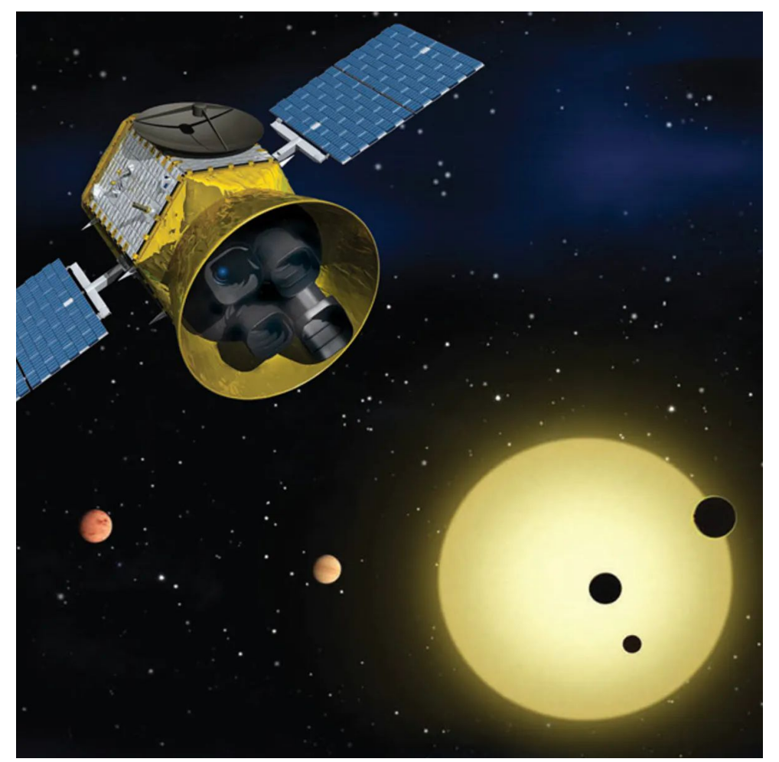
Is it a Planet?: Classifying TESS Observations

By: Kay Nam, Hannah Shane, Steven Tang, Maya Farhadi

Advisor: Peter Freeman



Introduction

Exoplanets, planets that exist outside of our Solar System, are almost impossible to discover through direct imaging. One method to detect them is the transit method, which observes potential exoplanets passing in front of their host stars. The Transiting Exoplanet Survey Satellite (TESS) is currently observing many stellar systems, looking for transiting planets, and if there is statistical evidence of one, the object becomes an "object of interest." All TESS objects of interest, or TOI, are classified as being confirmed planets (CP), as having been shown to not actually be a planet (FP), or as awaiting a final decision (PC). **We intend to learn a classifier to distinguish between confirmed planets and false positives, and use the model to determine which of the planetary candidates can be categorized as confirmed planets (CP) or false positives (FP), which can help inform astronomers which candidates are worthy of inspection first.[1]**

