Supervised Classification of Exoplanets into Confirmed or False Positive

Matthew Karnes
First Reader: Dungang Liu
Second Reader: Liwei Chen

University of Cincinnati

Abstract

We propose a method of determining whether an object of interest identified by telescope observation can be confirmed as an exoplanet or is a false positive. An open source dataset from the NASA Kepler Mission, which over its lifetime discovered 2,706 exoplanets, is used. Three models were developed in R using this dataset, and each were evaluated on the same criteria of accuracy and error. Of the models, logistic regression and random forest achieved promising results, while the k-nearest neighbors model did not. The logistic regression and random forest models were then used to predict the status of yet unclassified objects of interest. These models are useful for supporting scientific inquiries into exoplanet discovery and can help reduce cost and man-hours necessary to validate exoplanet confirmation.

Table of Contents:

1. Introduction

Astronomy is currently in a period where the amount of data being collected is surging to levels that cannot be dealt with by human hands alone. Telescopes that are being developed and constructed will take in upwards of ten terabytes of data every night in the form of images and signals which will need to be categorized and analyzed [12]. In the past, crowdsourcing projects have been employed and machine learning initiatives are underway, but new methods and systems will need to be created as information sizes continue to grow. Using data from the NASA Kepler Missions we create several classification models that aim to be able to predict whether or not an observed exoplanet is real or a false positive. Using these models we can then give the probability that a yet unclassified exoplanet is either real or a false positive.

This is done using the Kepler dataset, with the variable koi_disposition as our response. This variable contains three different classifications for the objects of interest, they being "Candidate", "Confirmed", and "False Positive". For the purposes of training our models, we will use only the "Confirmed" and "False Positive" categories, making this a binary classification task. We do not use the "Candidate" data as since it has not yet been classified, it will be not useful in determining whether or not a planet is real. Instead, once we have trained our models we will use this candidate data for predictions. We will report the candidates that are deemed most likely to be confirmed and the most likely to be false positives.

This project is valuable for the astronomy community as a whole, as previous techniques for finding exoplanets are generally done by hand. The most common form of detection is through the use of light curves being interrupted by transiting exoplanets [1]. Which is to say that an exoplanet moving in front of a star causes a shadow to be visible on the star, leading to a decrease in the light seen. Machine learning will allow for easier detection that removes the human element of the equation. In the face of increasing amounts of data, models such as the ones built for the project will be able to work faster and with more data. This will allow for more efficient allocation of resources and time in verifying these exoplanets.

2. Data Description

The NASA Kepler Space Telescope was a telescope first launched into orbit around the Sun in March of 2009 [10]. The telescope was designed to survey the portion of the Milky Way that the Earth resides in and attempt to discover exoplanets. Specifically, the telescope searched for Earth-sized planets within or near the habitable zone of the stars they orbited. The program gathered data from May 2009 until a mechanical error led to a halt in the program. In 2014 the program was restarted as the K2 Mission, which would use the now disabled Kepler telescope to survey red dwarf stars for exoplanets. The K2 Mission ended in November of 2018 when the telescope officially ran out of fuel. Over its lifetime the telescope observed 530,506 stars and 2,709 exoplanets were discovered from the observations [7]. A new mission, TESS, was also started in April 2018 and continues to search for new exoplanets. In total, 5,005 exoplanets have been discovered.

The dataset that this paper will make use of is found at the NASA Exoplanet Archive, which is operated by the California Institute of Technology [5]. Each point of data within the dataset refers to a "Kepler Object of Interest" or KOI. These KOIs are exoplanets that were observed, or thought to be observed by the Kepler telescope during its mission. For the purposes of this paper we will be using the KOI Cumulative list, including each column, as the data allows

for specific variables to be chosen and used for analysis. Initially, the dataset contains 9564 records and 83 variables. These variables can be broken down into several categories:

- Descriptive: These variables include categorical data for the objects of interest. This includes identification numbers, the team that first discovered the object, the time it was found, the location of the team that made the discovery, and the assigned name of the object. Of the 83 variables, 20 are categorical.

- Disposition: The koi_disposition and koip_disposition are two variables that describe the objects of interest and whether they are real exoplanets or not. Disposition contains three different states for the objects of interest, they being "Candidate", "Confirmed", and "False Positive". These classifications tell us whether or not an object is truly an exoplanet, has not been categorized yet, or is known to be a false positive and thus not a real exoplanet. Koip_disposition contains only "Candidate" and "False Positive" and is the most probable category for the objects based on koi_score, a score between 0 and 1 where 1 is more likely to be a real exoplanet.

