Is It a Planet?: Learning a Classifier for TESS Observations to Distinguish Between Confirmed Planets and False Positives

Steven Tang

December 3, 2021

Contents

Introduction

Exoplanets, also known as extrasolar planets, are planets which exist beyond our solar system. Instead of direct imaging, these planets are most commonly observed using two methods, either through measuring the radial velocity of a host star around the star system's center of mass, or through the transit method, which seeks to measure the degree to which a planet eclipses its host star (from our perspective). The *Transiting Exoplanet Survey Satellite*, launched in 2018, surveys the full sky and utilizes the transit method to search for exoplanets. It has revealed many objects of interest, some of which are confirmed to be planets or false positives, while others are still planetary candidates that have not been categorized. The objective of this report is therefore **to learn a classifier that is able to distinguish between false positives and confirmed planets, which will then be used to predict the status of the undesignated planetary candidates**.

Data Visualization and Exploratory Data Analysis (EDA)

General Information about the Data Set and Variables

The dataset analyzed in this report was downloaded from the NASA Exoplanet Archive on September 15, 2021. The data set contains 3393 and 16 variables, one categorical response variable and fifteen quantitative predictor variables. Below is a table displaying the various explanatory variables included in the data set:

Variable	Description	Units and Other Information
`ra`	celestial longitude	deg
`dec`	celestial latitude	deg
`st_pmra`	how "fast" host star moves in celestial longitude direction	milliarcseconds per year (mas/yr)
`st_pmdec`	how "fast" host star moves in celestial latitude direction	mas/year
`pl_orbper`	planetary orbital period	days
`pl_transdurh`	the duration of transit	hours
`pl_trandep`	the "light blocking amount" of transit	parts-per-million (ppm)
`pl_rade`	the radius of the planet	Earth radii
`pl_insol`	the light received by the planet	relative to what Earth receives
`pl_eqt`	the planet's temperature	K
`st_tmag`	the host star's brightness in TESS-band	units of magnitude
`st_dist`	the distance to host star	parsec
`st_teff`	the temperature of host star	K
`st_logg`	logarithm of the host star's surface gravity	cm/s^2
`st_rad`	the host star's radius	in solar radii

The response variable `label` describes the designation of the observation, and has the following breakdown:

Designation	Description
`PC`	Planetary Candidate
`CP`	Confirmed Planets
`FP`	False Positives

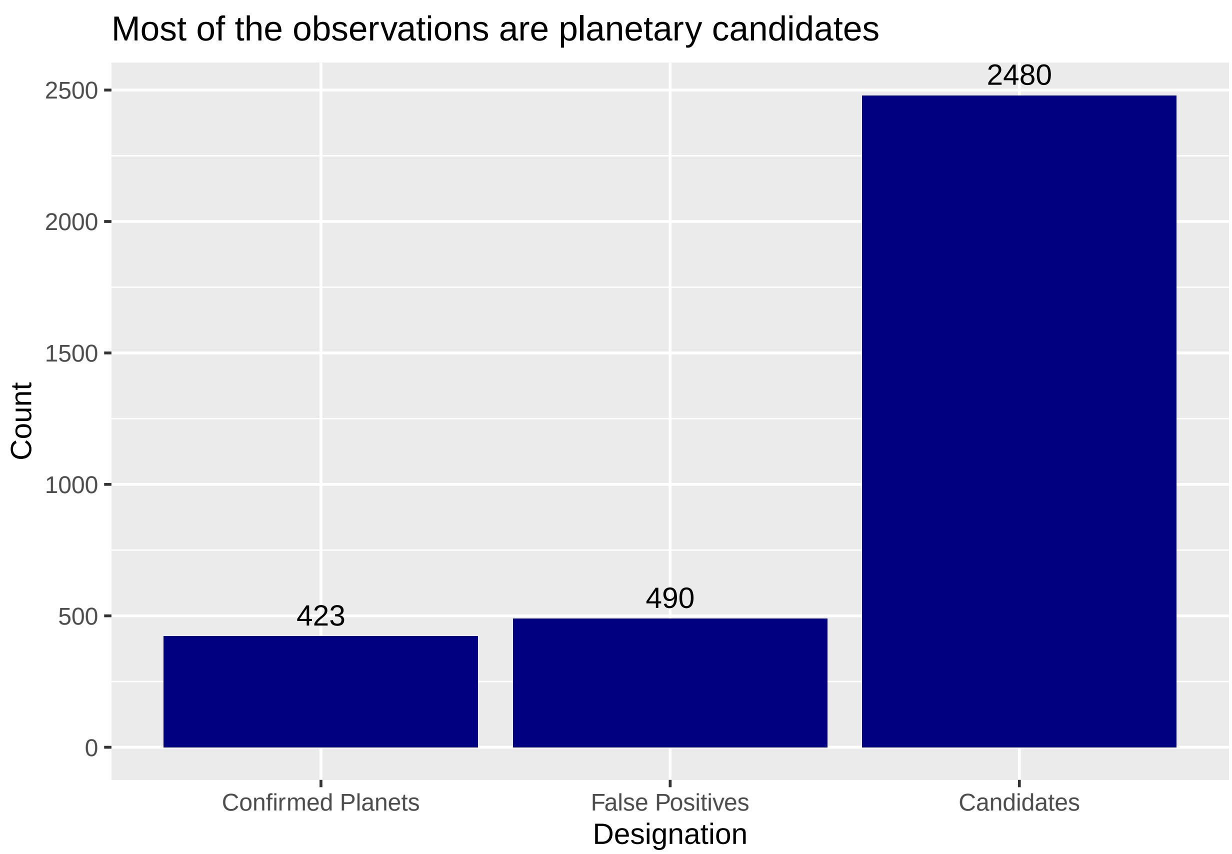
A sample of six rows of the data, with columns split in half and table formatted for readability, is shown below:

