Using NASA's Kepler Telescope Data to Identify Exoplanets

By: Sashank Yalavarthy, Javier Abollado, Marius Nwobi, Erick D. Cohen, Akshay Gupta, Alekhya Vittalam

Background & Introduction

- Identifying exoplanets is almost impossible via direct imaging due to their (relatively) small size
- To identify exoplanets we rely on indirect methods such as radial velocity (detecting a star's *wobble*) and transit (detecting a partial eclipse of that body on another star)
- NASA's Kepler satellite observed the Cygnus constellation and identified over 100,000 possible exoplanets between 2009 - 2013

Can we construct a classification model to correctly identify whether an extrasolar object is a true exoplanet?

Data Exploration & Pre-Processing

- Data processing software was used to analyze all the light curves (i.e., the brightnesses of each star as a function of time) and identified "objects of interest," i.e., stars with possible exoplanets.

- The Kepler observations, along with observations made independently, were used to take these objects of interest and label them as **CONFIRMED** (really an exoplanet) or **FALSE POSITIVE** (not an exoplanet)

- The data were initially heavily skewed, which we remedied using appropriate transformations such as applying logarithmic and inverse logarithmic functions to certain variables based on the properties of their skewness.

Predictor Name	Description
period	The interval between consecutive planetary transits
eccen	Orbital Eccentricity: Measure of the orbit's deviation from a perfect circle
incl	Inclination: Angle between the plane of the sky and the orbital plane of the planet
dor	Planet-Star distance divided by Star Radius
impact	Impact Parameter
duration	Transit Duration
depth	Transit Depth
ror	Planet radius / stellar radius
prad	Planetary Radius
teq	Equilibrium temperature (Kelvin)
insol	Insolation flux equilibrium temperature
srho	Fitted stellar density
steff	Star photospheric temperature
slogg	Base-10 log of acceleration on the star surface due to gravity
smet	Base-10 log of Fe to H ratio on the star surface (normalized)
srad	Star photospheric radius
smass	Star mass

Methods

- **Models Used**: Logistic Regression, SVM, KNN, Random Forest, Gradient Boosting and XGBoost.

