Transiting Planet Discovery in the Kepler Pipeline Using Automated Machine Learning

Martin Mohan

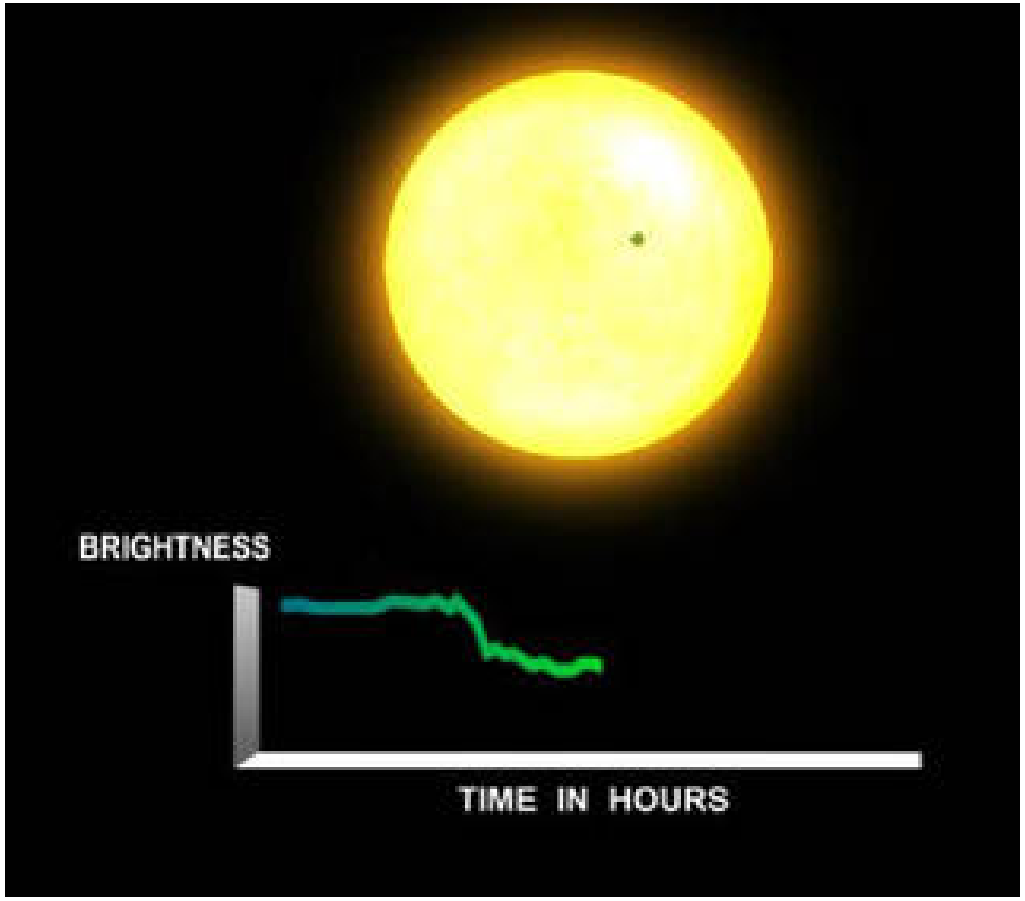
Research in Computing Presentation
MSC in Data Analytics - 2020
National College of Ireland