ra	dec	st_pmra	st_pmdec	pl_orbper	pl_trandurh	pl_trandep
112.3577	-12.695960	-5.964	-0.076	2.171348	2.017220	656.8861
122.5805	-5.513852	-4.956	-15.555	1.931671	3.180000	1030.0000
122.1782	-48.802811	-4.496	9.347	3.577575	2.934709	501.6029
112.7524	-4.463359	0.357	3.399	6.998921	3.953000	2840.0000
109.3828	13.395219	-17.900	1.300	2.049360	2.234000	249.5000
318.7370	-55.871863	12.641	-16.011	1.430343	1.637794	19976.6300

pl_rade	pl_insol	pl_eqt	st_tmag	st_dist	st_teff	st_logg	st_rad	label
5.818163	22601.949	3127.204	9.60400	485.735	10249.0	4.19000	2.169860	FP
10.316800	42432.800	3998.000	9.42344	295.862	7070.0	4.03000	2.010000	PC
5.050111	8092.969	2419.060	9.13550	356.437	9219.0	4.14000	2.150400	FP
14.775200	448.744	1282.000	8.87759	283.291	6596.0	3.71000	2.700000	PC
1.716070	2107.050	1887.000	8.93620	144.297	6858.5	4.20000	0.990000	FP
13.576368	1281.273	1525.914	12.40690	375.310	5600.0	4.48851	0.890774	CP

Basic Summary of the Categorical Variable `label`

Regarding the composition of the objects of interest, there are 423 confirmed planets, 490 false positives, and 2480 planetary candidates.



As our analysis is focused on the classification of planetary candidates, we will remove them for now and focus on learning a classifier to distinguish confirmed planets and false positives.

Below, we display the summary of the quantitative variables for the resulting data set:

```
##        ra               dec              st_pmra             st_pmdec      
##  Min.   :  0.1856   Min.   :-88.121   Min.   :-1053.980   Min.   :-990.311  
##  1st Qu.: 80.8961   1st Qu.:-45.898   1st Qu.:  -14.691   1st Qu.: -19.113  
##  Median :136.2389   Median :-20.246   Median :   -2.155   Median :  -3.640  
##  Mean   :166.5056   Mean   : -7.443   Mean   :    7.073   Mean   : -14.032  
##  3rd Qu.:277.4781   3rd Qu.: 38.242   3rd Qu.:   12.885   3rd Qu.:   7.316  
##  Max.   :359.9009   Max.   : 87.869   Max.   : 2074.520   Max.   :1048.840  
##    pl_orbper         pl_trandurh        pl_trandep          pl_rade        
##  Min.   :  0.2666   Min.   : 0.3458   Min.   :    90.1   Min.   :  0.8202  
##  1st Qu.:  1.6698   1st Qu.: 1.9090   1st Qu.:  1130.0   1st Qu.:  4.8218  
##  Median :  3.2429   Median : 2.6730   Median :  4657.9   Median : 11.0678  
##  Mean   :  5.4782   Mean   : 2.9290   Mean   :  8280.8   Mean   : 11.0892  
##  3rd Qu.:  5.5554   3rd Qu.: 3.5888   3rd Qu.: 11315.9   3rd Qu.: 14.8373  
##  Max.   :163.9874   Max.   :14.8448   Max.   :183631.0   Max.   :208.4890  
##     pl_insol            pl_eqt        st_tmag          st_dist        
##  Min.   :     0.54   Min.   : 239   Min.   : 4.628   Min.   :   6.531  
##  1st Qu.:   223.20   1st Qu.:1050   1st Qu.: 9.773   1st Qu.: 146.653  
##  Median :   902.20   Median :1445   Median :10.539   Median : 268.508  
##  Mean   :  5063.15   Mean   :1595   Mean   :10.756   Mean   : 350.658  
##  3rd Qu.:  2579.38   3rd Qu.:1909   3rd Qu.:11.795   3rd Qu.: 437.215  
##  Max.   :280833.00   Max.   :6413   Max.   :16.338   Max.   :7294.410  
##     st_teff         st_logg          st_rad      
##  Min.   : 2940   Min.   :2.360   Min.   : 0.150  
##  1st Qu.: 5375   1st Qu.:4.080   1st Qu.: 0.920  
##  Median : 5915   Median :4.290   Median : 1.300  
##  Mean   : 5986   Mean   :4.273   Mean   : 1.451  
##  3rd Qu.: 6406   3rd Qu.:4.480   3rd Qu.: 1.760  
##  Max.   :31000   Max.   :5.600   Max.   :12.230  
```

Notice that there are several variables with potential outliers, which according to the six number summaries of the variables as well as their respective histograms, include all variables besides `dec` and `ra`. Note that `dec` and `ra` are measurements that should not be too deeply looked into because they correspond to the location where the observation is taken, rather than it meaningful properties.