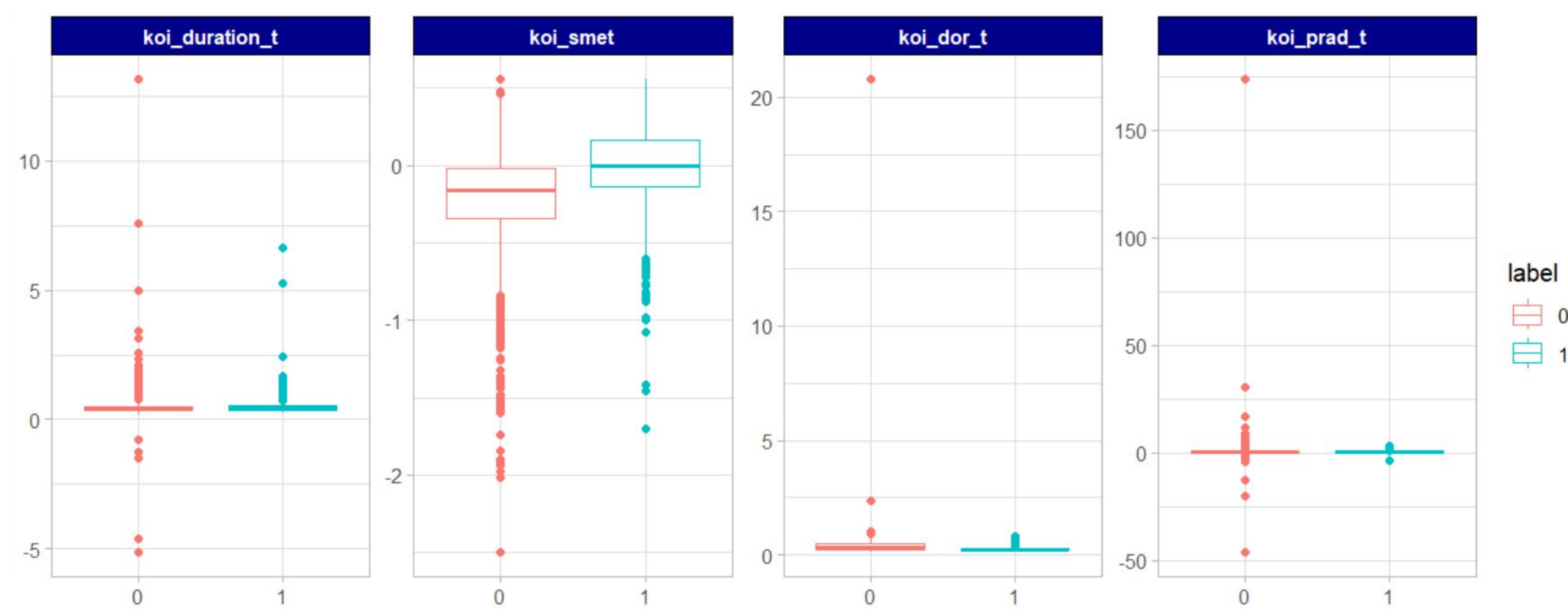
Analysis & Results

Confusion Matrix of Random Forest

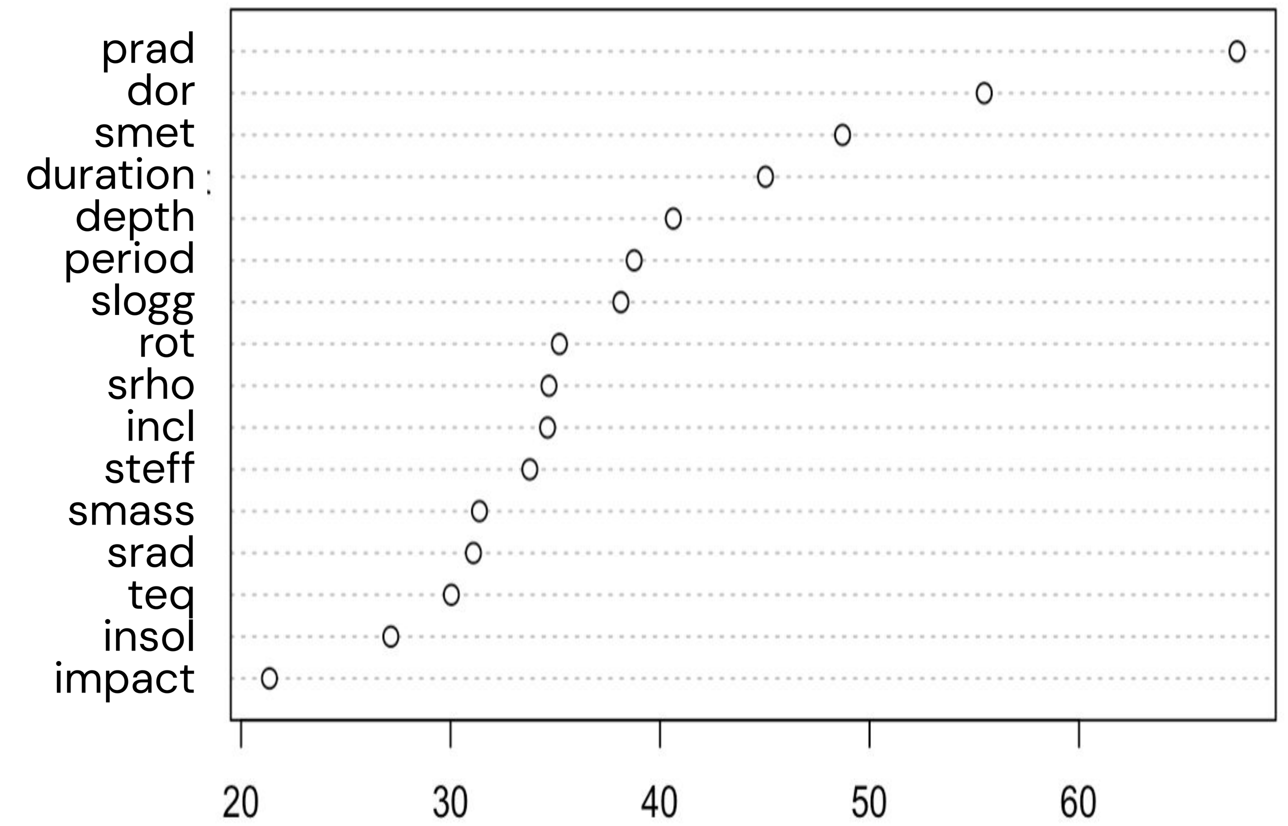
	Predicted	
Actual	CONFIRMED	FALSE POSITIVE
CONFIRMED	1271	48
FALSE POSITIVE	131	608

Table above shows the validation set performances with recall: **0.9298** and precision: **0.8243**. It was constructed by maximizing the sum of sensitivity and specificity for the random forest model.

