Introduction

This report will example NASA Kepler exoplanets dataset using machine learning classification. There will be 3 different classification models used: Linear Regression, Decision Tree Classifier and Random Forest Classifier. In order to score these different classifications these processes will be done: preprocess the data, tune the classifiers using Grid Search and scoring the different classifiers.

Assumptions

1. Removed features relating to: names, dates/times, parameters, reference links and comments.
2. Finding the best feature set in the linear regression classifier will be the best features to use for the other classifiers.

Feature Removal

|  |  |
| --- | --- |
| Removed Columns | Score |
| None | 0.5791579615972337 |
| ['koi\_time0bk', 'koi\_time0bk\_err1', 'koi\_time0bk\_err2' ,'ra', 'dec','koi\_kepmag'] | 0.5675812577798078 |
| ['ra', 'dec','koi\_kepmag'] | 0.5801422098587785 |

Linear Regression

The worst performing of the 3 classifiers and is not good enough to predict new exoplanets. This classifier does have speed advantage and was used to fine tune the features and setup the code structure.

Decision Tree Classifier

This classifier performed significantly better than the Linear Regression classifier with a small hit to performance. This classifier is not good enough to be used to predict new exoplanets. Using Grid Search did help in tuning the model, using different parameters could lead to a better score.

Random Forest Classifier

This classifier was the best performing classifier of the 3 with a slight improvement over decision tree classifier. However, there was a big hit to performance with grid search and only yielded a very small improvement. The small improvement is not worth the slow performance. Different parameters could be tested to see if tuning would improve the score. This classifier is not good enough to be used to predict new exoplanets.

Summary:

The Random Forest Classifier is the best classifier analyzed. Feature selection did not significantly improve the scores of the 3 classifiers and removing too many features did lower the scores. This suggests that the dataset should not be changed too much. Grid search did not significantly improve score but more investigation is needed to find better parameters to tune. All 3 classifiers are not good enough to be used to predict new exoplanets. Investigation on different classifiers are needed.

Summary Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name | Model Type | Score | Parameters | Score with Grid Search |
| Model1 | Linear Regression | 0.5791579615972337 | N/A | N/A |
| Model2 | Decision Tree Classifier | 0.8558352402745996 | Max\_depth, min\_samples\_split | 0.8748807934388709 |
| Model3 | Random Forest Classifier | 0.8913043478260869 | N\_esimators, max\_depth | 0.891474346748045 |