Transforming Predictor Variables

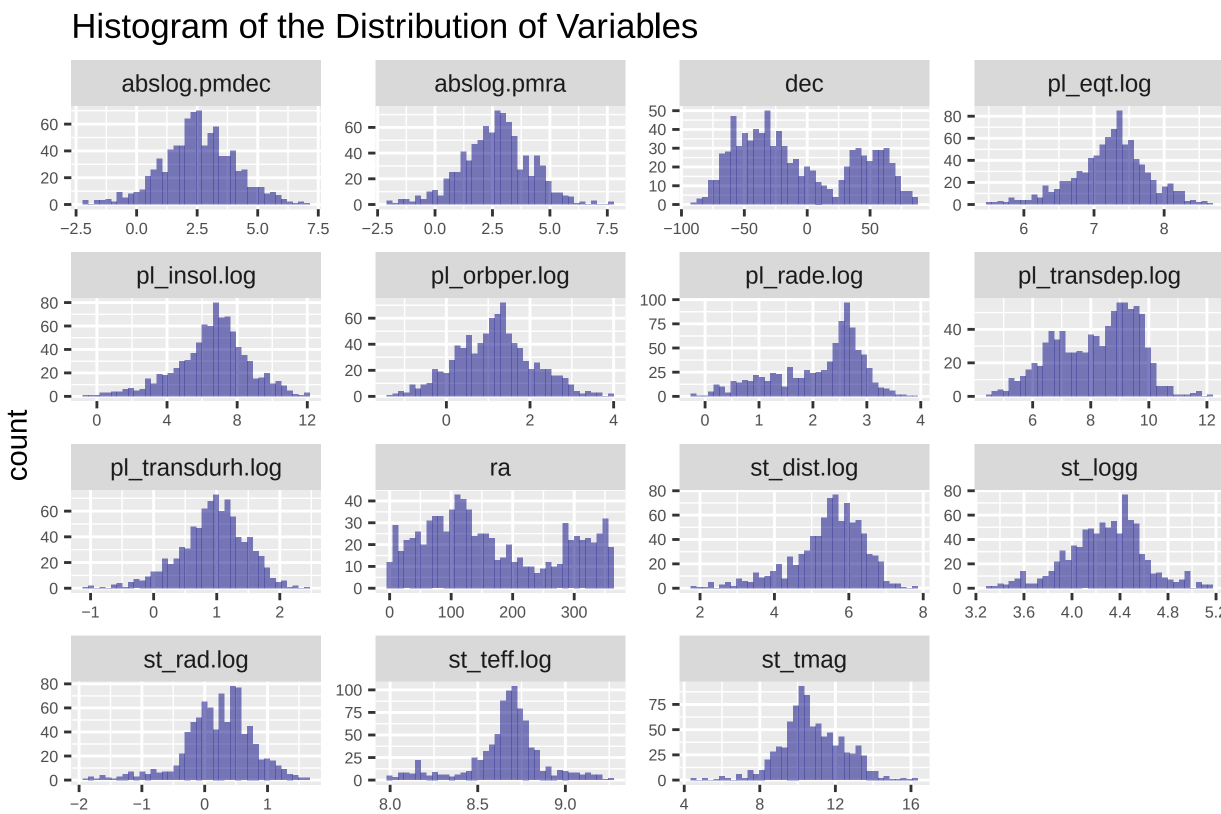
We would like to transform several variables to minimize the effects of these outliers, as well as to produce a more visualizable data set. We do so on both the planetary candidates and the non-planetary candidates data sets. All variables except `ra`, `dec`, `st_tmag`, `st_logg`, and `st_rad` are logarithmically transformed. On the other hand, for `st_pmra` and `st_pmdec`, the logarithm of the absolute value of the variables are calculated.

All the variables have been renamed, if necessary, with the previous name appended with `.log` for logarithmic transformations or prepended with `abslog.` for transformations involving the logarithm of the absolute value of the measurements

Even after transformation, we decided to remove 28 observations from the non-planetary candidate observations that were much farther away from the mean and median than the other observations of certain predictor variable distributions. We also removed the original variables that were trasnformed from both datasets.

Visualizing the Distribution of Predictor Variables

We now would like to visualize the resulting distributions of the transformed variables as histograms.



It appears that the logarithm of the absolute value of how "fast" the host star moves in the celestial latitude (`abslog.pmdec`) and longitude directions (`abslog.pmra`) appear to be approximately normal. Additionally, celestial latitude (`dec`), celestial longitude (`ra`), and the logarithm of the light-blocking amount of transit of the object of interest (`pl_transdep.log`) seem to be bimodal. Furthermore, the logarithm of the planetary orbital period of the object of interest (`pl_orbper.log`) and the logarithm of the host star's radius (`st_rad.log`) appear to unimodal and skewed to the right. Finally, all eight of the other variables appear to be unimodal and skewed to the left.