- Flags: There are four flags in the data that are set if the data is known to be a false positive. They tell the specific reasons as to why an object is a false positive, such as being visual artifacts from a nearby star or from being an eclipsing binary, or two stars that orbit one another blocking the light of each other.

- Planet Data: These variables describe the exoplanet object of interest, including various orbit data, such as the right ascension and declination of the planet and the orbital period and shape of the orbit. Nearly half of all variables in the dataset describe the supposed exoplanets.

- Stellar Data: The final category describes the star that the object of interest orbits. These variables include the temperature of the star, the metallicity of the star, the radius and mass, and how many known planets orbit the star.

Detailed descriptions of each variable can be found in the "Data Columns Definitions" documentation on the Exoplanet Archive [6].

3. Data Analysis and Variable Selection

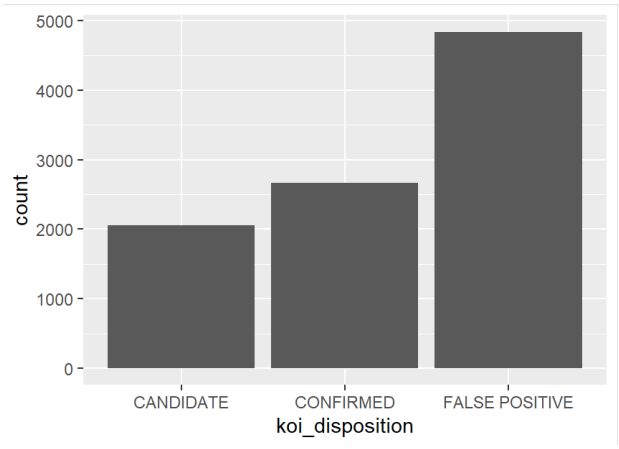

Fig. 1: Distribution of Candidate, Confirmed, and False Positive Objects of Interest

3.1 EDA

Initial data exploration of our response variable tells us that there are 2666 “Confirmed” exoplanets (not equal to the amount discovered from the Kepler Mission, as new exoplanets have been discovered since the creation of the dataset), 2058 “Candidate” objects, and 4840 “False Positive” objects. As can be seen in Fig. 1, our classes are imbalanced at this time. As the candidate objects will not be used in training our models, they are removed from the training dataset.

13 variables within the dataset contain either no information or contain information that would lead to data leakage within the models. The variables that would result in leakage are the four flag variables and koi_score. Each false positive in the Kepler data is marked with one or more of these flags meaning that if used within a model it will instantly tell the model that the object is a false positive. The same is true for koi_score, as it is a confidence in whether or not an object is an exoplanet and as such a higher score will instantly identify the object as real.

Within the dataset there are a number of variables which are a part of the same category of data. These are the “fwm”, “dicco / dikco”, and “mag” variables. The fwm variables are from an “FW Source” meaning that it is “flux-weighted”. This means that the variable is calculated by measuring the center of light and how it changes during a transit. This is done to counteract the effects of an eclipsing binary. There are seven “fwm” variables that contain the flux-weighted Declination and Right Ascension of the object, calculated through various means. Declination and Right Ascension are equivalent to the object moving around the star equivalently to changing one’s latitude and longitude on Earth. We check these variables for multicollinearity and find that only “pdeco” and “prao” have significant correlation. It is also found the variables “sra” and “sdec” have significant correlation with the non-flux weighted right ascension and declination variables.

The next category is the “mag” variables. Within optical astronomy one of the main forms of light viewed is infrared light. There are multiple “bands” of near infrared light which are based on the wavelength of the light viewed. Within our dataset there are eight “band” variables that are all multicollinear, as can be seen in Fig. 1.

