Identifying misclassified exoplanets

Jane Hsieh - Leo Chen - Srujana Rao Yarasi - 36-600: Overview of Statistical Learning and Modeling

Introduction

Problem: Identifying exoplanets, or planets that lay outside our Solar System, is a difficult problem due to the distances involved and the limitations of our imaging technology.

The transit method: One technique for identifying exoplanets is imaging stars to detect when an orbiting exoplanet passes in front of its host star, ergo eclipsing our view of the star.

Data collection: Between 2009 and 2013, NASA's Kepler satellite observed over 100,000 stars to detect potential exoplanets with the transit method. Scientists later classified the observations as "confirmed" exoplanets or "false positives."

Goal: In this project we train a classifier that predicts whether exoplanets exist or not using observed properties of the planetary candidates and the stars they orbit taken from the Kepler satellite.

Data

This dataset comes from the Kepler satellite and contains 6859 data points. It contains 17 predictors, though we removed koi_eccen (the orbital eccentricity value) as that column contained zero for all rows.

Variable	Definition
koi_prad	Radius of the planet
koi_ror	Planet radius divided by the stellar radius
koi_slogg	Log10 of acceleration due to gravity at surface of star
koi_smass	Mass of the star
koi_smet	Log10 of Fe to H ratio at surface of the star, normalized by the solar Fe to H ratio
koi_srad	Photospheric radius of the star
koi_srho	Fitted Stellar Density
koi_steff	Photospheric temperature of the star
koi_teq	Approximation for the temperature of the planet

Variable	Definition
koi_period	The interval between consecutive planetary transits
koi_depth	Fraction of stellar flux lost at minimum planetary transit
koi_dor	Distance between planet & star at mid-transit divided by stellar radius
koi_duration	Duration of observed transits
koi_impact	Sky-projected dist. b/w centers of of stellar disc & planet disc at conjunction, normalized by stellar radius
koi_incl	Angle b/w plane of the sky (perpendicular to the line of sight) & the orbital plane of the planet candidate
koi_insol	Equilibrium temperature based on stellar parameters

Analysis

We compared the accuracy of multiple models:

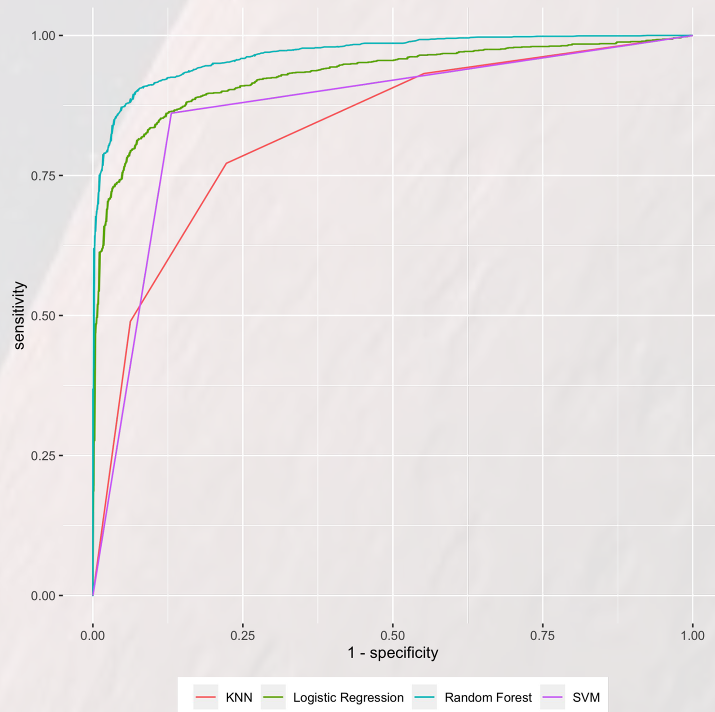
1. Logistic regression: uses a logit function to model probabilities for a binary response variable
2. Random Forest: trains multiple decision trees on subsets of the data and aggregates the probabilities derived from all the trees
3. KNN: uses similarity metrics to place new data into classes
4. SVM: maps the predictor variables into a higher dimensional space before constructing a linear boundary for class separation