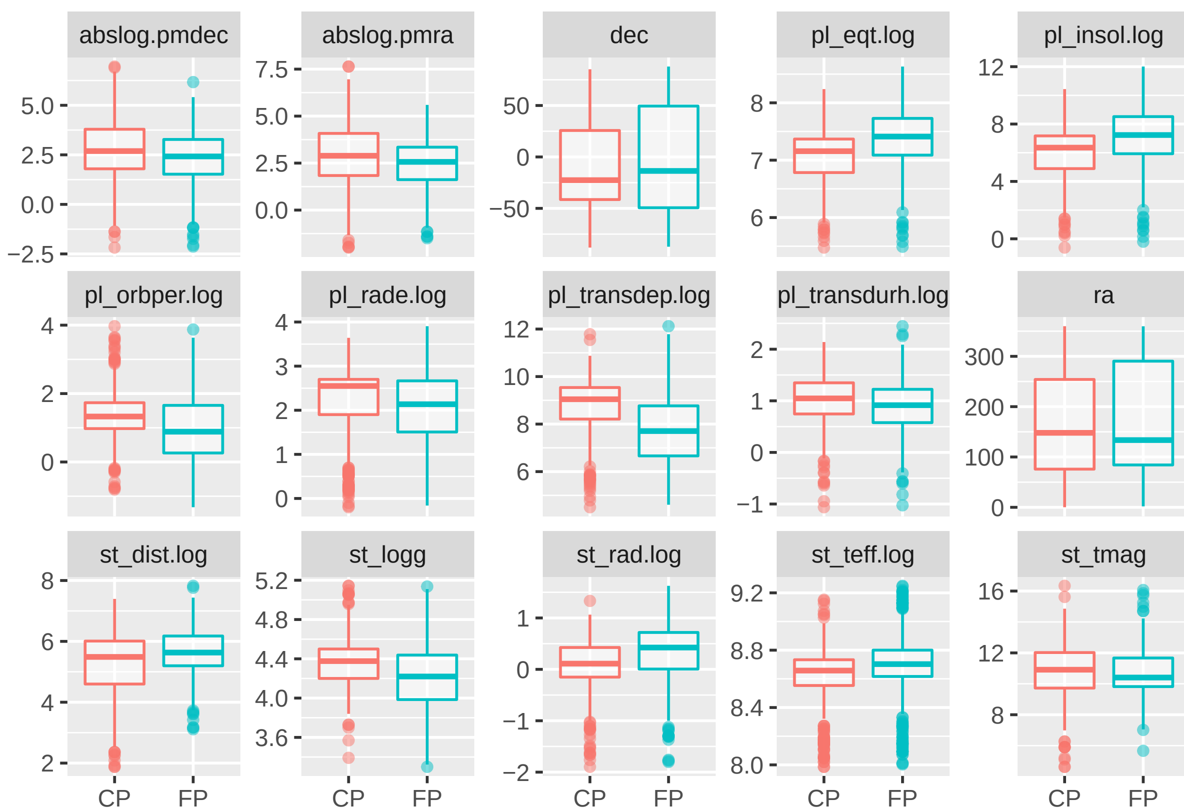
Here are the six number summaries for each of the transformed predictor variables:

```
##        ra               dec              st_tmag            st_logg
##  Min.   :  0.1856   Min.   :-88.121   Min.   : 4.628   Min.   :3.300
##  1st Qu.: 81.3990   1st Qu.:-45.426   1st Qu.: 9.773   1st Qu.:4.082
##  Median :140.0568   Median :-19.420   Median :10.548   Median :4.300
##  Mean   :167.4430   Mean   : -7.011   Mean   :10.762   Mean   :4.276
##  3rd Qu.:279.1010   3rd Qu.: 38.431   3rd Qu.:11.828   3rd Qu.:4.480
##  Max.   :359.9009   Max.   : 87.869   Max.   :16.338   Max.   :5.143
##    pl_eqt.log      pl_insol.log      pl_orbper.log        pl_rade.log
##  Min.   :5.476   Min.   :-0.6121   Min.   :-1.3218   Min.   :-0.1982
##  1st Qu.:6.960   1st Qu.: 5.4196   1st Qu.: 0.5127   1st Qu.: 1.5731
##  Median :7.275   Median : 6.7993   Median : 1.1689   Median : 2.4091
```

```
##  Mean    :7.232   Mean    : 6.5649   Mean    : 1.1571   Mean    : 2.1117
##  3rd Qu.:7.548   3rd Qu.: 7.8318   3rd Qu.: 1.7062   3rd Qu.: 2.6967
##  Max.    :8.632   Max.    :12.0084   Max.    : 3.9699   Max.    : 3.9056
##  pl_transdep.log  pl_transdurh.log   st_dist.log      st_rad.log
##  Min.   : 4.501   Min.   :-1.0619   Min.   :1.877   Min.   :-1.89712
##  1st Qu.: 7.056   1st Qu.: 0.6468   1st Qu.:4.972   1st Qu.:-0.08338
##  Median : 8.465   Median : 0.9832   Median :5.588   Median : 0.25251
##  Mean   : 8.203   Mean   : 0.9403   Mean   :5.439   Mean   : 0.21716
##  3rd Qu.: 9.334   3rd Qu.: 1.2689   3rd Qu.:6.076   3rd Qu.: 0.55099
##  Max.   :12.121   Max.   : 2.4460   Max.   :7.825   Max.   : 1.62531
##   st_teff.log     abslog.pmdec     abslog.pmra
##  Min.   :7.986   Min.   :-2.180   Min.   :-1.995
##  1st Qu.:8.587   1st Qu.: 1.610   1st Qu.: 1.719
##  Median :8.685   Median : 2.516   Median : 2.697
##  Mean   :8.658   Mean   : 2.546   Mean   : 2.649
##  3rd Qu.:8.764   3rd Qu.: 3.508   3rd Qu.: 3.544
##  Max.   :9.249   Max.   : 6.955   Max.   : 7.637
```

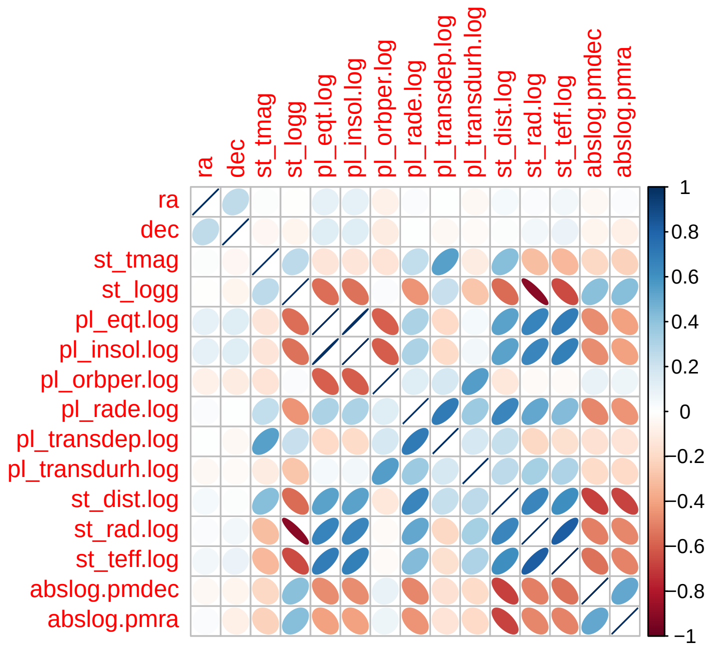
Visualizing the Differences Between Confirmed Planet and False Positives Among Predictor Variables

Below we have displayed a side-by-side boxplot of each predictor variable. These visualizations show how predictor variables can be differentiated by the designation of the observation. Notice that we have removed planetary candidates because those contain observations that are to be confirmed as planets or false positive in the future. After all, we are interested in seeing how planetary candidates are categorized.



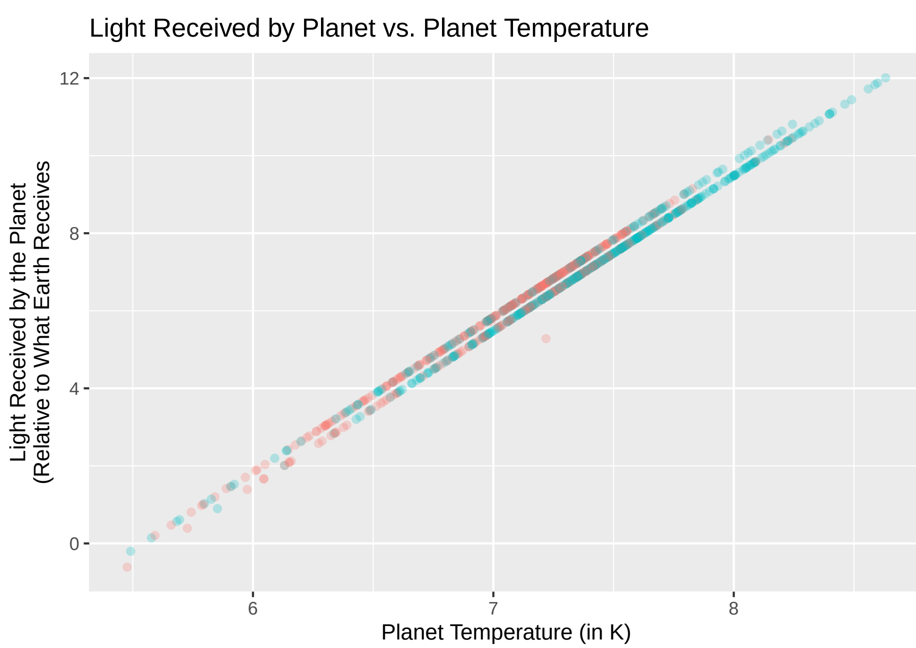
Visualizing the Relationships between variables

