Using Machine Learning Algorithms to Model Planetary Collisions and Predict the Masses of the Largest Remnants

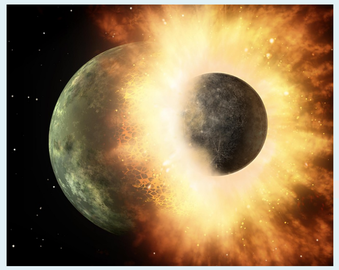
Emaad Paracha[1], Diana Valencia[1,2], Alan Jackson[3]

(1)Physics and Astrophysics, Department of Physical & Environmental Sciences, University of Toronto Scarborough

(2) Astronomy Department, University of Toronto, (3) Centre for Planetary Sciences, University of Toronto

Introduction

Planetary formation is an integral part of planetary science and involves the processes that lead to the creation and formation of planets. The best way to study the whole process thoroughly involves N-Body Simulations, computerized simulations of planets as particles with certain features and under certain forces. A very large percentage of planets owe their formation and creation to collisions.



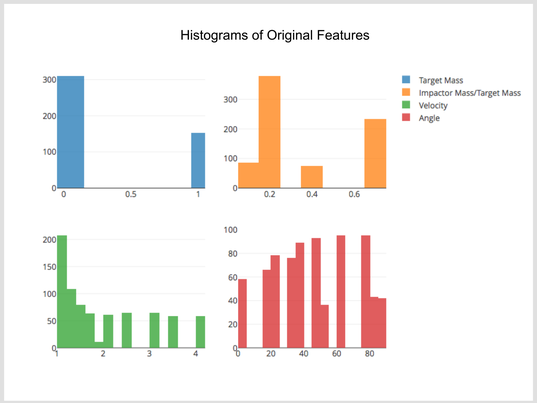
These collisions can vary substantially with respect to multiple parameters, including masses of the bodies involved, impact velocities, impact angles, and composition of the embryos, and variation in such parameters can result in many different outcomes. With the use of N-Body simulations, we can study different collisions and outcomes. Where we look towards Machine Learning is to understand the parameters and their relationships, and make predictions based on a model. This can help make simulations more efficient and unravel relationships between the parameters, ultimately developing a better and faster way to study planetary formation and collisions.

Objective

The objective of this research was to study planetary and gravity-dominated body collisions and to create a model using various machine learning algorithms that would use basic parameters such as masses of the bodies, impact velocity, and impact angle and predict the masses of the largest remnants of the collisions. These machine learning models would first be trained on data from N-Body simulations, and then refined to be able to give an accurate estimation of the remnants of the collisions.

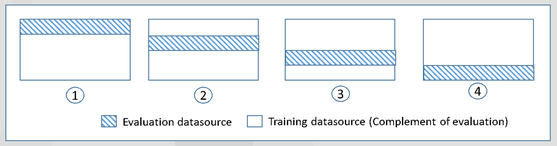
Methods

Data from N-Body simulations of collisions with different parameters (the mass of the two bodies, impact angle, impact velocity) was used. This dataset was split into a Train Set containing 75% of the data and a Test Set containing the rest of the 25%. scikit-learn and XGBoost were used to create the regression models. They were trained on the data in the Train Set, tested on the Train Set using Cross Validation to check the efficiency of the models. Additional features were created based on scaling laws provided in research articles Leinhardt & Stewart 2012 and Stewart & Leinhardt 2012.



Methods cont.

Three different "targets" were chosen that were to be predicted, the mass of the largest remnant normalized to total mass, normalized to target mass, and accretion efficiency. Grid Search was used to find the best parameters for the machine learning algorithms, and the most optimal features and parameters were used to create the best models for the different targets.



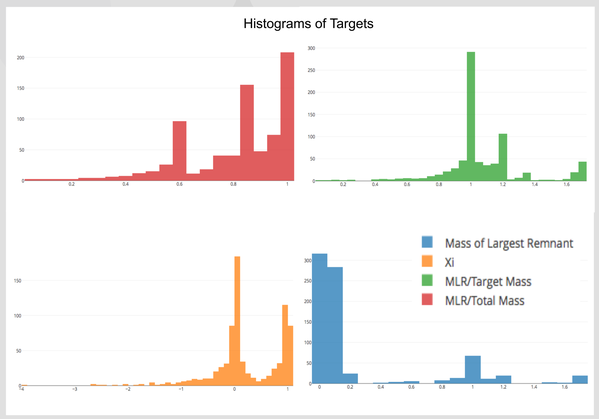
Cross Validation

Results and Discussion

Three different targets were chosen, the masses of the largest remnants normalized, once to the total mass and once to the target mass only. Another value that was predicted was the Accretion Efficiency, ξ:

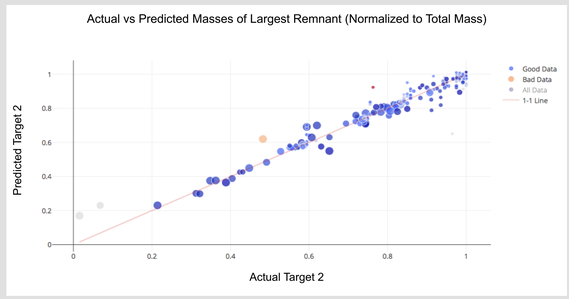
$$\xi = \frac{M_{LR} - M_{targ}}{M_{Imp}}$$

Initially, the models were trained and tested with just the 4 initial features



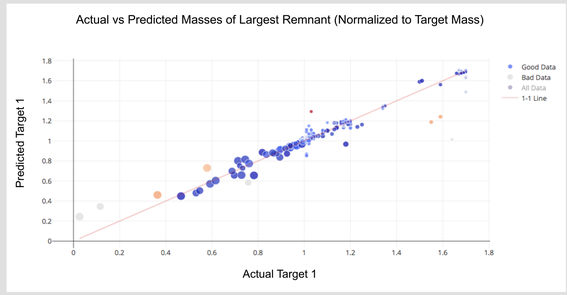
Histograms of the initial machine learning targets

With 10-Fold Cross Validation, the Mean MSE showed that Target 2 was being more accurately predicted. 4.26% of the predicted Target 1 values (11 datapoints) had more than 20% error, 4.26% again for Target 2, and 38.91% (122 datapoints) for Target 3.