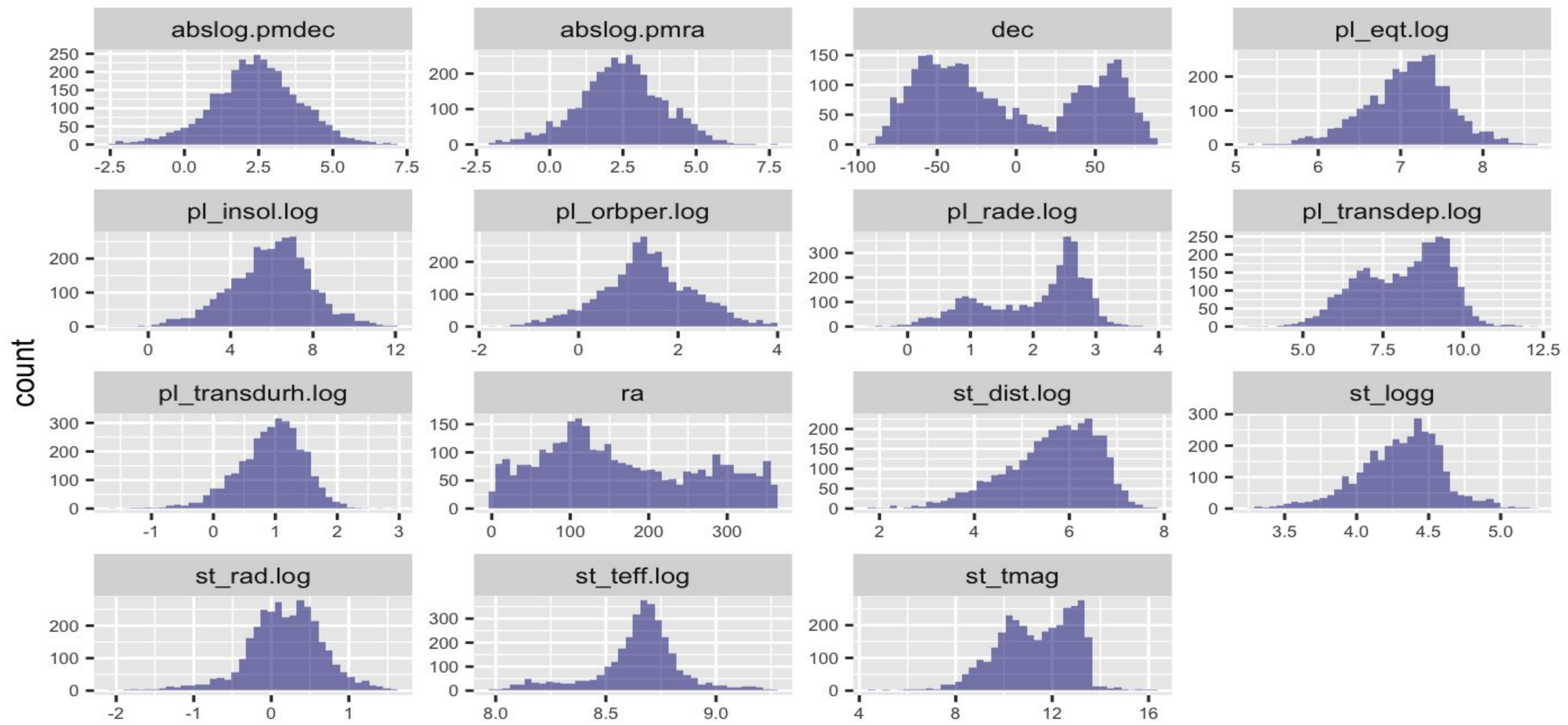
Data

The TESS satellite collects different measurements from potential exoplanets and its host star. In addition to the predictors below, the potential exoplanets are labeled Planetary Candidate, False Positive or Confirmed Planet. There are 423 confirmed planets, 490 false positives, and 2480 planetary candidates that have yet to be classified as any of the previous two categories. In addition to performing the necessary transformations, we remove 141 observations that are considered as outliers. The resulting data are visualized as histograms below.

Variables(s)	Description
(ra,dec)	celestial longitude and latitude
(st_pmra,st_pmdec)	how "fast" the host star moves in the ra and dec directions (mas/yr)
pl_orbper	the planetary orbital period (days)
(pl_tranduh,pl_trandep)	the duration and "light-blocking amount" of the transit (hours and ppm)
pl_rade	the radius of the planet in Earth radii
pl_insol	the amount of light the planet receives, relative to what the Earth receives
pl_eqt	the planet's temperation (K)
st_tmag	the host star magnitude in the "TESS band"
st_dist	the distance to the host star, in parsecs
st_teff	the temperature of the host star (K)
st_logg	the host star's surface gravity (cm/s^2)
st_rad	the host star's radius, in solar radii

Table 1: Predictor Variables[1,2]

Figure 1: Histogram of Variables



Methods

- We remove planetary candidates and train our model on 70% of the data and test on 30% of the data that remain.
- We use a number of classification techniques to build classifiers: logistic regression, log-forward subset selection based on Akaike Information Criteria (AIC), decision tree, random forest, gradient boosting, K-nearest neighbors, Naive Bayes, support vector machines with linear, polynomial, and radial kernels.
- We generate final predictions using our random forest model because it has the greatest AUC value (0.941). Random forest models are an ensemble learning method in which many decision trees are aggregated together to form a unified model.

Analysis

Model	AUC
Random Forest	***0.941***
Support Vector Machine (Linear Kernel)	0.926
Gradient Boosting	0.926
Support Vector Machine (Radial Kernel)	0.903
Logistic Regression (Log-Forward Subset Selection based on AIC)	0.846
Logistic Regression	0.845
Support Vector Machine (Polynomial Kernel)	0.844
Decision Tree	0.812
Naive Bayes	0.802
K-Nearest Neighbors	0.749

Table 2: AUCs for the models attempted

	CP	FP
CP	101	14
FP	16	135

Table 3: Confusion matrix of the random forest model (Columns are predictions and rows are its true designation)