Below is a correlation plot displaying how variables in the original data set are correlated with one another:



It appears most of the variables are well correlated with another, with some exceptions, perhaps due to multicollinearity. However, celestial latitude and longitude do not appear to be correlated with any of the variables. Multicollinearity is not of much concern because our intent is prediction.

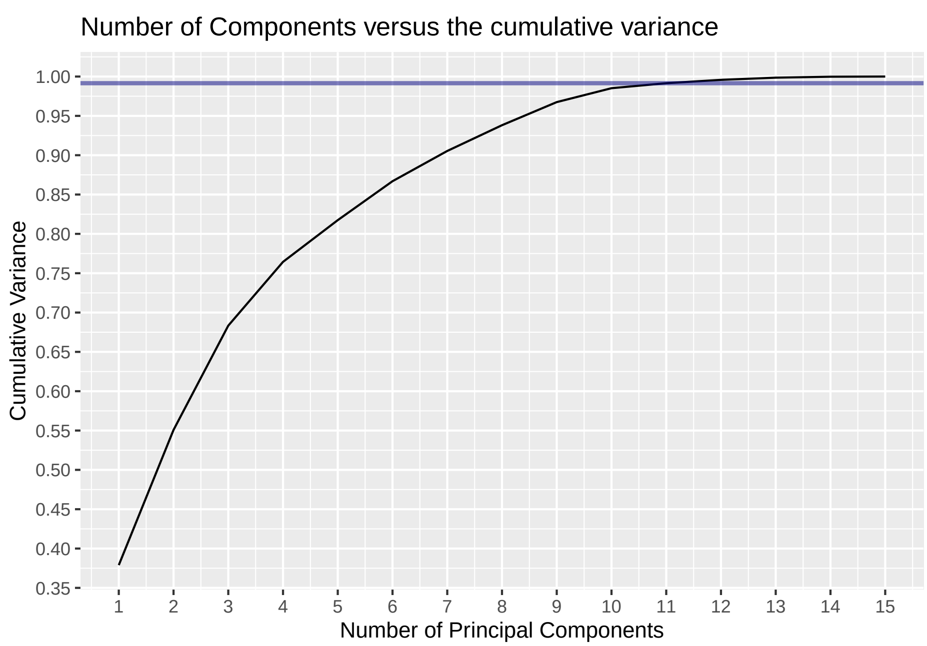
A particularly interesting relationship, however, is that of the light received by the planet and the planet's temperature, which appears to be perfectly correlated. This implies we could remove one of the variables, especially if our purpose is inference. Our model would therefore incorporate the one which yields the most predictive power of the two. However, since our intent is prediction, we can ignore this concern.



Principal Component Analysis on the Data Set

Before we proceed with fitting regression model, we perform principal component analysis. Since there are several variables underlying which observations are confirmed planets or false positives, we intend to determine how many dimensions (principal components) of the data set contribute the most information about the variance encountered within the data set. Our goal is to therefore minimize the number of components such that any more information provides more minimal information than do the previous information. The following cumulative variance plot describe how many variables we should keep in out analysis, in other words, when the variance "elbows" off. A table of specific cumulative variance is also provided:

	Cumulative Variance
1	0.3788958
2	0.5508039
3	0.6833660
4	0.7644246
5	0.8174342
6	0.8669471
7	0.9053876
8	0.9381426
9	0.9674953
10	0.9851886
11	0.9915305
12	0.9957960
13	0.9985694
14	0.9997897
15	1.0000000



We would like to cut off the number of principal components when the cumulative variance is greater than 99 percent. This occurs when we consider eleven principal components, which has a cumulative variance of around 0.9915. However, because dimensionality is only marginally reduced, we can still proceed with regression and other techniques of analysis in the original space without having to worry about multicollinearity.

Regression Models

We will now present the results of regression analysis. Recall that we have already separated PC objects of interests. Before completing regression analysis and applying machine learning techniques, we randomly split that remaining data set, in which 75 percent of the observations corresponds to the training data and the other 25 percent corresponds to the test data. This yields a splits of 222 objects of interest in the test data and 663 in the training data. Later on, we will determine the most effective regression and later on machine learning model which maximizes the area under the ROC curve after learning models using the training data. Afterwards, we will test them against the test data, and finally use those learned models to label the planetary candidates.

Fitting Logistic Regression Models

After fitting a logistic regression model on the data, we obtain the following output summary:

```
##
## Call:
## glm(formula = factor(nonpc.train$label) ~ ., family = binomial,
##     data = nonpc.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1781  -0.8152   0.2385   0.7398   4.5137
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -6.403e+01  1.594e+01  -4.017 5.90e-05 ***
## ra                4.409e-04  9.800e-04   0.450   0.6528
## dec               2.108e-03  2.314e-03   0.911   0.3624
## st_tmag           2.291e-01  1.510e-01   1.517   0.1292
## st_logg          -6.146e-01  7.625e-01  -0.806   0.4202
## pl_eqt.log        1.556e+01  2.299e+00   6.767 1.32e-11 ***
## pl_insol.log     -3.874e+00  5.854e-01  -6.617 3.66e-11 ***
## pl_orbper.log    -4.595e-01  2.673e-01  -1.719   0.0856 .
## pl_rade.log       5.794e+00  9.707e-01   5.969 2.38e-09 ***
## pl_transdep.log  -3.633e+00  5.061e-01  -7.178 7.06e-13 ***
## pl_transdurh.log  4.091e-01  2.868e-01   1.426   0.1538
## st_dist.log       5.992e-01  3.342e-01   1.793   0.0730 .
## st_rad.log       -6.003e+00  1.117e+00  -5.374 7.68e-08 ***
## st_teff.log      -8.244e-01  1.223e+00  -0.674   0.5002
## abslog.pmdec      3.944e-02  9.260e-02   0.426   0.6702
## abslog.pmra       7.404e-02  8.804e-02   0.841   0.4004
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 913.12  on 662  degrees of freedom
## Residual deviance: 652.75  on 647  degrees of freedom
## AIC: 684.75
##
## Number of Fisher Scoring iterations: 5
```

