Classifying Kepler Objects of Interest

By: Ananya Vasudev, Andrew Furlong, James Lederman, Lajja Pancholy

Project Supervisor: Peter Freeman

Background & Introduction

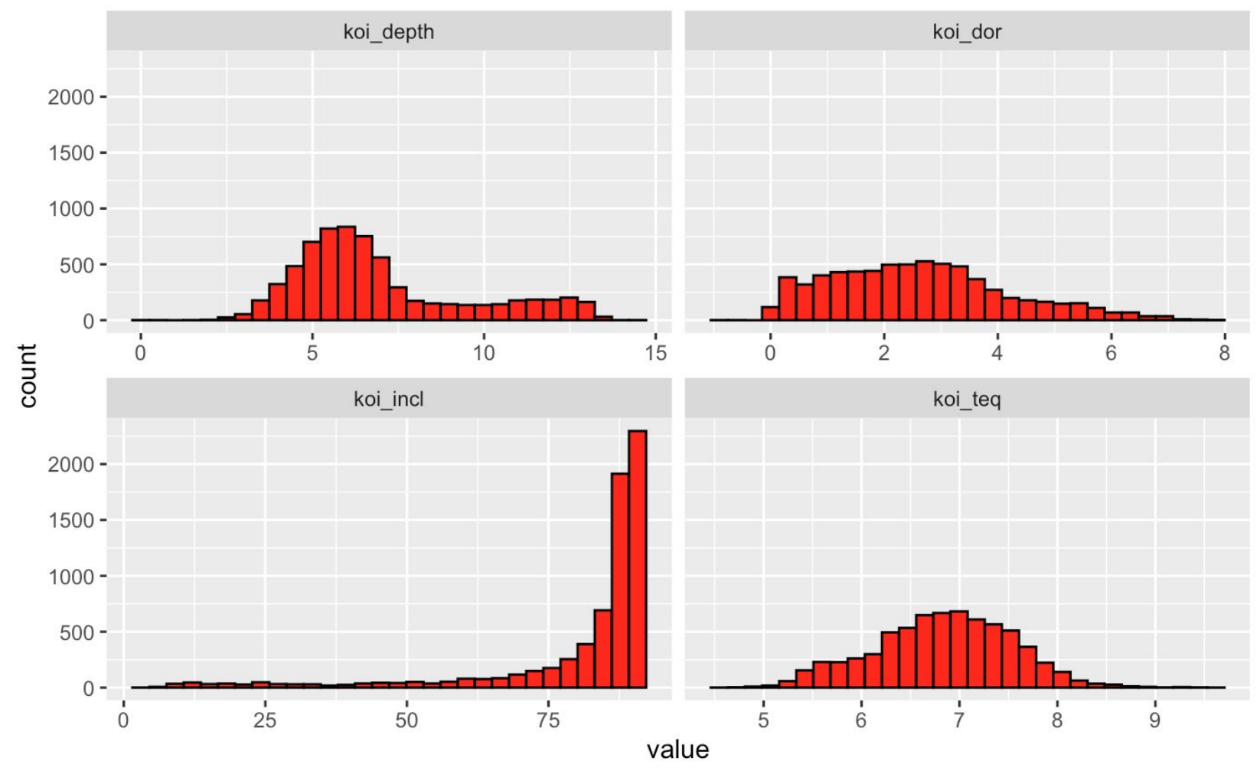
Exoplanets are planets that exist outside of the bounds of our solar system. Due to their small size, exoplanets are almost impossible to identify directly. Between 2009 and 2013, NASA's Kepler satellite kept tabs on hundreds of thousands of stars. From there, "objects of interest," or stars with possible exoplanets, were identified. **The goal of this project is to learn a classification model which can accurately differentiate between stars which do have exoplanets ("confirmed"), and stars which do not have exoplanets ("false positive").**

category	variables
exoplanet orbit-related	period, eccen, sma, incl, dor
transit/eclipse-related	impact, duration, depth
exoplanet property-related	ror, prad, teq, insol
host-star property-related	srho, steff, slogg, smet, srad, smass

