Predicting Exoplanet Disposition Using NASA's Kepler Space Observatory Data and Machine Learning

Capstone Project 1: Proposal

Springboard Data Science Career Track
January 6th, 2019

Logan Rudd

1 Background and Justification

Scientists and researchers in organizations across the globe are currently searching for exoplanets (planets that orbit a star outside our solar system) in the habitable zone of their stars where liquid water might exist on the surface of the planet in order to answer one of the most timeless questions, 'Are we alone?' NASA's Keppler space observatory in particular was developed to survey the region of the Milky Way galaxy closest to us to reveal hundreds of Earth-size and smaller planets in or near the habitable zone so that we can determine the fraction of star systems in our galaxy that might contain such planets.

Throughout it's mission duration, which lasted from May 12, 2009 to November 15 2018, the Kepler space observatory has successfully located 9,564 KOIs (Kepler Objects of Interest) for which the data is publicly available online from the NASA Exoplanet Archive.[1] Of the 9,564 KOIs, 2,298 are confirmed exoplanets, 4,841 are false positives, and 2,425 are classified as candidates. For each entry in the KOI dataset there are 140 columns/features that organized into 8 categories ranging from identification columns to pixel based vetting statistics.

In order to aid in the discovery of such exoplanets, machine learning classification algorithms have already been developed to quickly and accurately determine which KOIs are exoplanets and which are false positives based on light curve data from the Kepler space telescope.[2] This project however will specifically focus on the binary classification of KOIs as either confirmed exoplanets or false positives using the following categories from the Cumulative KOI Data table publicly available on the NASA Exoplanet Archive: Transit Properties, Threshold-Crossing Event (TCE) Information, Stellar Parameters, KIC Parameters, and Pixel-Based KOI Vetting Statistics.

2 Project Plan

The project will begin by downloading and importing the data into a Jupyter notebook, and then cleaning and preparing the data by making any necessary transformations such as engineering new features. The data will then be thoroughly explored to reveal correlations and trends between features of the data and disposition/classification. Categorical features will be binarized if necessary, missing values will be imputed, and features will be selected and normalized accordingly. Lastly, the data will be modeled using the following supervised learning algorithms: Support Vector Machine, Logistic Regression, Random Forest, and XGBoost.

Since this is an unbalanced binary classification problem and false negatives are equally as bad as false positives, the models will then be tuned and evaluated using MAP (Mean Average Precision), also known as the area under the Precision-Recall curve. The parameters and methods will then be revised and tuned as needed, and the performance of the algorithms will be summarized in a table.

This project will conclude with a Jupyter notebook in github, milestone report, final report, and slide deck.

References

[1] NASA Exoplanet Archive.
Cumulative KOI Data, 2018
`https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=cumulative`

[2] Shallue, Christopher J.; Vanderburg, Andrew *IDENTIFYING EXOPLANETS WITH DEEP LEARNING: A FIVE PLANET RESONANT CHAIN AROUND KEPLER-80 AND AN EIGHTH PLANET AROUND KEPLER-90* Harvard Center for Astrophysics. (Dec. 2017) `https://www.cfa.harvard.edu/~avanderb/kepler90i.pdf`