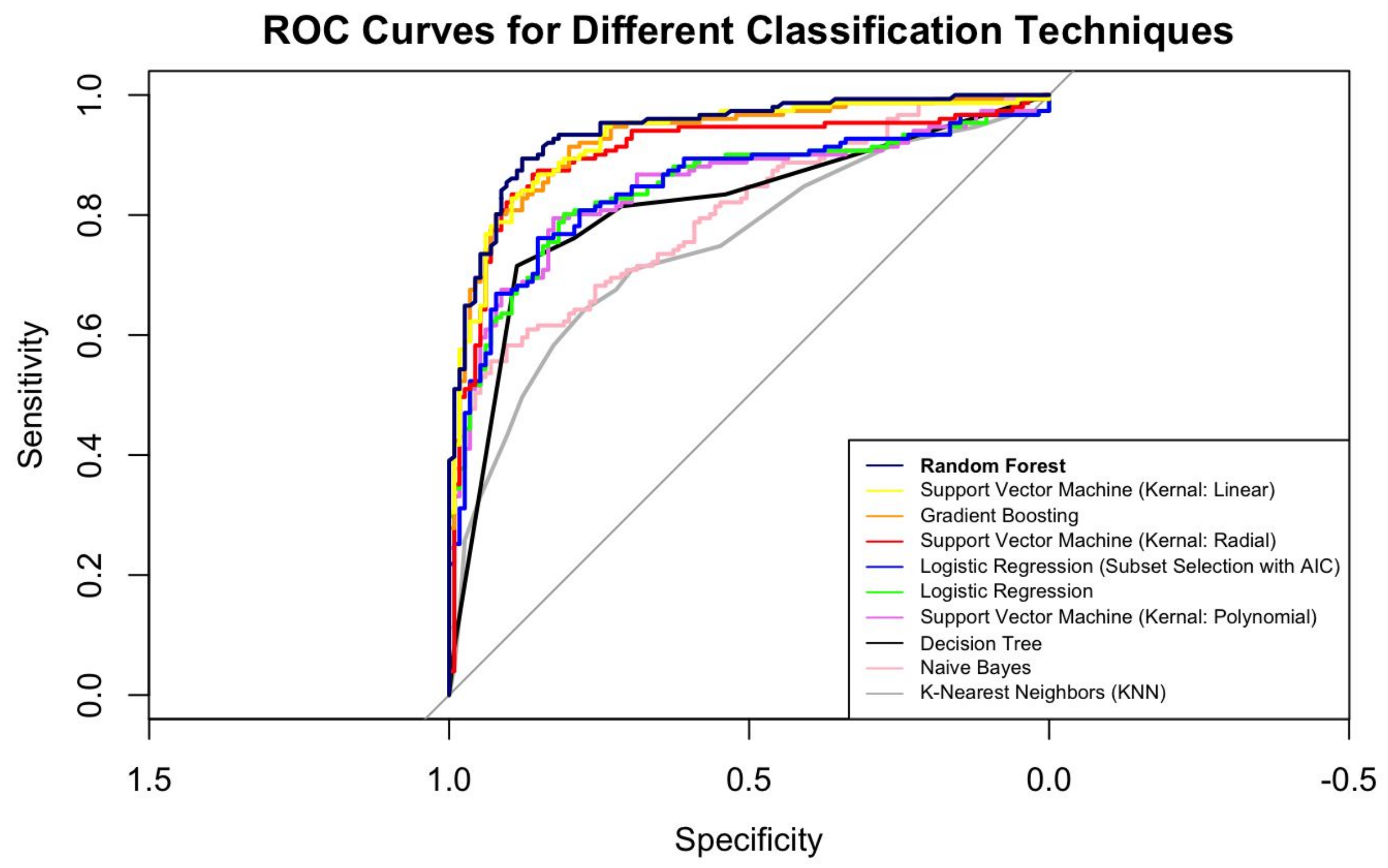


Figure 2: ROC Curve of all the models learned

- According to Table 2 and Figure 2, the random forest model performs the best compared to the other models. Its AUC is 0.941.
- The ROC curve for the random forest model is shown in Figure 2. From the results, we have chosen a optimum threshold based on Youden's statistic, which ensure the optimal predictive accuracy for both classes
- The specificity (accuracy in predicting false positives) is 0.894. The sensitivity (accuracy in predicting confirmed planets) of the random forest model is 0.878. The final misclassification rate is 0.113. See Table 3.

- We also display the variable importance plot. Mean decrease in accuracy is used to measure how much accuracy is lost if the variable is excluded, thereby measuring the importance of the variable.
- According to Figure 3, it appears that the "light-blocking amount" of the transit (in ppm) appears to be the most important variable in classifying TESS observations.

