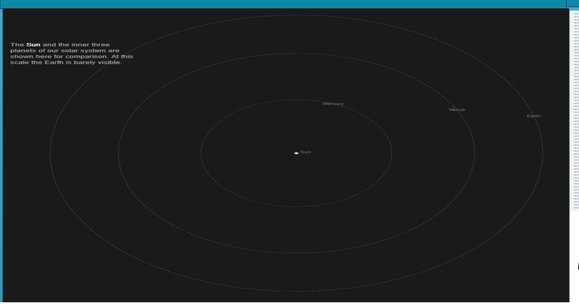
Planet Classification

Elijah Andrushenko & Jake Gergen

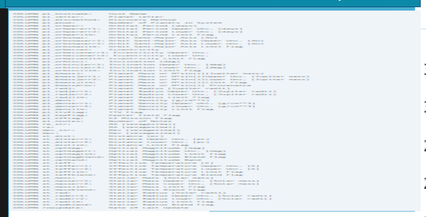
Objective

Our object with this project is to take the Kepler exoplanet database to train two different classifiers and classify unseen exoplanets as either "In the habitable zone" or "Not in the habitable zone". We chose to use two different classifiers because we wanted one to be a high bias model and the other to be a high variance model. So we chose a perceptron and a random forest ensemble.



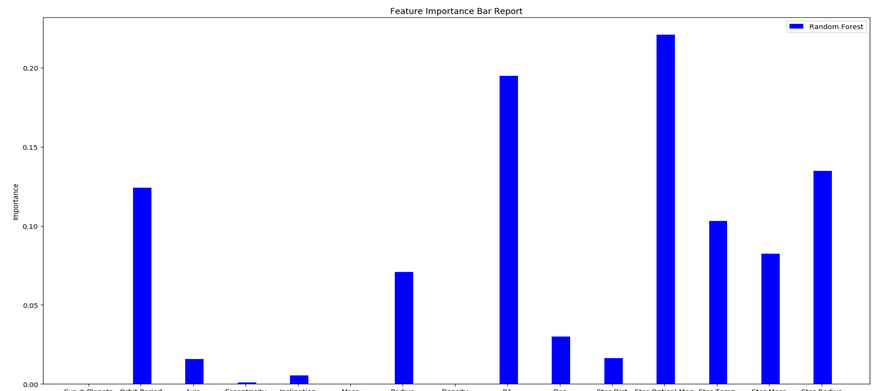
K Means Clustering

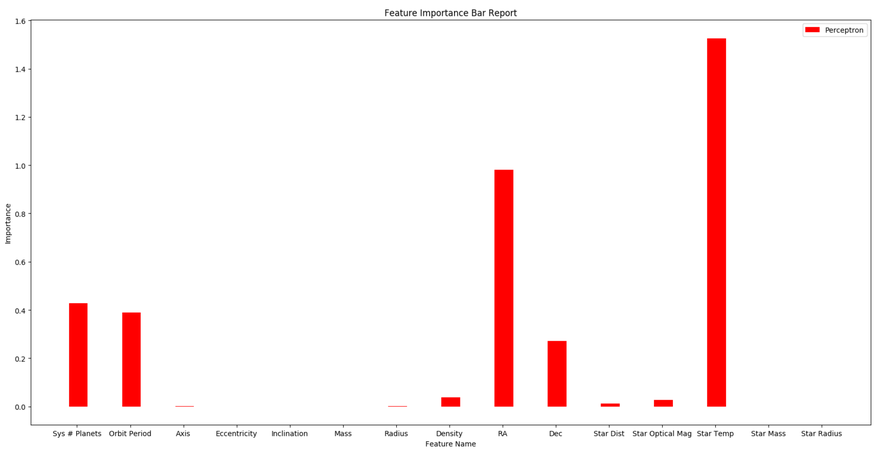
Implementation of K-Means was another approach that was considered for classification of planets. The issue was that if we had small clusters; it led to a vast number of false positives and a very large number of clusters was simply useless. Implementation of K Means is fast and easy.