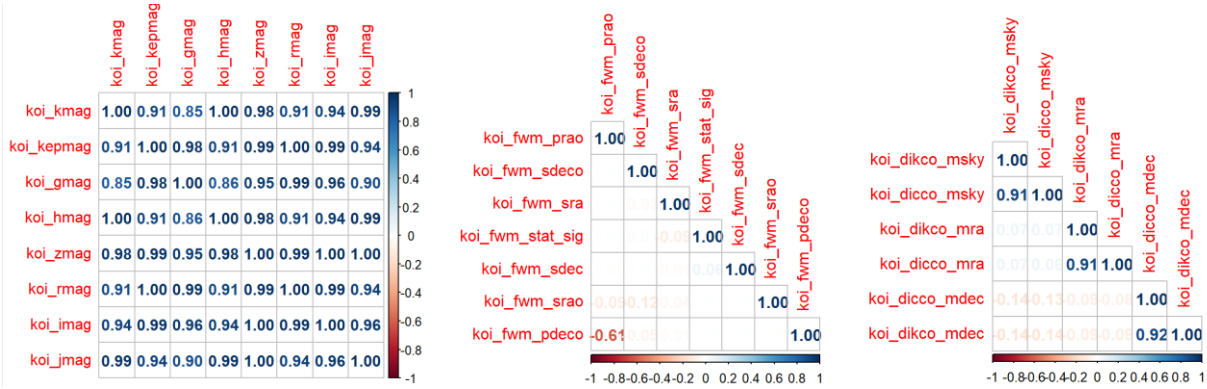

Fig. 2: Correlation Plots for "mag" variables (left), "fwm" (center), and "dicco"/ "dikco" (right).

The final category is the "dicco" and "dikco" variables. These six variables each detail different angular offsets for the Declination and Right Ascension of the objects of interest. Of these variables there is a "dicco" and "dikco" variant of the Declination, Right Ascension, and the angle of the plane of the sky, and each are significantly correlated between the two types.

3.2 Variable Selection

There are, in total, 83 variables within the Kepler Cumulative List dataset. In order to select the variables for use in our models, we must first consider what parts of our dataset we wish to use. Of the data available, the variables that provide descriptive information about the objects of interest do not have value within our models. The assigned name and who discovered the object do not tell us whether or not the object is an exoplanet or false positive and as such will be removed for the purposes of the model. This is also true of the 13 variables that are either null or contain no variation in the data. There is no information that is beneficial to our models within these variables and as such they are removed from the training data.

The next set of variables to consider is the flags that are set for false positives. If we were to include these within our training data, they would most likely provide too much information to the model, causing data leakage and too perfect of a model. In reality, these flags are set only after a false positive has been confirmed to be a false positive, and as such would not exist when using the model to predict on new data. We wish to build off of the attributes of the objects and the stars they orbit and not these flags, so they are removed.

Finally, in each of the three categories described in the previous section there is heavy multicollinearity. Multicollinearity is a problem that arises whenever a predictor is highly correlated with one or more other predictors [4]. This can lead to inflated standard errors and statistical insignificance in predictors that should be significant. In the case of the categories shown in the previous section, the high correlation amongst the variables may lead to our models being less accurate than they could be. However, for initial model building the correlated variables are not removed, and as they are shown to be detrimental to our models they are removed.

Step AIC is then implemented set to "both". This means that after an initial selection of variables is given for the model, the stepAIC package will iterate through the available variables,

both increasing and decreasing the amount used, to find the best model. The AIC, or Akaike information criterion is then used as the scoring method for a "better" model [4].

3.3 Preprocessing

An important question that must be answered with our data is what to do with missing data. Within the Kepler dataset there are 2143 samples which contain some amount of missing data. Options for dealing with this problem include dropping the null values and imputing a new value. Imputing would entail creating our own data, such as a median or mean, or possibly performing multiple imputations in order to create an accurate estimate based off of similar observations. However, for this project we will be instead removing the observations with null values. While this does lead to a loss of information, the large variations that occur naturally in exoplanets means that imputation could lead to incorrect data. So, we remove the 2143 samples that contain null values. This has the side effect of balancing our data as the majority of samples with missing data are false positives. This brings our count to 2938 false positives and 2425 confirmed exoplanets.

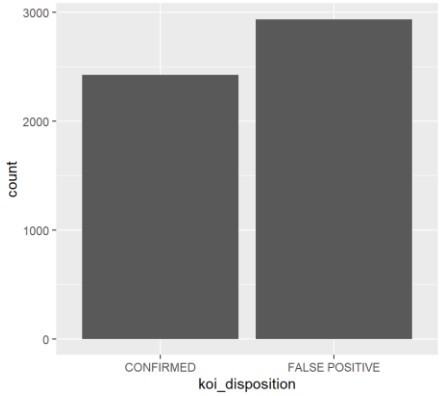

Fig. 3: Distribution of Confirmed and False Positive after removing null values.