category	data
Confirmed	2296
Candidate	2318
False Positive	4563

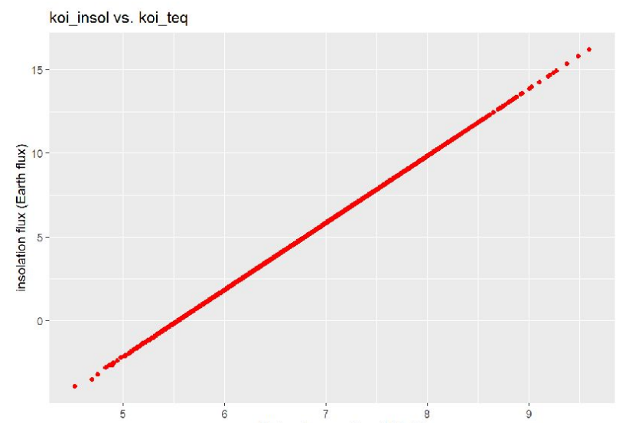
Data Pre-Processing

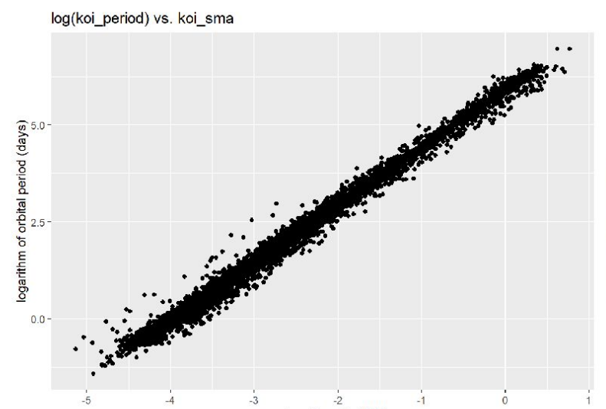
- **EDA**
  - *Univariate Analysis*: After removing `koi_eccen` (which contained only zeroes), we conducted log transformations on the significantly right-skewed predictors. For our visualization, we focused on the distributions of one variable from each of the four categories.

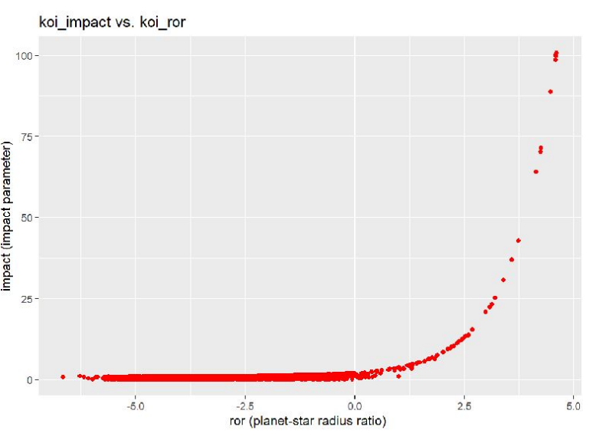


Following our log transformations for `koi_depth`, `koi_dor`, and `koi_teq`, we see that they have somewhat symmetric and unimodal distributions. `Koi_incl` appears to have a significantly left-skewed distribution.

  - *Bivariate Analysis*: We observed deterministic relationships in the `koi_period` vs. `koi_sma`, `koi_ror` vs. `koi_impact`, and `koi_teq` vs. `koi_insol` relationships, and hence decided to remove `koi_period`, `koi_ror`, and `koi_insol` to avoid multicollinearity.







- **Variable Selection**
  - *Principal Component Analysis*: Given that we had a large number of variables in our dataset, we conducted PCA to obtain an idea of whether we could reduce the dimensionality of our predictor space. We found that we would ideally retain only our first 7 principal components.
  - *Best Subset Selection*: Using best subset selection, we found that the optimal model made use of 12 predictors out of the 14 predictors we were working with, with `koi_impact` and `koi_slogg` being omitted.

Methods

- We attempted to fit eight potential models to our data, namely logistic regression, best subset selection, a classification tree, linear discriminant analysis, Naive Bayes, SVM, XGBoost, and random forest.
- Random forest - a machine learning method used for classification and regression that involves constructing and aggregating multiple decision trees - yielded our lowest misclassification rate (~7.5%).
- Thus, we decided to use our random forest model to generate final predictions.