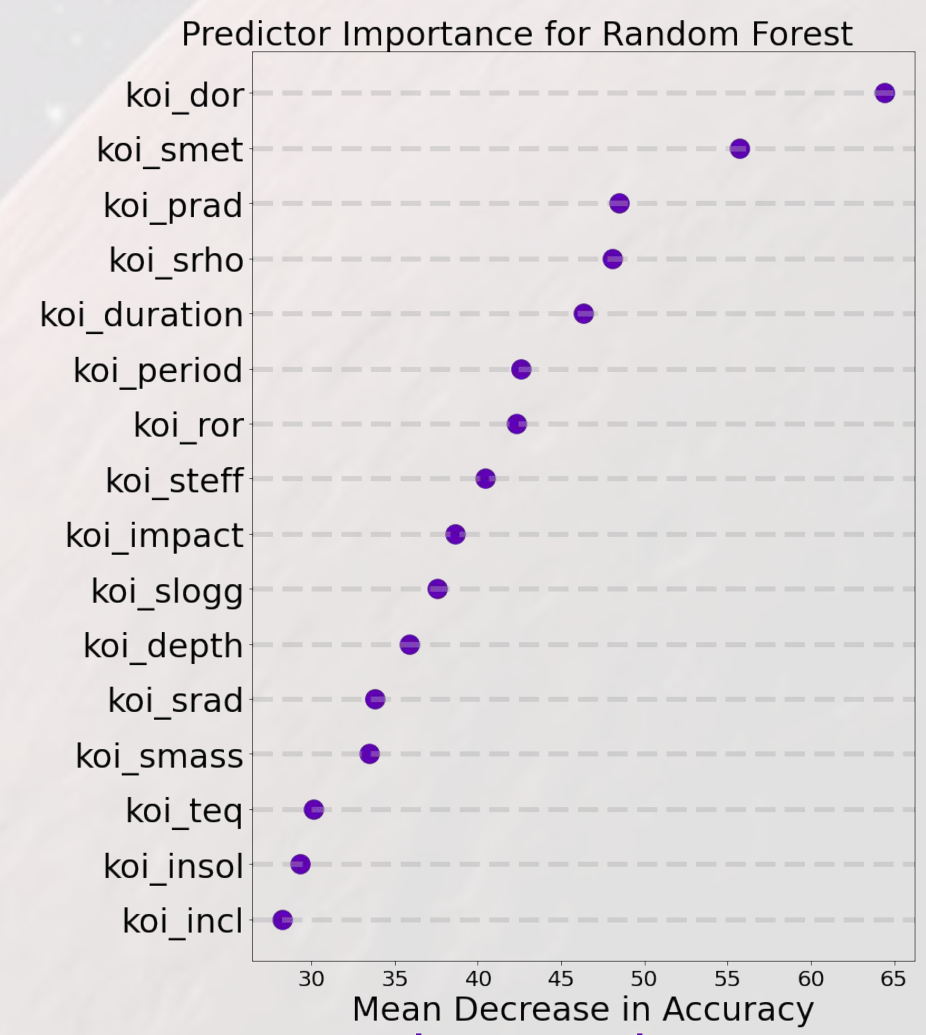
For each of these models, we output the probability that the probability that a planetary candidate is "confirmed" or "false positive." To determine the ***separation threshold***, or cutoff probability for classifying a planetary candidate as "confirmed" or "false positive," we use the Youden's J statistic, which optimizes the sensitivity and specificity metrics achieved at different separation thresholds. To compare model performances, we calculate the ***area under the curve* (AUC)**, which measures model sensitivity and specificity under various separation thresholds.

We measured the ***misclassification rate (MCR)*** - the percentage of incorrect classification. Overall, the random forest performed the best with the MCR and AUC metrics.

We also determined which predictors were the most important for our random forest model, as seen on the graph to the right.

Model	Logistic regression	Random Forest	KNN	SVM
MCR	0.143	0.097	0.241	0.133
AUC	0.929	0.968	0.830	0.865
Separation Threshold	0.613	0.605	0.500	NA





Confusion Matrix	Confirmed	False Positive
Confirmed	510	312
False Positive	146	1055

Above Top: Plot showing the effect each predictor has on the accuracy of the random forest classification.
Above Bottom: Confusion matrix
Far Left: Table showing measures of accuracy for the four classification models. Out of the four, random forest performs the best with the lowest misclassification rate and highest area under curve.
Left: The Receiver Operating Characteristics (ROC) Curve for all the models. The AUC statistic is the integral from these curves.

Conclusion

Overall, the random forest performed better than the other classification models used in this project.

From the predictor importance analysis for the random forest, it appears that all 16 of the non-zero predictors played a role in the classification. The most important variables are **koi_dor** – the distance between the planet and the star, and **koi_smet** – the ratio of iron and hydrogen at the star's surface.

These conclusions can be used to inform future techniques for identifying exoplanets. Since koi_dor and koi_smet are the most important predictors for differentiating "confirmed" and "false positive" cases, these observations can be given more scrutiny in future analyses that try to identify exoplanets.

Bibliography

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Data columns in Kepler Objects of Interest Table. (2021, February 11). NASA Exoplanet Archive. Retrieved December 2, 2021, from https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html

Freeman, P. (2021). Weeks 1-13. Retrieved from CMU Canvas site.