Final Project Data Memo

TJ Sipin and Preeti Kulkarni

1/23/2022

Overview of dataset

The dataset is a cumulative record of all observed Kepler “objects of interest”. The observations are made by NASA’s Kepler Space Observatory, dedicated to finding possible habitable planets outside of Earth.

The dataset is available on Kaggle, but was published as-is by NASA (https://www.kaggle.com/nasa/kepler-exoplanet-search-results?select=cumulative.csv (https://www.kaggle.com/nasa/kepler-exoplanet-search-results?select=cumulative.csv)).

There are about 10,000 observations (potential exoplanet candidates), with 49 total predictors, but with 5 notable predictors according to Kaggle: - kepoi_name - kepler_name - koi_disposition - koi_pdisposition - koi_score

Within the 50 columns, there are 36 decimal variables, 5 integer variables, 5 string variables, and 4 “other” variables according to Kaggle.

After running nrow(na.omit(df)), we found that we have 7803 observations with complete data. That means there are 1761 observations that we’d probably need to remove. Due to the large amount of observations, we plan on just removing the rows with missing values.

An overview of the research questions

We want to predict koi_score, but only for observations classfied as CANDIDATEs under the KOI disposition variable. Given a certain set of predictors, what KOI score might that observation have? A KOI score is a value between 0 and 1 indicating the confidence in the KOI disposition. A KOI disposition has 4 possible values: CANDIDATE, FALSE POSITIVE, NOT DISPOSITIONED, or CONFIRMED. For CANDIDATEs, a higher value correlates to more confidence in its disposition. For FALSE POSITIVEs, a higher value indicates less confidence in its disposition. Thus, we wish to only test on a subset of the original data set, that is, only observations that are CANDIDATEs.

Based on the high amount of quantitative variables, we plan on using regression for predicting the KOI score. Some variables of interest to act as predictors might be those under the Transit Properties, such as koi_depth, koi_ror, koi_prad, koi_teq/koi_insol, koi_dor. Other variables of interest might be under the Stellar Parameters catergory: koi_steff, koi_slogg, koi_smet, koi_smass, koi_sage, koi_sparprov.

Proposed project timeline and group work

We plan on working together for the most part. That is, performing tasks at the same time, as in coding together and comparing and reviewing each other's code. However, when tasks don't require comparing code, we plan on asking each other to complete tasks that are preliminary to the main task. We can begin doing exploratory data analysis sometime next week.

Questions/Concerns

We might need help with subsetting our data set so it works with what we're looking to solve.