For our model building, the data is split into two sets for training and testing. We use an 80/20 split, meaning that 80% of our data, or 4290 observations, for the training data, and 20%, or 1073 observations, as our testing data. It is important to split into to training and testing groups so that our model will be tested on new data that it has not yet seen. The data is not scaled or transformed except in the case of our k-nearest neighbors model. KNN relies on the distance between data points and as such needs standardized data. This is not true of our other models and as such will not be done.

4. Models

4.1 Logistic Regression

The first of the models that we will build is a logistic regression model. Logistic regression is a relatively simple model that is used for binary classification. Within a logistic regression the logistic curve relates our independent variable to the probability of being 1 or 0. The dependent variable is a logit or the natural log of the odds such that

$$\text{logit}(p) = \log(\tfrac{p}{1-p}) = \beta_1 + \beta_2 X \quad (1)$$

We can then assume that the logit scale is linearly related to our independent variable and that

$$p = \frac{e^{\eta}}{1+e^{\eta}} \quad (2)$$

where η is $\beta_1 + \beta_2 X$. This gives us our probability scale from 0 to 1 in the form of a sigmoid, on which we can set a probability threshold to determine whether or not an exoplanet is confirmed or not [2]. Logistic regression was chosen for its ease of use and power, as well as for its clarity and interpretability.

4.2 K-Nearest Neighbors

The next model to be built is a k-nearest neighbors classifier. This model was chosen due to its general ease of use and its interpretability. The model, taken from the FNN package, uses the Euclidean distance between the k nearest training set vectors to determine the classification of the current observation being decided, as seen in formula (3) [11]. This is to say that the point is decided upon by what the majority of the nearby other points. If the majority of nearby points are confirmed then the unknown point is also a confirmed exoplanet. For the purposes of our model, the k is set at 65, the square root of the number of training observations [9].

$$\hat{Y}(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i \quad (3)$$

4.3 Random Forest

The third and final model employed is a random forest model. This model was selected due to its high power and ease of use. Unlike a decision tree, a random forest requires no pruning as each tree is grown to the largest size possible [3]. When choosing to split a tree, the model will take a selection of the variables available and make a random selection. The split is then made on the best split from the random assortments tried. The forest also does not require any sort of cross-validation due to the fact that as the model trains it automatically builds each tree on a bootstrap of the available data. Approximately one-third of observations are left out for each tree constructed. Variable importance and probabilities are also inherently calculated leading to easy interpretation. 500 trees are fit to the training data for our model.

5. Results

5.1 Metrics

For each of our models the accuracy and predictive power is measured through two different ways. Firstly, the models are evaluated using the Area Under the Receiver Operating Curve or the AUROC. This is to show how well our models are able to separate the two classes from one another, with a score closer to 1 being the best [4]. Next, the models are evaluated on the Brier Score or the error of our models when it comes to the probabilities of the outcomes.

The lower the Brier Score, the more accurate the probabilities of our predictions [8]. When combined with the AUROC, we can determine both if our models are able to correctly predict which class the observations are and see the accuracy of the probabilities of those predictions.

5.2 K-Nearest Neighbors

The first model to look at is the k-nearest neighbors. The model was first trained on the 80% training data before being evaluated on the 20% testing data. The model performs admirably with a Brier Score of 0.38, or an error of 0.38, and an AUROC of 0.65. As what stated before, the lower a Brier Score and the higher an AUROC the better the model is able to predict the classes. From these scores we can tell that the model has only middling predictive power and would most likely not be considered to have satisfactory power. The ROC curve in Fig. 4, tell us that as the probability cutoff is higher the model cannot accurately classify the data points. This is corroborated by the calibration curve in Fig. 4, which shows that as the probability grows closer to 1, the model has a tendency to predict lower probabilities as higher than they actually are. The classification table can be found in Table 1.