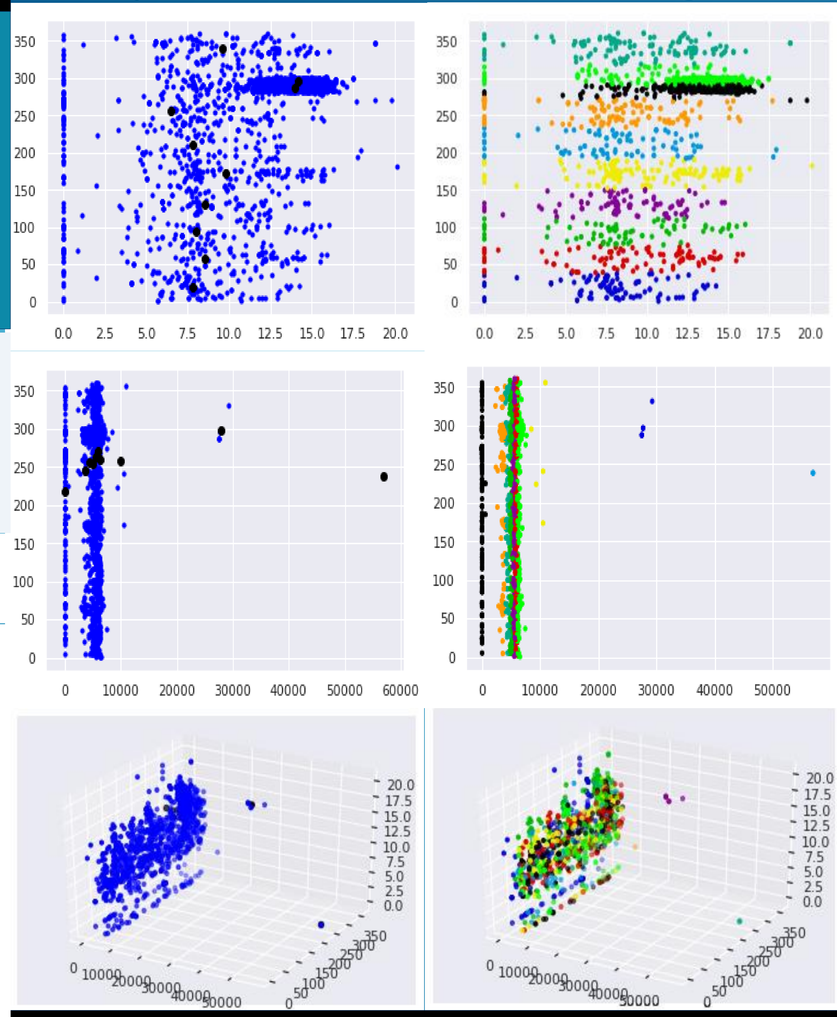


$$k \approx \sqrt{n/2}$$

N = 3372
k = 41







Solution Approach

Before the Kepler data could be used for machine learning, it had to be stripped of any unique features and useless values. Then we normalized all our values because a perceptron performs optimally with normalized data. We also had to label all this data because the techniques we wanted to use both are supervised learning algorithms. We accomplished this by finding a list of planets from the Kepler data set on the NASA website that are known to be in the habitable zone, and labeling all those with a binary label of yes and the rest with a binary label of no. We also used resampling to improve the size of our dataset because we only had about 20 example of planets labels with a yes. Even though Earth was not one of the planets the Kepler probe analyzed we still calculated the respective values and added it to the dataset as a good testing example.

Future Plans

In the future we plan on further implementation of K Means Clustering. We have two solutions in mind. Since one of our problems is a large amount of false positives along with true positives we could run K-Means recursively on the focused cluster, though we suspect that our results would look similar of that when we set our k value very high which would result in useless data. Our other solution was to run K Means many times on many different feature combinations. We then would implement a scoring system for every planet. Planets that end up in the target cluster in a certain graph get a plus to their score and those that aren't don't have their score changed. Although we would now have to consider planets that are similar in everyway except so dissimilar in one specific graph that it would no longer be considered habitable at that point. Thus we would need to check for that in some way and automatically disqualify said planet from appearing as a potential candidate for being "In the habitable zone"

Results

The results we received were unfortunate because we believe both are models to be overfitting on this problem because they both express basically perfect accuracy. We thought that if we boosted our dataset significantly we would have enough data to avoid overfitting but this doesnt seem to be the case. Even though overfitting may be occurring we still wanted to know what it was our models were learning. We created two bar graph reports that show how much each feature value is considered when making the classification and what we saw was interesting. We expected both bar graphs to look basically identical but that was not the case, the perceptron favored slightly different features than the random forest. This result although unexpected does make sense and supports our end goal, the perceptron is considering less things overall because its learning a higher bias model and the random forest tries to consider many features giving it a low bias. We found both models to be highly influenced by features that would influence how hot the surface of the planet would be which we found as compelling evidence that our models were learning to correctly classify a planet as either in the "habitable zone" or "not in the habitable zone".

Acknowledgements to Data World, NASA and Kepler Science for providing planet data for this project