The summary appears to suggest that in this logistic regression model for the labelled data set, the logarithm of planetary temperature, light received by the planet, planetary radius, and the light blocking amount of the planetary transit, distance to the host star, and the host star's radius, are statistically significant predictors of the odds that a certain object of interest has a certain designation.

For the given logistic regression model, the area under the ROC curve (AUC) is 0.881.

Best Subset Selection

We now would like to use best subset selection for determining the set of variables to keep for regression. Since we have too many variables in our data set, we use the greedy log-forward subset selection algorithm, selecting based on Akaike Information Criterion (AIC). Using this metric, however, tends to include not only all important variables but also those that are not as important.

Note that we were able to keep `pl_eqt.log`, `pl_insol.log`, `pl_orbper.log`, `pl_rade.log`, `pl_transdep.log`, `st_dist.log`, `st_rad.log`, and `st_teff.log`, while other variables were removed. It is concerning to see the both `pl_eqt.log` and `pl_insol.log` were kept even though they exhibited high multicollinearity and appear to be measuring the same thing, but as said before, this issue can be ignored for our purposes.

Nevertheless, we generate a new logistic regression model with this subset of variables. The resulting output is

```
##
## Call:
## glm(formula = factor(nonpc.train.sub$label) ~ ., family = binomial,
##     data = nonpc.train.sub)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2430  -0.8194   0.2627   0.7312   4.2832
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -50.6219    14.1459  -3.579 0.000346 ***
## pl_eqt.log       15.0502     2.2519   6.683 2.34e-11 ***
## pl_insol.log     -3.6516     0.5624  -6.493 8.42e-11 ***
## pl_orbper.log    -0.2253     0.2103  -1.072 0.283886
## pl_rade.log       5.2941     0.9223   5.740 9.47e-09 ***
## pl_transdep.log  -3.3598     0.4777  -7.034 2.01e-12 ***
## st_dist.log       0.9676     0.1735   5.577 2.45e-08 ***
## st_rad.log       -5.5154     0.9783  -5.638 1.72e-08 ***
## st_teff.log      -2.4635     0.9329  -2.641 0.008274 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 913.12  on 662  degrees of freedom
## Residual deviance: 658.73  on 654  degrees of freedom
## AIC: 676.73
##
## Number of Fisher Scoring iterations: 5
```

It seems that nearly all of the variables are statistically significant, with the only exception being the logarithm of the planetary orbital period. We repeat the same process as before and calculate the AUC of this simpler logistic regression.

We ultimately find that the area under the ROC curve for the subset selected logistic regression model is marginally smaller than that of the initial logistic regression model (0.8806).

Attempting Lasso and Ridge Regression on the Data Set

As an alternative to best subset selection, we should attempt logistic regression with lasso and ridge regression. These two shrinkage methods penalize models, and the goal in this attempt is to tune the parameter λ.

Lasso regression yielded a λ value of 0.0000809, and ridge regression yielded a λ of 0.0166. These λ values, both approximately 0, suggest that lasso and ridge regression is not needed, as doing so is essentially equivalent to utilizing the initial logistic regression model.

Learning Models through Machine Learning Algorithms

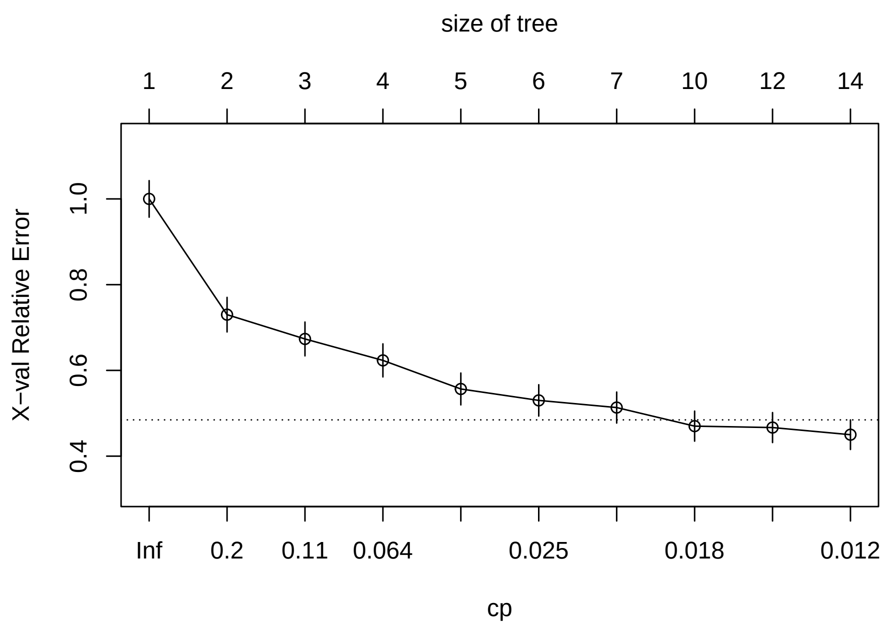
As we are essentially classifying the set of observations as confirmed planets and false positives, we need to incorporate several machine learning techniques, namely:

- Decision Trees
- Random Forest
- Gradient Boosting with XGBoost
- Naive Bayes
- KNN (K-nearest neighbors)
- SVM (Support Vector Machines)—with linear, polynomial, and radial kernals