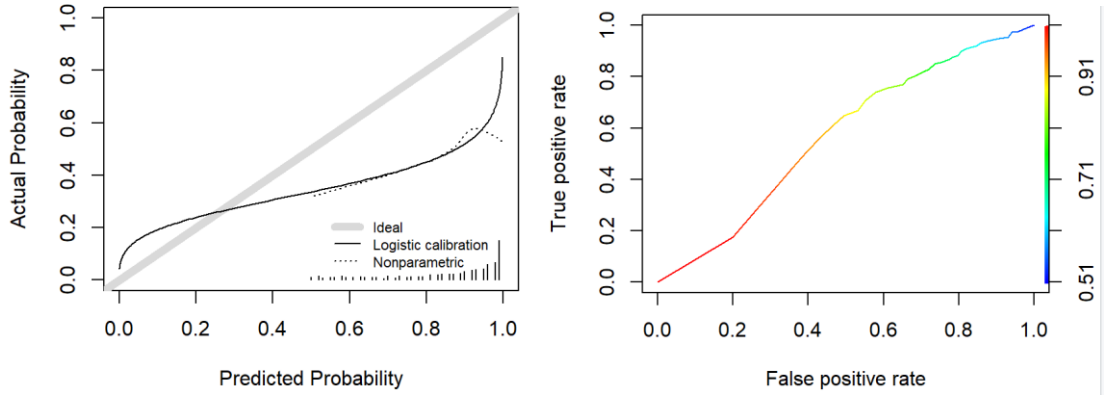

Fig. 4: Calibration Curve of KNN model (left) and ROC curve of KNN (right)

5.3 Random Forest

The random forest model performed at a level much higher than that of the KNN model. The AUROC of the model was found to be 0.98 and the Brier Score comes out to be 0.06. These are incredibly high scores and tell us that the model has significant predictive power. From the ROC curve we can see that the model is nearly perfect, but is not perfect which is an important distinction telling us that the model is not overfit, since this is on our testing data. From the calibration curve we can see that the random forest has a tendency to predict the probabilities as slightly under the actual probability, but when compared to the KNN model, we can see vast improvement. A classification table can be found in Table 1.

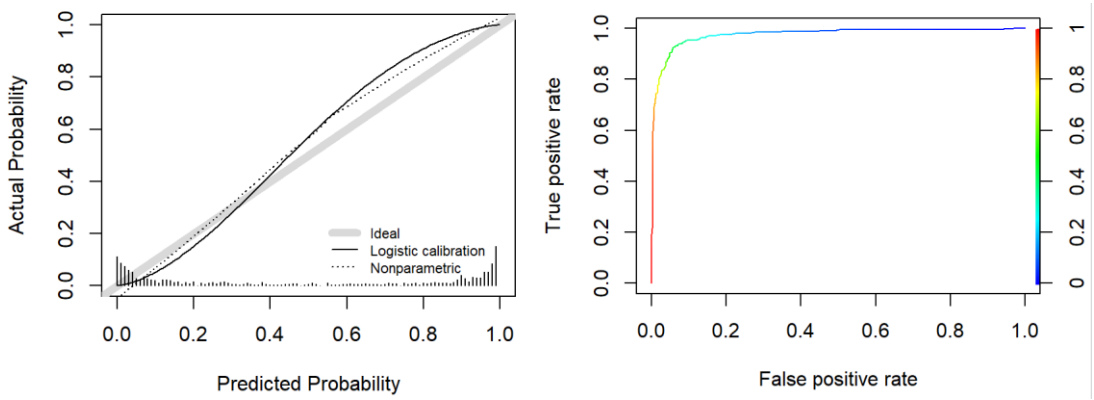

Fig. 5: Calibration Curve of Random Forest model (left) and ROC curve of Random Forest (right)

5.4 Logistic Regression

The last model to test is the logistic regression model. When testing on the 20% test split the model produces an AUROC of 0.95 and a Brier Score of 0.09. This tells us that the model has high predictive power and that the probability error is low as well. The model has nearly the same predictive power as the random forest, but from the calibration curve in Fig. 6, we can see that the probability error manifests in different ways. Whereas the random forest falls off as the probabilities grow closer to one, the logistic regression model is consistently under the actual probability throughout the entire spectrum. So, while the random forest is overall more accurate, the logistic regression may be more interpretable, as we know the model is always slightly under the actual probability. This is in contrast to the random forest which may be slightly over or under the actual depending on which class you are observing.

