KEPLER EXOPLANETS CLASSIFIER

By: Nelson Genao

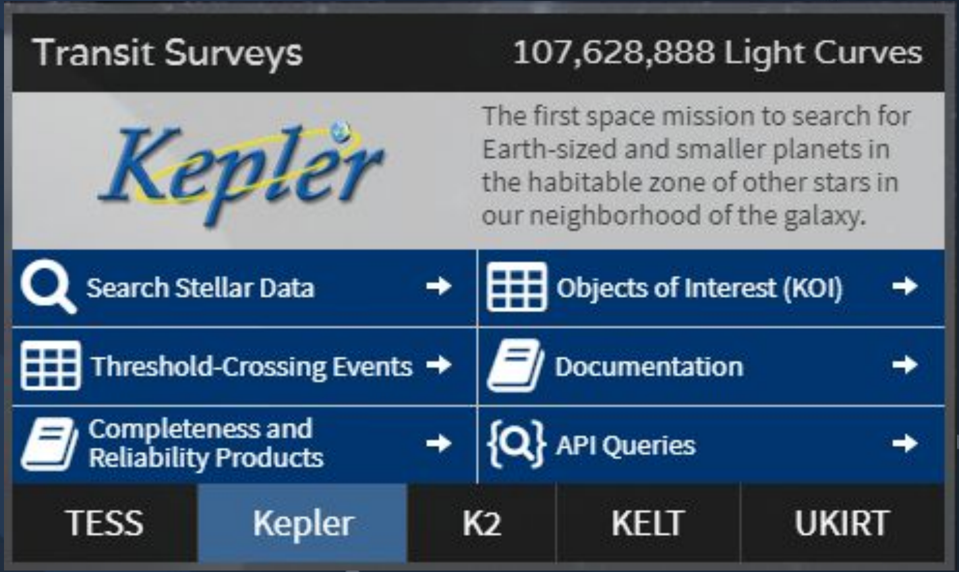
BACKGROUND AND OVERVIEW

- The Kepler mission was designed to locate Earth sized planets by analyzing an object's transit data as it orbits a star
- These objects are classified as confirmed, false positive or candidate exoplanets.
- Candidate planets require additional research in order to classify. The goal of the model is to predict which candidates would likely be confirmed exoplanets so resources and focus can be directed towards them.

SOURCES AND DATA

- NASA Exoplanet Archive
  - Kepler Objects of Interest Database
  - User friendly
  - Lots of documentation

- Final Model Includes:
  - 33 features
  - 2900 False Positive Objects
  - 2200 Confirmed Planets
  - Random Forest Classifier



Transit Surveys 107,628,888 Light Curves

Kepler

The first space mission to search for Earth-sized and smaller planets in the habitable zone of other stars in our neighborhood of the galaxy.