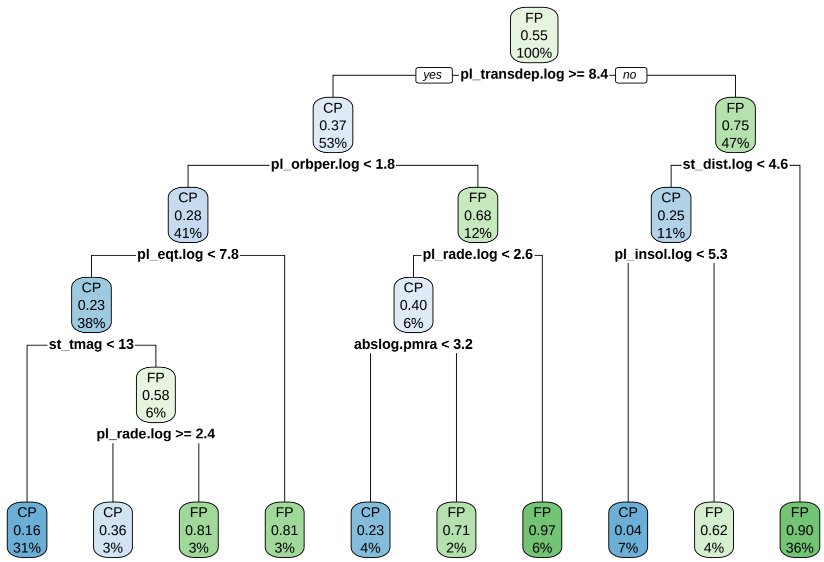
We intend to determine which of the models learned through these techniques, as well as the previous two logistic regression models, performs the most effectively. The most effective model would be one that maximizes the area under the ROC curve.

Decision Trees

We begin by learning an unpruned decision tree and displaying the relationship between the complexity parameter (cp), which is used to determine the size of the tree, and the X-val relative error. Recall that the complexity parameter measures how much of extra leaves in the decision tree improve error measures. Therefore, we decide to cut the tree if extra leaves do not improve the errors.



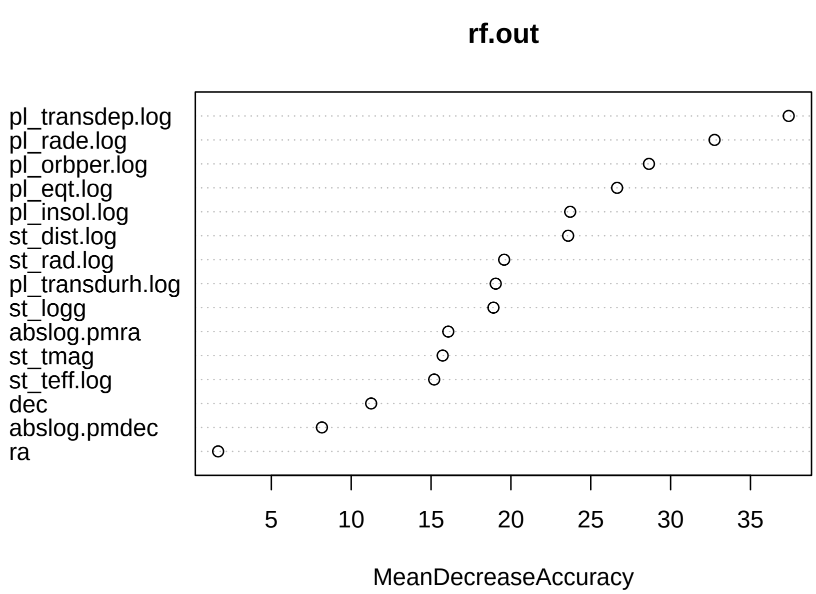
According to the graph above and our previous discussion of the complexity parameter, we need to prune the tree, as the X-val relative error is below the horizontal line when the cp is 0.018. We prune the tree as follows, ultimately creating the following tree:



This decision tree has an AUC of 0.845. The most important variables are those which appear in the tree, with those at higher nodes having greater importance.

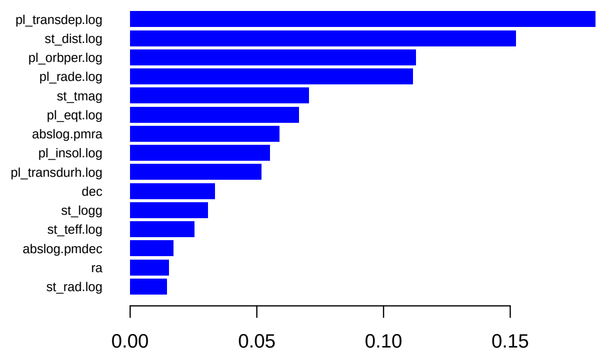
Random Forest

A random forest model yields an AUC of 0.9132. The following importance plot, as determined through mean decrease accuracy, shows that the model agrees with decision trees in that `pl_transdep.log` is the most important variable. It also appears that the same variables appear to be important to both models.



Gradient Boosting with XGBoost

We then attempt learning a model generated through gradient boosting using the XGBoost library. The optimum number of splits is 18. It appears that the AUC is 0.9113. Like the decision tree and random forest models, for the gradient boosting model, `pl_transdep.log` is the most important variable in distinguishing false positives from confirmed planets. The same variables deemed important in the decision tree and random forest models appear in a similar fashion for the gradient boosting model.



Naive Bayes

After learning a model using the Naive Bayes algorithm, we obtain an AUC of 0.8441.

Support Vector Machine

Linear Kernal

We have determined that support vector machines with a linear kernal require a cost of 3.981, resulting in an AUC of 0.9032.