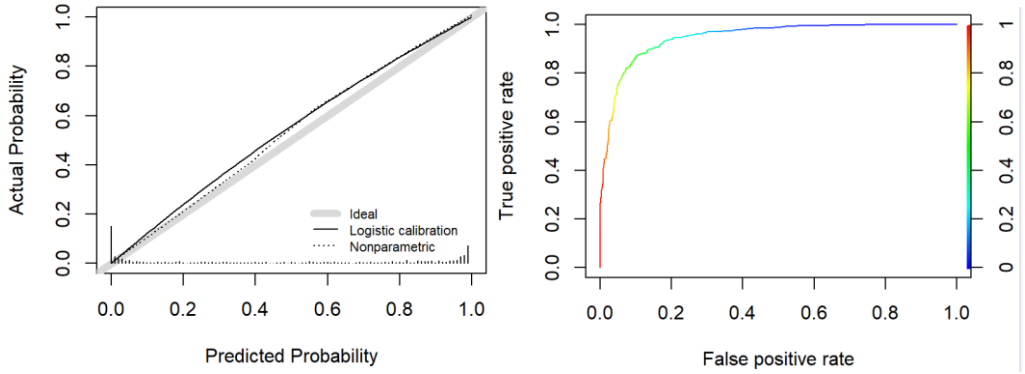

Fig. 6: Calibration Curve of Logistic Regression model (left) and ROC curve of Logistic Regression (right)

K-NN		Predicted		Logistic Regression		Predicted		Random Forest		Predicted	
		Confirmed	False Positive			Confirmed	False Positive			Confirmed	False Positive
Actual	Confirmed	471	108	Actual	Confirmed	61	494	Actual	Confirmed	26	529
	False Positive	9	485		False Positive	442	76		False Positive	467	51
AUC		Brier Score		AUC		Brier Score		AUC		Brier Score	
0.65		0.38		0.95		0.09		0.98		0.006	

Table 1: Confusion Matrix of K-NN (left), Logistic regression (center), and Random Forest (right) Models on Test Split

5.5 Variable Importance

An important metric to take into account is the variable importance within each model. While this is not inherently measured within the k-nearest neighbors model it is defined by the logistic regression model and the random forest. For the purposes of this report we will focus on the logistic regression most important variables, or the variables that most influence its decision making process. The top ten most important variables are found in Fig. 7. We will now go more in depth on the top five variables so as to provide a reasoning as to why these variables influence the model so much.

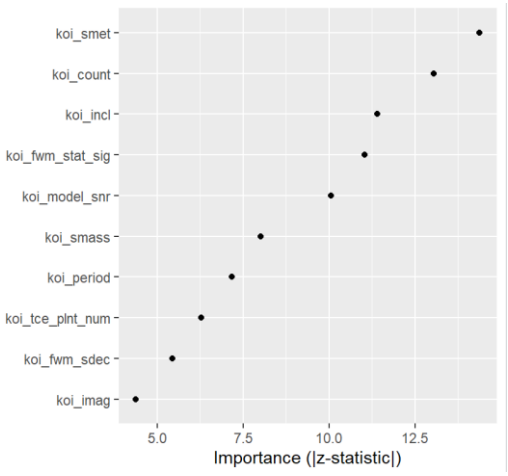

Fig. 7: Variable Importance Graph for Logistic Regression

5.5.1 koi_smet

The most important variable within our logistic regression model is found to be koi_smet. This feature is defined as the base-10 logarithm of the Iron to Hydrogen ratio at the surface of the star the object orbits, normalized by the solar Iron to Hydrogen ratio. This is to say it is the amount of chemical makeup of the surface of the star based on the wavelengths of light observed. In Fig. 8, we can see that in our data false positives will span the entire range of koi_smet values, from 0.5 to -2, while confirmed will only occur from 0.5 to -1, with some outliers. This is why the variable is of such great importance, as an occurrence below -1 will then be much more likely to be a false positive.

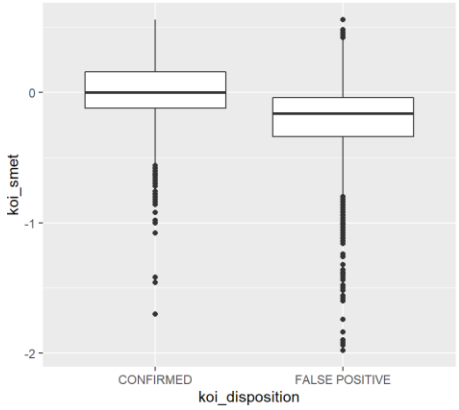

Fig. 8: Distribution of koi_smet for Confirmed and False Positive objects

5.5.2 koi_count

Next is koi_count, which is the amount of exoplanets that have been confirmed to be orbiting the star the object in question orbits. That is to say that a higher koi_count means that there are more confirmed exoplanets orbiting this star, than a lower koi_count. From Fig. 9, we can see that false positives will generally have no other exoplanets orbiting the star in question, aside from in the case of outliers, while confirmed exoplanets will very often have high amounts of other exoplanets orbiting the star. This may indicate some amount of bias in the process of discovering new exoplanets, as the mission may have allocated more resources towards studying stars that were already confirmed to have exoplanets.

