Predicting Diameter and Physical Harm of Asteroids using Machine Learning

Procedures

Dataset

The dataset for this project is from kaggle, taken from a NASA JPL (Jet Propulsion Laboratory) database. The columns describe the physical and basic properties of an asteroid along with their diameter and label for whether they are physically hazardous to earth. The diameter feature will be used as the regressive target and the physically hazardous feature will be modeled for classification. There are over 800,000 data points and 31 features in this dataset.

Regression

Regression models will be utilized to predict the diameter of asteroids. The models being considered are SGD, Random Forest, KNN Regressor, and a voting regressor to average the results and balance out the bias within each method.

Classification

Classification models will be utilized to predict whether the asteroid is or is not physically hazardous to earth. The models being considered are KNN, Decision Trees, SVC, and a voting classifier to choose the majority rule to have a higher chance of classifying such an important feature.

Team Responsibility

Each member of my team oversees two models, one regression and one classification. These are still being determined as we clean and learn more about the data. My personal role is to lead the project as well as my two models as well as create the voting models to run through the HPC cluster. The Gantt chart in Figure 1 is shown below for the project. My main priorities will be Random Forest Regressor, voting method for Regression, finding the best way to balance the classes, and the Voting method for Classification. The other teammates will oversee their models as well as help with data cleaning, tuning and validation.

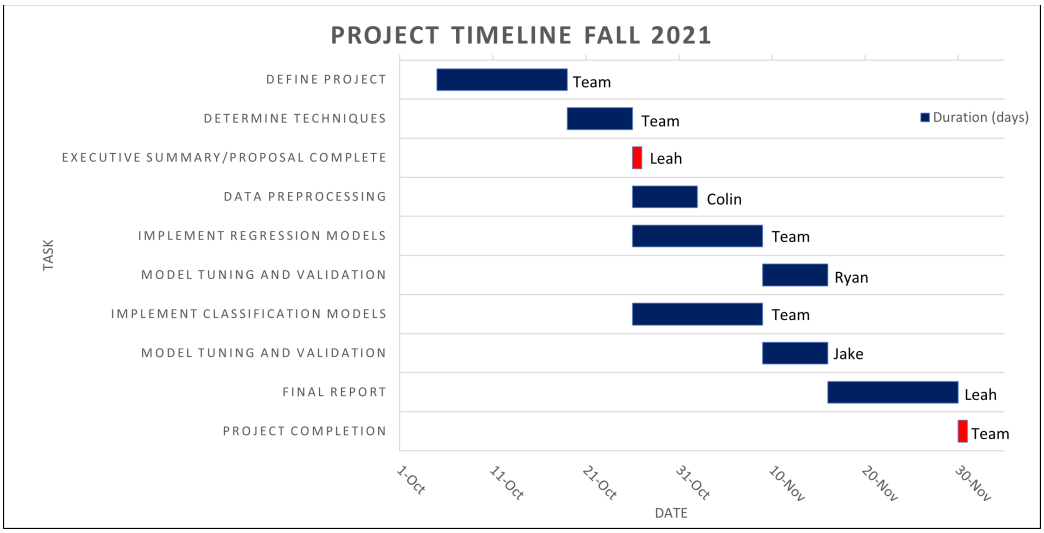

Figure 1: Gantt Chart for Project

Validation Plan

Data Preparation

To test the efficiency of the models, each model will go through their own hyperparameter tuning using gridsearchCV to get the best possible models. To deal with the large number of null values in the dataset, especially for the classification models, a comparison between different dimension reduction algorithms such as PCA and LDA. Some handling will be needed for the class imbalance of "Y" vs. "N" when determining if they are hazardous. We will look at sampling with replacement, stratified shuffle split and a variant autoencoder to balance "Y" labels and decrease the probability of bias in our classification models

Model Efficiency

For the regression models R^2, Mean Squared Error and the cross-validation score will be used to see how close our predictions are to the true values. The classification models will create a confusion matrix along with recall, precision and F1 scores.

Future Steps

Once we have a working model with our dataset from Kaggle, we will grab a larger set of data from the source to run with larger sample sizes for training, validation, and testing to see how that effects our models. Hopefully we can also run the models in the LEAP cluster once we have finalized the models and want to run with a larger amount of data.