Polynomial Kernal

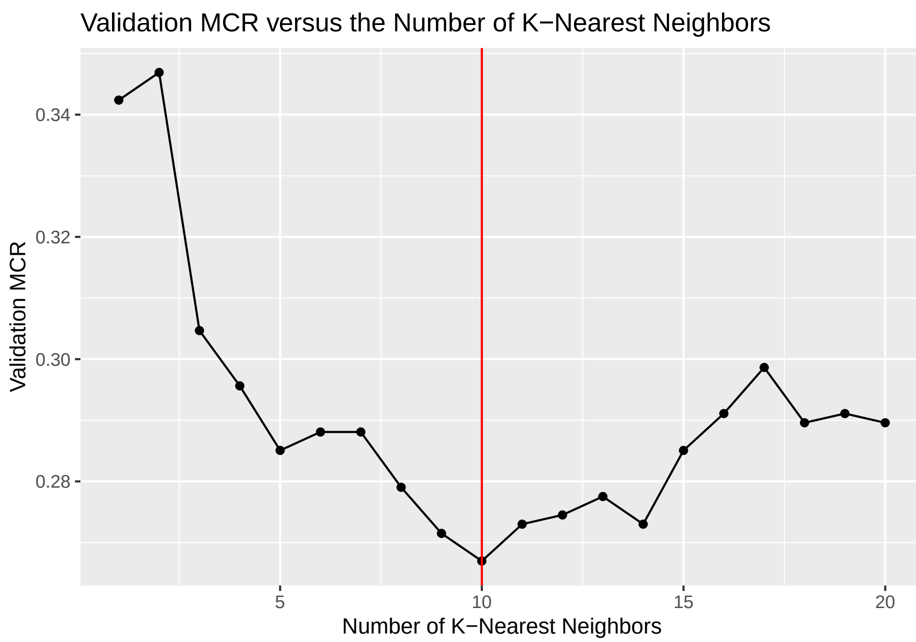
Similarly, the support vector machine with a polynomial kernal requires a cost 562.3413 and degree 2, yielding an AUC of 0.859.

Radial Kernal

The model learned through a support vector machine with a radial kernal requires a cost of 31.623 and a gamma of 0.0316, which results in an AUC of 0.9169.

K-Nearest Neighbor

We now learn a model using the K-nearest neighbor algorithm. It appears that a K of 10 minimized the mis-classification rate. Using $K = 10$ results in an AUC of 0.7868.

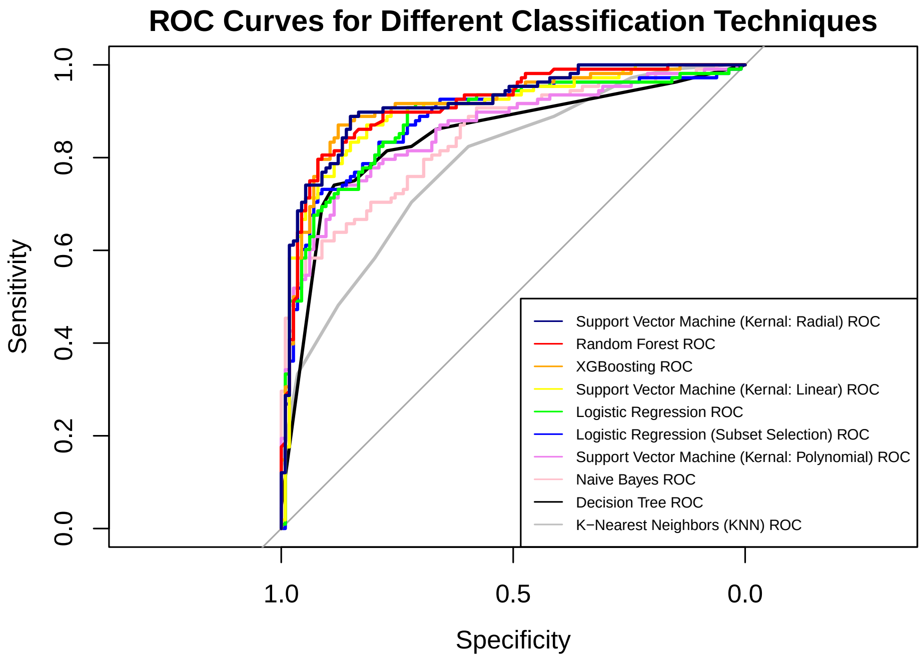


Summary of Results and Best Model Selection

Below is a table consisting of the results from learning regression and machine learning models learned on the data, sorted according to the area under the ROC curve (i.e. how well the model is able to distinguish between Confirmed Planets and False Positives):

Classification Model	AUC
Support Vector Machine (Kernal: Radial)	0.9169
Random Forest	0.9132
Boosting	0.9113
Support Vector Machine (Kernal: Linear)	0.9032
Logistic Regression	0.881
Logistic Regression (Subset Selection)	0.8806
Support Vector Machine (Kernal: Polynomial)	0.8592
Decision Tree	0.845
Naive Bayes	0.8441
K-Nearest Neighbor (KNN)	0.7868

The ROC curves for each model is also plotted on the same graphic below:



The best model appears to be a support vector machine using a radial kernal, with a cost of 31.623 and a γ value of 0.0316. Larger values of cost correspond to higher toleration for violations (overlaps), while smaller values of γ correspond to higher influence from observations that are farther away from the boundary.

We now calculate the Youden's J value in order to determine the threshold for classification which minimized the misclassification rate. We determine that the threshold for classification should be 0.577. The resulting classifcation threshold yields a misclassification rate of 0.131, as well as the confusion matrix below:

```
##                preds
## nonpc.test.resp CP FP
##              CP 97 17
##              FP 12 96
```

Conclusion

It appears that the model learned through support vector machines using radial kernals performs the best among all the models learned in this report with regards to classifying planetary candidates into confirmed planets and false positives. The resulting model has a misclassification rate of 0.13, with a classifcation threshold of 0.577. In other words, when the probability of an observation being a false positive is greater than 0.577, then it is a false positive; otherwise, the observation is a confirmed planet. Using the resulting model, we find that 1254 of the planetary candidates seem to be confirmed planets while 1226 of them appear to be false positives. Although these planetary candidates are not confirmed to be either planet or false positives, the modeled learned seems to approximate to the best of its ability how many of the observations from the *Transiting Exoplanet Survey Satellite* are confirmed exoplanets and which ones are not.

Bibliography and Sources

James, G., et al. 2013, An Introduction to Statistical Learning, Springer

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