Boxplots of Best Predictor Variables

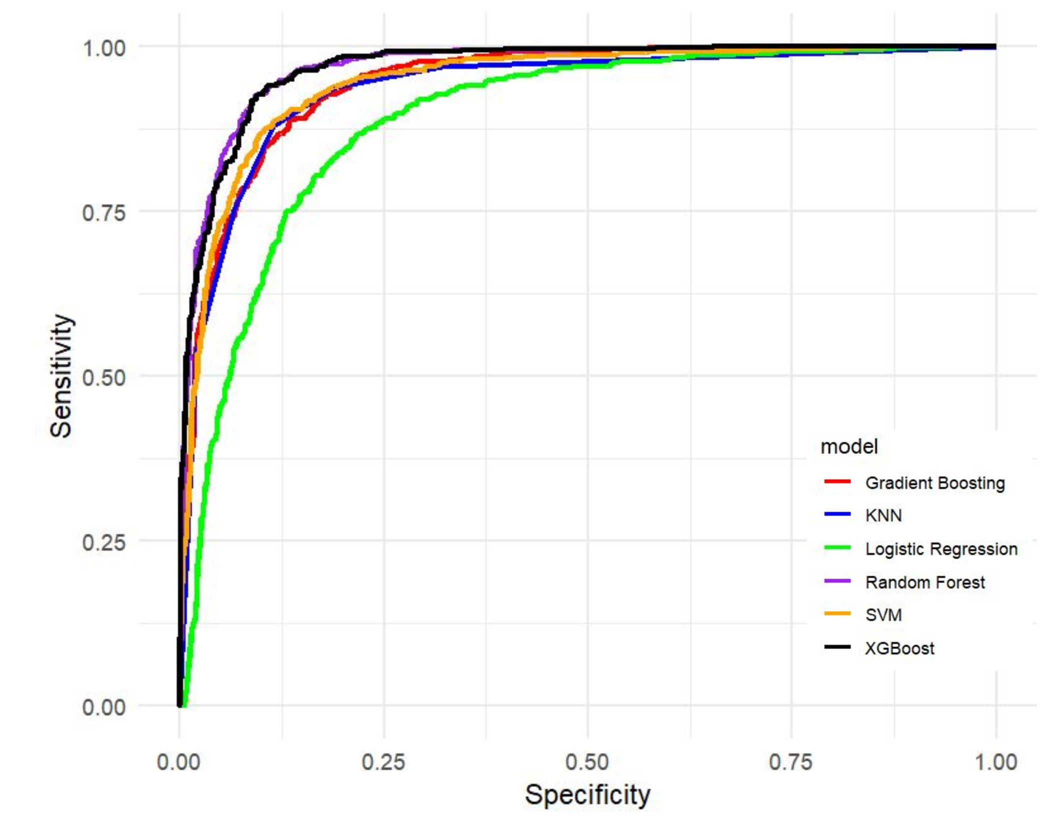


Best Predictor Variables Selected by Random Forest



The planet's radius stands out as the most significant feature in the importance plot, which aligns with expectations for exoplanet detection. Its high importance indicates that the data are particularly sensitive to small errors. In scenarios where the radius is derived from radial velocity measurements (inversely proportional to R^2), this metric must be extremely precise.

ROC Curves



	Accuracy	Precision	Recall	F1	AUC	Optimal Threshold
Logistic Regression	0.8095238	0.6506849	0.8689024	0.7441253	0.8895752	0.5755720
Support Vector Machine	0.8853256	0.7815013	0.8887195	0.8316690	0.9467420	0.3460900
K-Nearest Neighbors	0.8833819	0.7818428	0.8795732	0.8278336	0.9373559	0.5000000
Random Forest	**0.9130224**	**0.8227334**	0.9268293	**0.8716846**	**0.9691936**	0.4100000
Gradient Boosting	0.8726919	0.7551813	0.8887195	0.8165266	0.9466289	0.4428734
XGBoost	0.9086492	0.8054830	**0.9405488**	0.8677918	0.9690631	0.2957910

Random Forest was the best performing model with the highest scores amongst all metrics except for recall. XGBoost had the highest recall score, meaning it identified more of the true exoplanets at the cost of more false positives

Conclusions

- Our best model achieved an 87% F1 Score & 91% accuracy, showing how data from powerful satellite instruments can be analyzed with machine learning tools.
- Future research could consider creating models to predict continuous numerical features relating to exoplanets, such as transit duration.
- The best predictions are made by Random Forest. Linear models yielded accuracies of approximately 80%, but given the non-linear nature of the problem, tree-based models increased our performance metrics to more than 90%.

References

1. NASA Exoplanet Archive. (2021, February 11). *Kepler candidate table columns*. Retrieved December 5, 2024, from https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html
2. Erickson, B. H., & Nosanchuk, T. A. (1992). *Understanding data*. Open University Press.
3. NASA Science Mission Directorate. (n.d.). *Kepler and K2 Missions*. Retrieved December 5, 2024, from https://science.nasa.gov/mission/kepler

Temporary notes