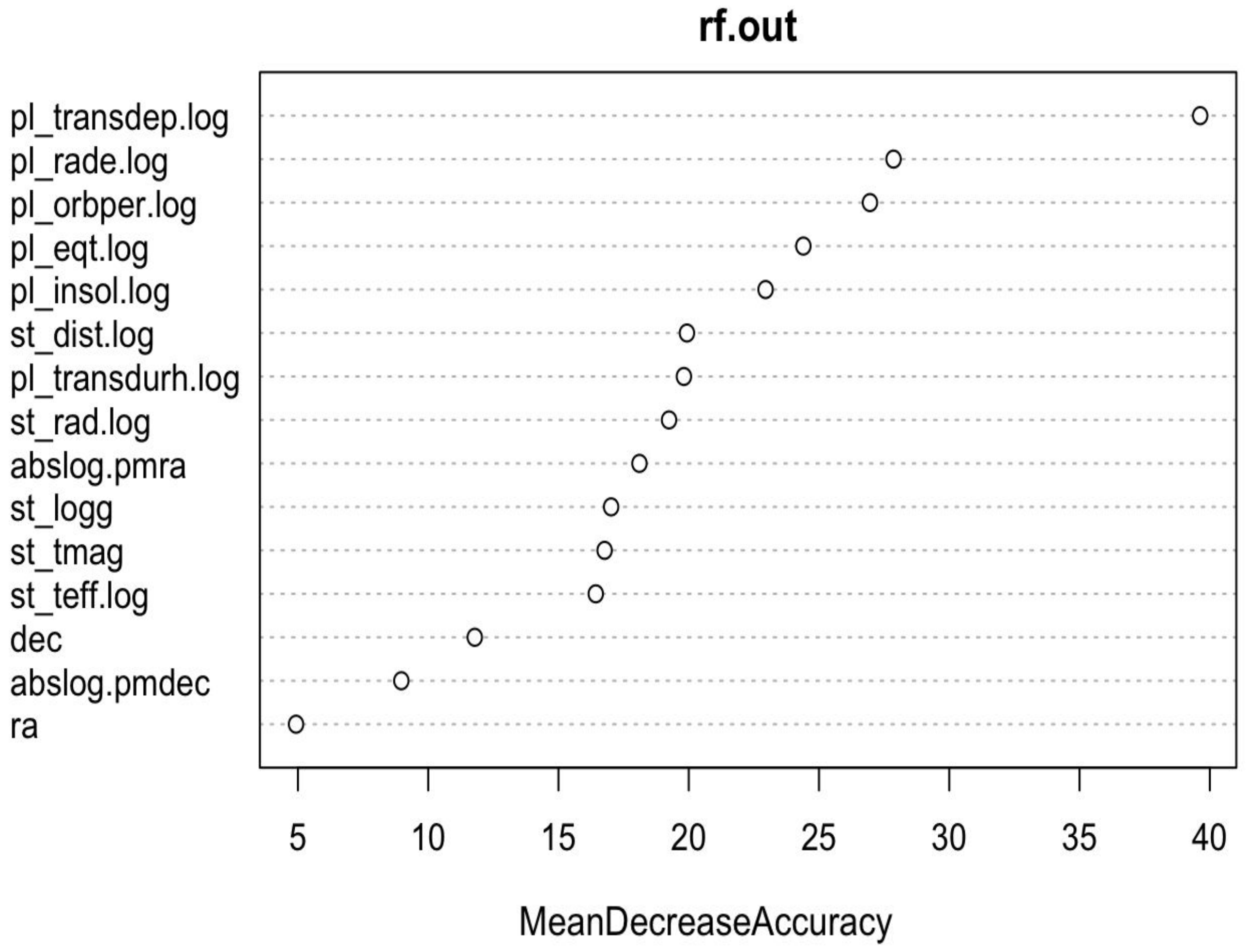


Figure 3: Variable Importance Plot of the Random Forest Model

Conclusions

We conclude that distinguishing between objects of interests that are false positives and those that are confirmed planets is possible. The best performing model for the classification task is the random forest model, having a classification rate of 0.113. So as to determine how many of the 2480 planets that have not been classified in our dataset as of yet, we ran these observations against our model and determined that 1392 of the Planetary Candidates will be eventually confirmed as confirmed Planets, whereas the other 975 objects of interest are false positives.

References

[1] Freeman, P. E. 2021, online at https://github.com/pefreeman/36-290/blob/master/PROJECT DATASETS/TOI/README.md

[2] NASA Exoplanet Archive, online at https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=TOI