Search Stellar Data →

Objects of Interest (KOI) →

Threshold-Crossing Events →

Documentation →

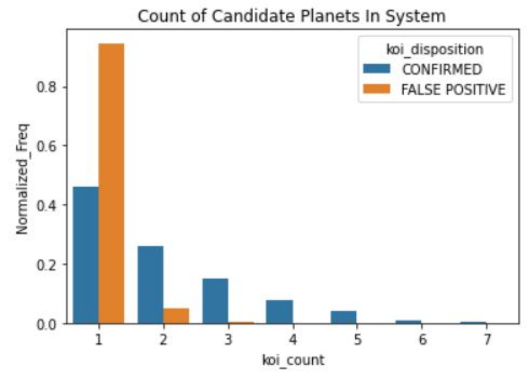
Completeness and Reliability Products →

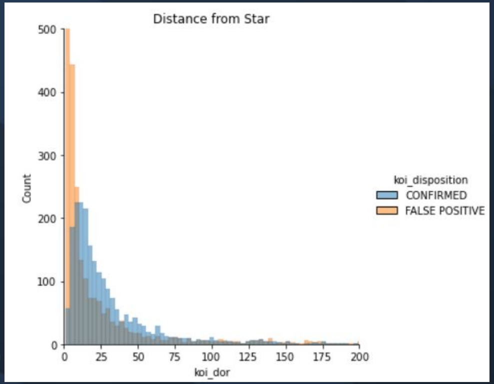
API Queries →

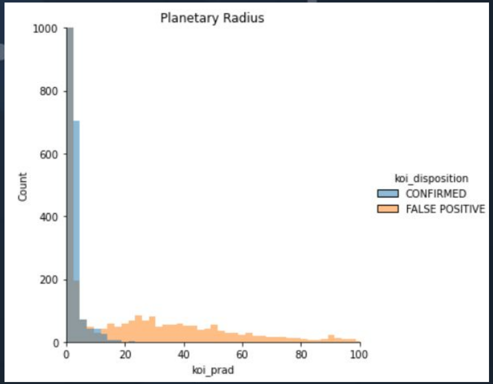
TESS | Kepler | K2 | KELT | UKIRT

EXPLORATORY DATA ANALYSIS

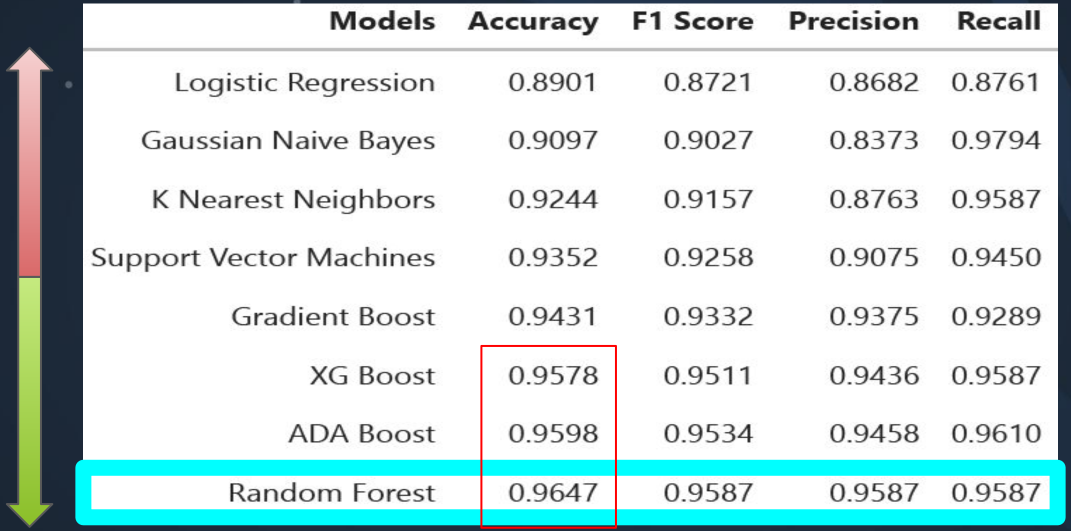
- Dropped any rows with missing data
- Several features contained many outliers
  - Used Robust Scaling when modeling to counteract
- Removed multicollinear features to help speed up modeling time
- Removed confounding features used in calculating disposition
- Top Features Include:
  - Planetary Radius
  - Distance from Star
  - Count of Planets in System


Count of Candidate Planets In System


Distance from Star


Planetary Radius

TUNED MODEL - TEST RESULTS



Models	Accuracy	F1 Score	Precision	Recall
Logistic Regression	0.8901	0.8721	0.8682	0.8761
Gaussian Naive Bayes	0.9097	0.9027	0.8373	0.9794
K Nearest Neighbors	0.9244	0.9157	0.8763	0.9587
Support Vector Machines	0.9352	0.9258	0.9075	0.9450
Gradient Boost	0.9431	0.9332	0.9375	0.9289
XG Boost	0.9578	0.9511	0.9436	0.9587
ADA Boost	0.9598	0.9534	0.9458	0.9610
Random Forest	0.9647	0.9587	0.9587	0.9587

- Top 3 models were very close in performance
- Could have improved scores if I allowed more time for tuning and training

MODELING - BALANCED FOR ACCURACY

BASE MODEL

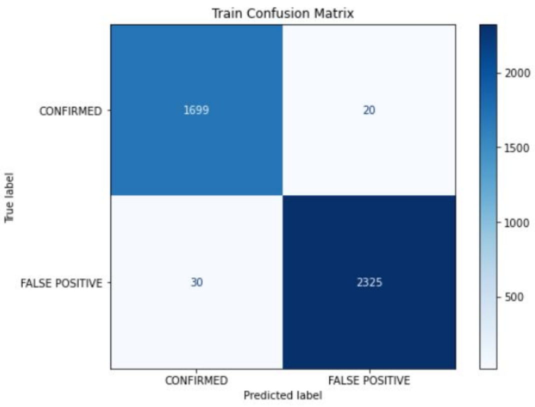
- Logistic Regression
- Train Results:
  - ○ Accuracy = 91%
  - ○ Precision = 91%
  - ○ Recall = 91%
- Test Results:
  - ○ Accuracy = 89%
  - ○ Precision = 89%
  - ○ Recall = 89%

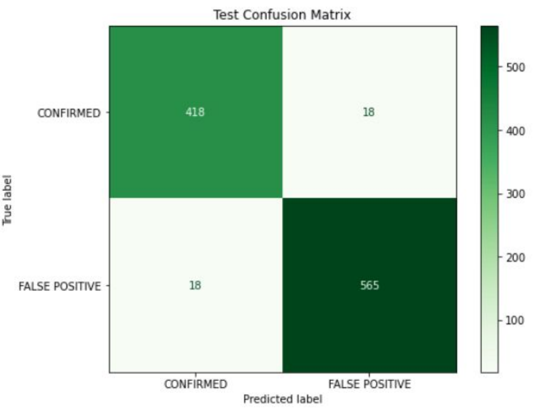
FINAL MODEL

- Random Forest Classifier
- Train Results:
  - ○ Accuracy = 99%
  - ○ Precision = 99%
  - ○ Recall = 99%
- Test Results:
  - ○ Accuracy = 96%
  - ○ Precision = 96%
  - ○ Recall = 96%

** Accuracy chosen to balance costs and benefits of True/False Positives

- Maximizing True Positives can help locate as many viable Confirmed planets as possible
- However, time and data collection can be very limited when observing celestial objects





CANDIDATE PREDICTIONS

~1600 CANDIDATE OBJECTS

57% CONFIRMED

~ 912 POTENTIAL CONFIRMED PLANETS

43% FALSE POSITIVES

NEXT STEPS

IMPROVE DOMAIN KNOWLEDGE

Consult with subject matter experts to improve models

HABITABLE PLANETS

Help identify potentially habitable planets

OTHER MISSIONS

Apply models to other exoplanet search missions

ADAPT MODELS

Utilizing different exoplanet searching techniques

THANK YOU

https://github.com/NelGen/NG-NASA-Exoplanet-Classifier-Project

https://exoplanets.nasa.gov/