Analysis and Results

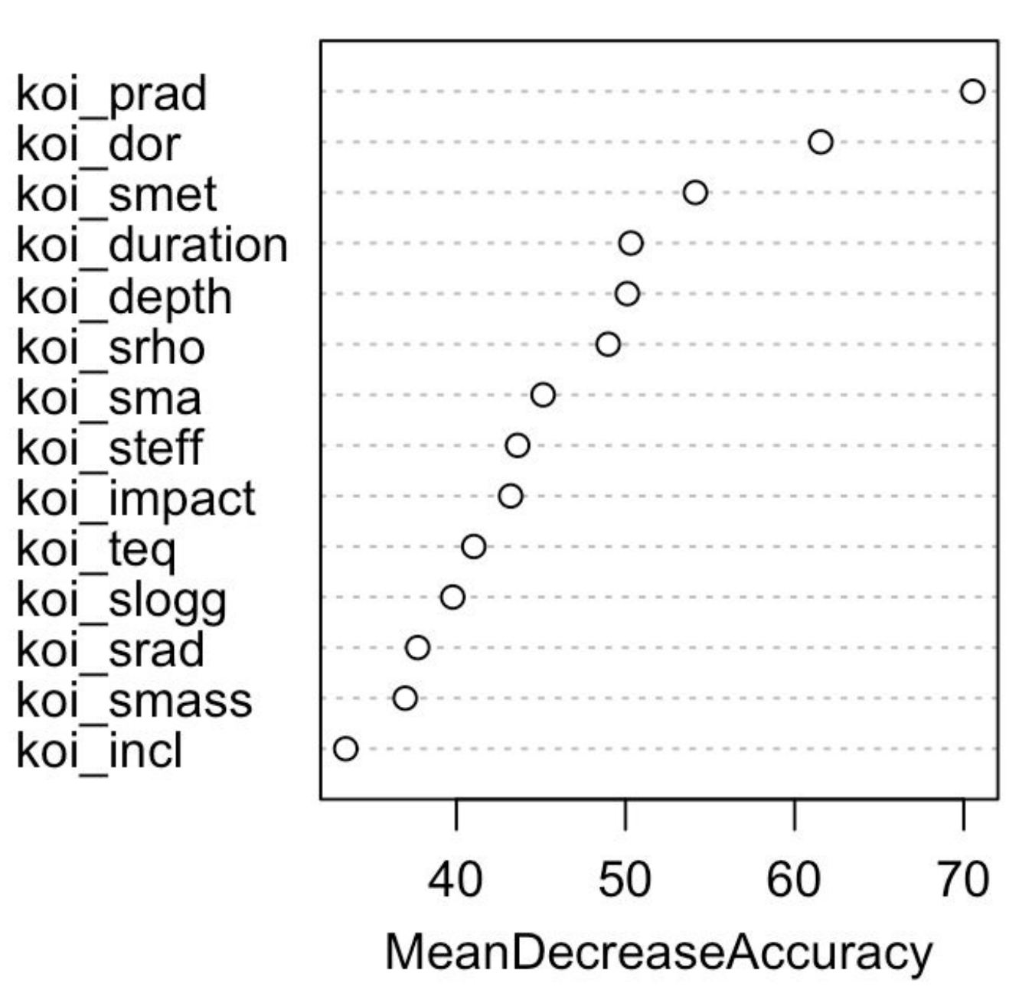
	CONFIRMED	FALSE POSITIVE
'PREDICTED' CONFIRMED	433	85
'PREDICTED' FALSE POSITIVE	18	836

- The table above shows the performance of our random forest model.
- The table to the right shows the performance of all models built.
- Below is the variable importance plot for our random forest model, which identifies the variables that were most important for classification of Kepler objects of interest.