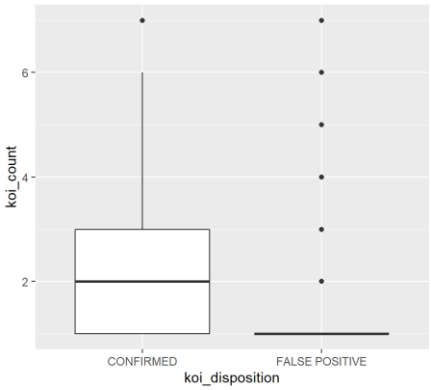

Fig. 9: Distribution of koi_count for Confirmed and False Positive objects

5.5.3 koi_incl

The third most important variable is the inclination of the object detected. The inclination is the angle between the plane of the sky and the orbital plane of the object of interest in degrees. This is essentially the tilt of the object in relation to its orbit, much like how the Earth is tilted on its axis. From Fig. 10, it is apparent that confirmed exoplanets have a smaller range of inclination than what is seen in false positives. Aside from outliers, confirmed objects have a mean of around 80 degrees with an interquartile range smaller than that of false positives. Confirmed planets also rarely have an inclination smaller than 50 degrees while false positives go well below that.

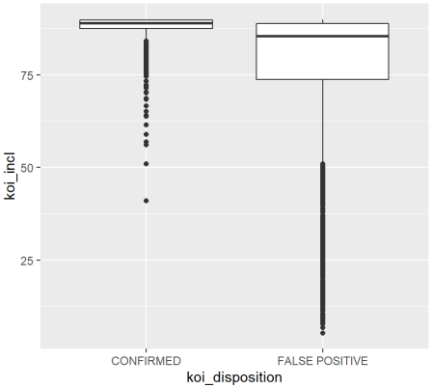

Fig. 10: Distribution of koi_incl for Confirmed and False Positive objects

5.5.4 koi_fwm_stat_sig

koi_fwm_stat_sig is the fourth of our important variables and it is the indication of whether there is significant flux-weighted offset between in and out-of-transit images. This is the confidence that the image seen of an object of interest transiting is in fact on the star that is being studied. Lower confidences happen when there is significant interference. Fig. 11 shows us that false positives will generally have a very low significance while confirmed will have a mean of around 0.010 and a large interquartile range.

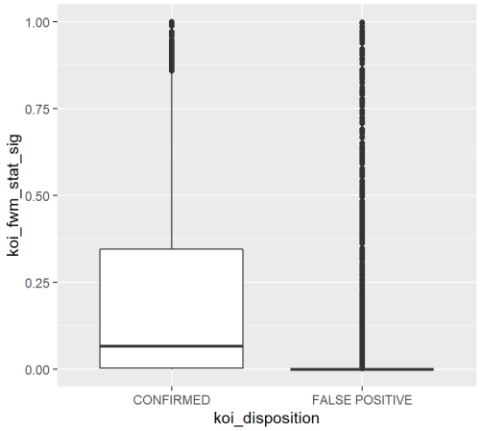

Fig. 11: Distribution of koi_fwm_stat_sig for Confirmed and False Positive objects

5.5.5 koi_model_snr

Finally, is koi_model_snr, which is the change in brightness of a star due to the movement of an exoplanet or object orbiting the star, known as the transit depth. It is the ratio of the area of the object and the area of the star. As seen in Fig 12, a confirmed exoplanet will have a relatively small transit depth, while false positives will have a transit depth above 5000, which a confirmed exoplanet will never have. This tells us that a true exoplanet will be relatively small in comparison to the star it orbits, while false positives are often much larger in comparison.

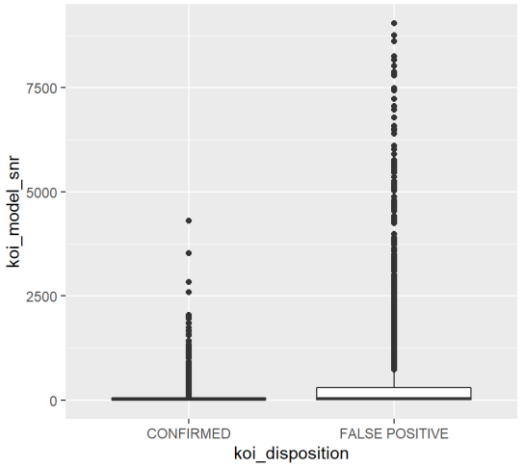

Fig. 12: Distribution of koi_model_snr for Confirmed and False Positive objects.