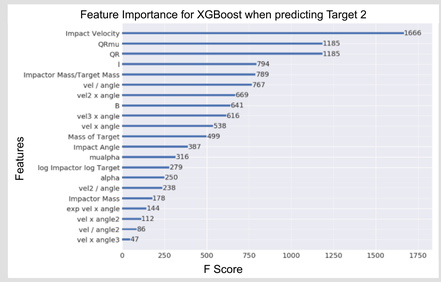
The MSE for MLR Normalized by Total Mass was 0.003274, 1.94% of the test data was outside a 20% error range.

Results and Discussion cont.



The MSE for MLR Normalized by Target Mass was 0.008026, 3.49% of the test data was outside a 20% error range.

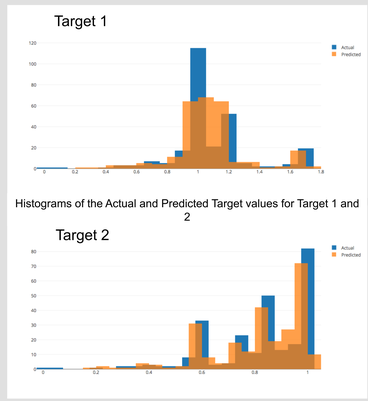
Additional features were then created and used. With the Test Set, 3.49% (9 datapoints) of the data had more than 20% error for Target 1, 1.94% (5 points) for Target 2, and 37.21% (96 points) for Target 3.



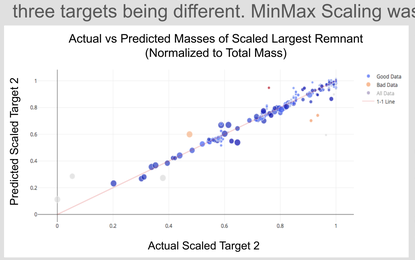
Similar features were noted as important for Target 1 and 3 as well

The model performed best when given the 4 initial features, energy, the size of the mass fraction of the impactor impacting the target and its energy, and combinators of impact angle and velocity.

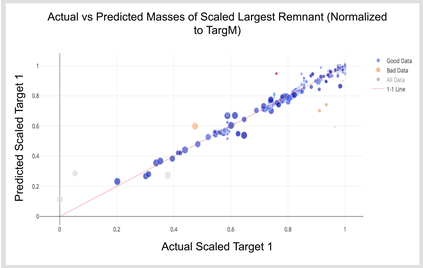
Histogram plots also show the discrepancy between actual and predicted values for Target 1 and 2

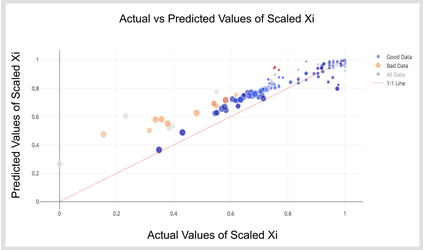


However, the scaling could have been an issue with the bounds of all three targets being different. MinMax Scaling was used to normalize the all three Targets between 0 and 1, and when the Test Set was predicted, Target 2 had 3.10% (8 points) of the points above 20% error, 3.49% (9 points) for Target 1, and 7.75% (20 points) for Target 3.

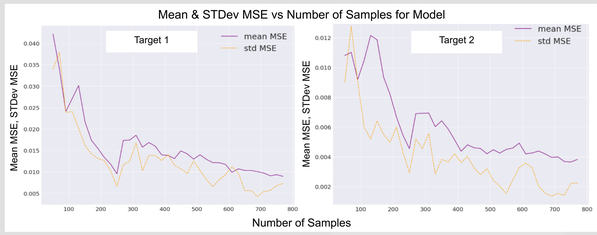


Results and Discussion cont.





Target 2 was most important with additional features, and Target 3 was the most inaccurate. Graphs of change in mean and standard deviation of MSE vs the number of samples the model is trained on, and the graph did seem to dip down, however the trend and fluctuating plateaus showed that more data would help make the model more efficient and better.



Conclusion and Further Work

Even though the XGBoost model that predicted Target 2 ended up predicting the values most accurately, the model is not 100% accurate.

As can be seen from the Mean and STDev for MSE vs Number of Samples graphs, **it also shows that the models could benefit from more data inputted** that would help make the model stronger and have it predict more efficiently.

Machine Learning algorithms are known to be stronger with more Gaussian fits. The data and the targets aren't representative of normal distributions, and that is something to be further looked into, **to make the data fit more of a normal distribution**, for more accurate analyzation.

Additionally, the mass of the second largest remnant will also be used to **classify different regimes of collisions** and then trained upon to create multiple more refined models.

The next steps are now to classify the regimes of collisions, with more data, and to have the targets scaled to more Gaussian fit to further enhance this project and to find more efficient machine learning models for the predictions.

References

1. Leinhardt, Z., & Stewart, S. 2012, ApJ, 745, 79
2. Chambers, J. E. 2001, Icarus, 152, 205 3
3. Kokubo, E., & Ida, S. 2002, ApJ, 581, 666
4. Asphaug, E. (2010). Chemie Der Erde - Geochemistry, 70(3), 199-219
5. Stars Collision Planets [Digital image]. (2010, December 5). Retrieved September 25, 2018.