction and training process of the model but also placed a special emphasis on the importance of performance evaluation, including calculating accuracy, precision, recall, and F1 scores, providing essential metrics for a comprehensive understanding of model efficacy. Importantly, by monitoring the loss and accuracy during the training and validation phases, we detailed the model's performance at different stages, thereby identifying and preventing overfitting phenomena to ensure the model has good generalization ability. Moreover, by visualizing the structure of the neural network and the changes in performance during the training process, we not only enhanced the transparency of the analysis but also improved the interpretability of the results.

4. Results

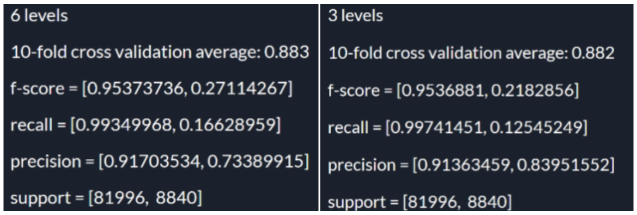
Decision Trees:

When applying decision trees to the dataset, there were some surprises in the results. We applied two different level depths, which were 3 and 6. For level 3, our 10-fold cross validation gave us 0.882 as a whole. The f-scores were 0.95 and 0.21. For level 6, the 10-fold cross validation gave us 0.883 as a whole. The f-scores were 0.95 and 0.27. The rest of the results are shown below. The precision did good, however recall (which is more important in this scenario), extremely struggles. These results show good f-scores, but poor predictability for what we were looking for. The model severely struggles for finding actual hazardous asteroids to be hazardous in the predictions. The level 3 tree is shown below. Given that the dataset was extremely lopsided, we decided to remove most of the non-hazardous entries to balance the dataset to get an idea of how much better we could get in terms of predictions. The results are shown below.





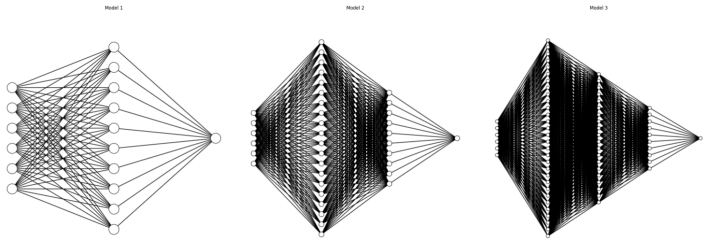
The results here are much better than before. However, in a situation like this, we cannot use the second analysis we just did. There are a lot more non-hazardous asteroids compared to hazardous asteroids, and that’s what makes this analysis so hard to predict. With respect to the levels of the results, the more levels, the higher the f-scores. What is not shown is a test run with a max depth of 25 levels, which gave us a 10-fold cross validation of 0.40, which is much higher than the depths ran at in this analysis. However, this can cause overfitting of the data. It likely is not capturing the ‘unseen’ data. With regards to bias and variance, bias will decrease and variance will increase with high depths. Bias will decrease because it can capture the finer details, but variance will increase because the model would be more sensitive to small fluctuations of the data. You may also notice that the algorithm is choosing different attributes to compare on at the same levels. Why is that? Well, the algorithm is choosing the best attribute, out of all attributes, that will reduce impurity for the next nodes. That is, it tries to discern the best way to separate the two classes.



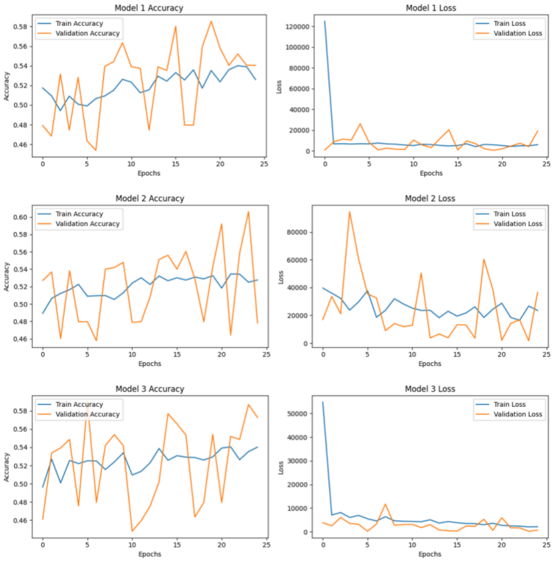
```
6 levels
10-fold cross validation average: 0.883
f-score = [0.95373736, 0.27114267]
recall = [0.99349968, 0.16628959]
precision = [0.91703534, 0.73389915]
support = [81996,  8840]
```

```
3 levels
10-fold cross validation average: 0.882
f-score = [0.9536881, 0.2182856]
recall = [0.99741451, 0.12545249]
precision = [0.91363459, 0.83951552]
support = [81996,  8840]
```

Neural Networks:



For model 1, the training accuracy and validation accuracy fluctuated greatly during the training, indicating that the model had some instability when fitting the data set. For model 2, the training accuracy and verification accuracy are improved compared with model 1, and the fluctuation range is reduced, indicating that the stability of the model in the training process is improved.The decline trend of training loss and validation loss is relatively stable, but the increase of loss at some points indicates that there may be overfitting phenomenon. For Model 3, the training and validation accuracy improved further, and the fluctuations in validation accuracy became smoother, possibly due to deeper network structures that better capture complex relationships in the data.



Model	Accuracy	Precision	Recall	F1-score
Model 1	0.550475	0.536645	0.998289	0.698045
Model 2	0.481888	0.528571	0.042213	0.078183
Model 3	0.581354	0.559036	0.926412	0.697295

The loss plot shows a continuous decline and a relatively stable trend, which indicates that model 3 has good consistency and generalization ability when fitting the data. All three neural network models performed with slightly higher accuracy than random guesses, while Model 3 performed best, showing a more balanced accuracy and recall. The high recall rate of Model 1 indicates that it tends to predict more positive classes, but it also brings more false positives. Model 2, although with greater accuracy, has an extremely low recall rate, which suggests that it is too conservative in labeling positive classes. Model 3 improves accuracy while maintaining recall compared to the other two models, showing the advantage of the deep network for this task. Although Model 3 performed the best, the overall results show that the model still needs to improve performance through more detailed feature engineering, network architecture adjustments, and data processing techniques.