6. Candidate Predictions

The logistic regression model and the random forest are both found to have significant predictive power and low error in their probabilities. As such, we will use these two models in order to predict the Candidate data that was initially removed from our training and testing data. The K-NN model is not going to be used for prediction as it was not sufficient in its metrics. For the predictions we will be using a cutoff of 0.9 probability, as if this were to be used for an actual NASA mission, then we would wish to only have the most likely candidates. The real world validation necessary means that we would not want to waste our time with less likely objects of interest. Table 2 shows us that the logistic regression is more likely to classify candidates with a probability above 90% than the random forest. This may be due to the fact the random forest was more likely to have lower probabilities than actual as was seen in the calibration curve in Fig. 5. Also included in Table 2, is a sample of the top ten most likely candidates as agreed upon by both models. Direct comparison of the probabilities shows that the models can disagree on the classification and as such we present candidates both models classify as confirmed.

Logistic Regression		Random Forest		Ten Most Likely Candidates	
				Name	
				K046747.01	K02037.03
Predicted				K01175.02	K02732.04
Confirmed	False Positive	Confirmed	False Positive	K01082.01	K04647.01
				K01590.01	K02220.04
311	1008	203	1116	K04032.05	K02579.02

Table 2: Model predictions for Candidate data and ten most likely candidates

7. Conclusion

As Astronomy moves forward into an age of increased data production new methods and means of analyzing this data will be necessary. Within this paper we develop three models to predict whether or not an object of interest observed by the Kepler Mission from NASA is a false positive or a confirmed exoplanet. A dataset from the Kepler Mission was used, which after data preparation contained 2938 false positives and 2425 confirmed exoplanets. This binary classification task was completed by three different models and then the results of each model were compared. Analysis shows that the logistic regression model and random forest model, each built with default settings, had significant and sufficient predictive power, while the k-nearest neighbors model was found to have insufficient power. The random forest model was found to have the highest AUROC and lowest Brier Score, but calibration curves show that the logistic regressions probabilities may be more in line with reality. These two models were then used for predicting whether candidate objects of interest were confirmed or false positives with a 90% probability cutoff. The logistic regression found 311 of the candidates to be confirmed and the random forest confirmed 203.

This project is important for Astronomy analysis and classification. The confirmation and validation of objects of interest costs the operation time of the various ground and space based telescopes used, and is limited by budgetary constraints. Using a model such as these will lead to better and more efficient use of time and resources in the search for exoplanets. Less time will be spent finding false positives and more confirmed exoplanets can be discovered.

References

[1] 5 Ways to Find a Planet (n.d.). In *NASA Exoplanet Exploration*. Retrieved from https://exoplanets.nasa.gov/alien-worlds/ways-to-find-a-planet/

[2] Brannick, M. (n.d.). Logistic Regression. Retrieved from http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html

[3] Breiman, L., & Cutler, A. (n.d.). Random Forests. Retrieved from https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[4] Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data Scientists* (Second ed.) Sebastopol, CA: O'Reilly Media, Inc.

[5] Data Columns in Kepler Objects of Interest Table (n.d.). In *NASA Exoplanet Archive*. Retrieved from https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html

[6] Cumulative KOI Data (n.d.). In *NASA Exoplanet Archive*. Retrieved from https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=cumulative

[7] Exoplanet and Candidate Statistics (n.d.). In *NASA Exoplanet Archive*. Retrieved from https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html

[8] Goldstein-Greenwood, J. (n.d.). A Brief on Brier Scores. In *University of Virginia Library*. Retrieved from https://data.library.virginia.edu/a-brief-on-brier-scores/

[9] Joby, A. (n.d.). What Is K-Nearest Neighbor? An ML Algorithm to Classify Data. In *g2*. Retrieved from https://learn.g2.com/k-nearest-neighbor

[10] Kepler Mission Information (2021, February 10). In *NASA Exoplanet Archive*. Retrieved from https://exoplanetarchive.ipac.caltech.edu/docs/KeplerMission.html

[11] What is the k-nearest neighbors algorithm? (n.d.). In *IBM*. Retrieved from https://www.ibm.com/topics/knn

[12] Zhang, Y. and Zhao, Y., 2015. Astronomy in the Big Data Era. *Data Science Journal*, 14, p.11. DOI: http://doi.org/10.5334/dsj-2015-011