Performance across Classifiers

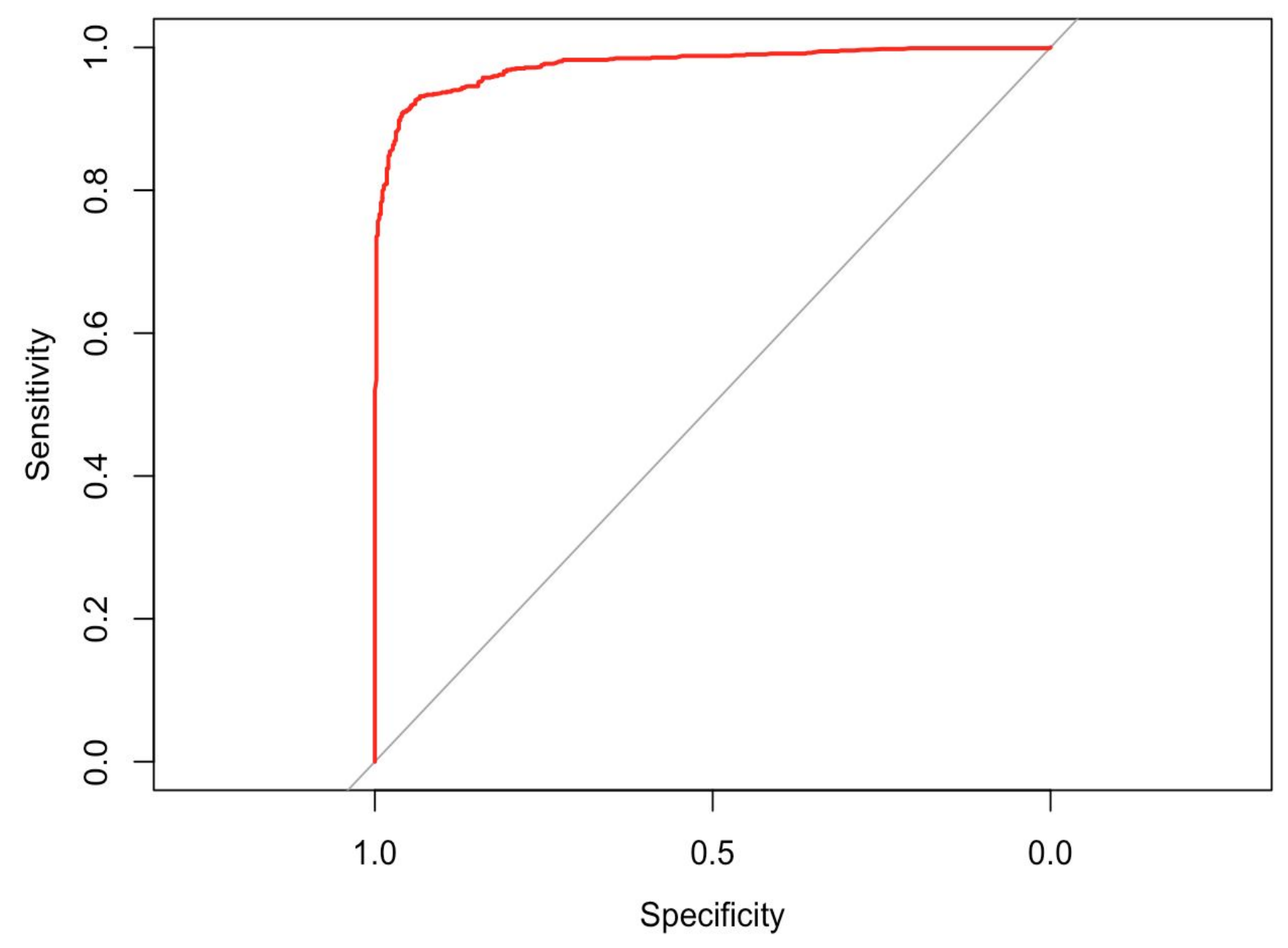
Name of Model	MCR (in percent)	AUC (in percent)	Sensitivity (in percent)	Specificity (in percent)
Logistic Regression	14.2	91.3	84.5	88.5
Logistic Regression with PCA Data	22.8	84.4	77.9	74.9
Classification Tree	13.3	92.9	83.2	93.8
LDA Analysis	23.8	86.2	69.1	90.9
Random Forest Analysis	7.5	97.6	90.8	96
Naive Bayes	13.5	92.4	86.6	86.3
SVM	8.7	96.6	94.9	89.6
XGBoost	8.5	96.4	90.2	94.2
Best Subset Selection	14.1	91.3	88.5	84.6

Variable Importance Plot for Random Forest Model



- Using the ROC curve on the right, we derived a Youden's J statistic value (i.e sensitivity plus specificity minus one) of 0.868 for our random forest model. This is fairly high and indicates that our model has good performance.

ROC Curve for Random Forest Model



Conclusions

- When building a binary classifier, the variables `koi_prad, koi_dor, and koi_smet` were most significant.
  - This is confirmed by the variable importance plot for our random forest model.
- We determined that the random forest model was superior for accurate classification of data.
- When run on current exoplanetary candidates, our model yielded the results on the right.

Confirmed	False Positive
765	1553

Reference: https://exoplanetarchive.ipac.caltech.edu