x18191339@student.ncirl.ie

August 17, 2020

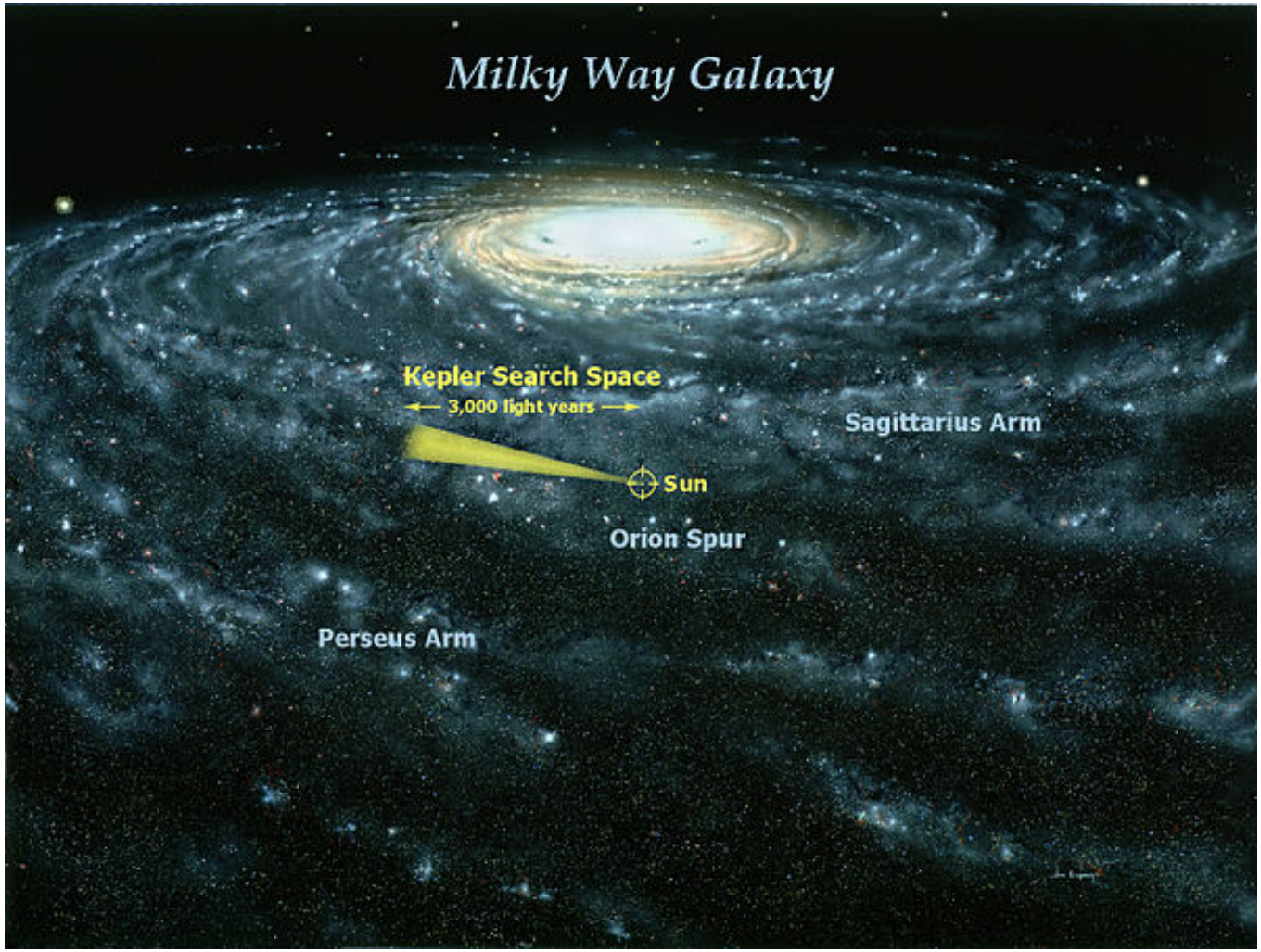
Overview

Transit Photometry: Approx 77% of all planets discovered with this method

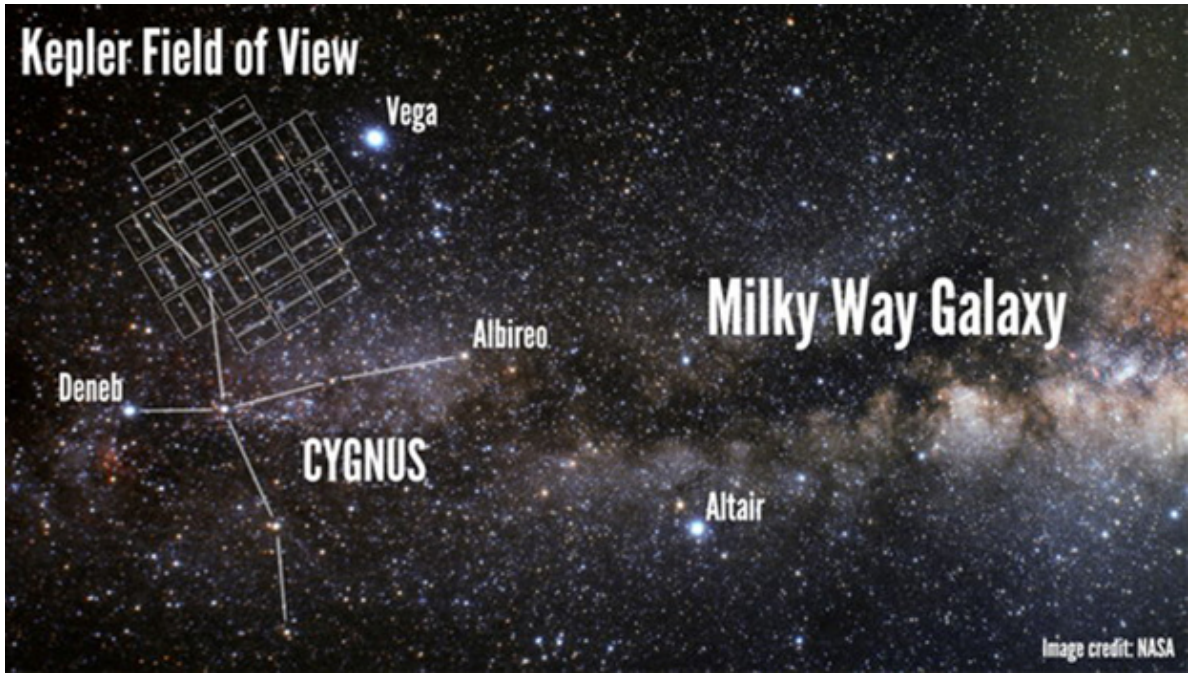


- Planet passing in front of sun causes a slight dimming
- Over 4000 planets discovered by all methods
- **Kepler** discovered over 2,300 planets using this method

Keplers Field of View (Artists Rendition)

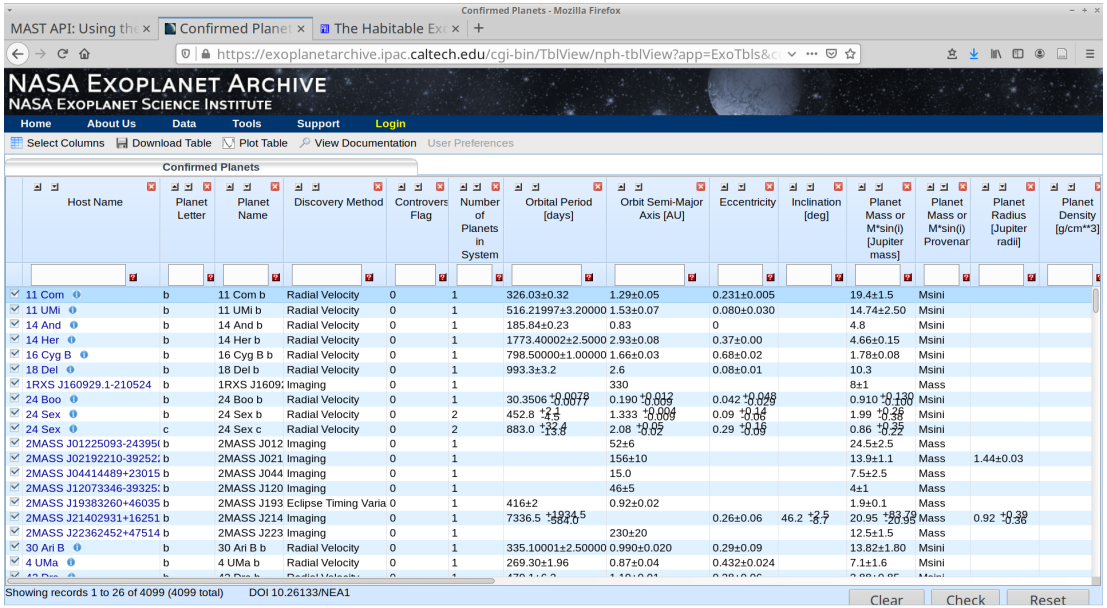


Keplers Field of View (Near)



- Approx 100 billion stars in the milky way
- Kepler surveyed 200,000 stars between Cygnus and Lyra Constellations
- Over 34,000 were identified as Transit Crossing Events (TCE)
- Over 8000 were identified as Kepler Objects of Interest (KOI)

NASA Exoplanet Archive: TCE and KOI are publicly available



- `https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=planets`
- **T**ransit **C**rossing **E**vents (**TCE**. Over 34,000 cases)
- **K**epler **O**bjects of **I**nterest (**KOI**) (CONFIRMED,CANDIDATE or FALSE POSITIVE)

Research Question

- RQ: Can an automated machine learning pipeline be used to improve exoplanet prediction using Kepler data.
- Sub-RQ: Would the methodology work on other transit missions such as K2 and TESS

Research Answer

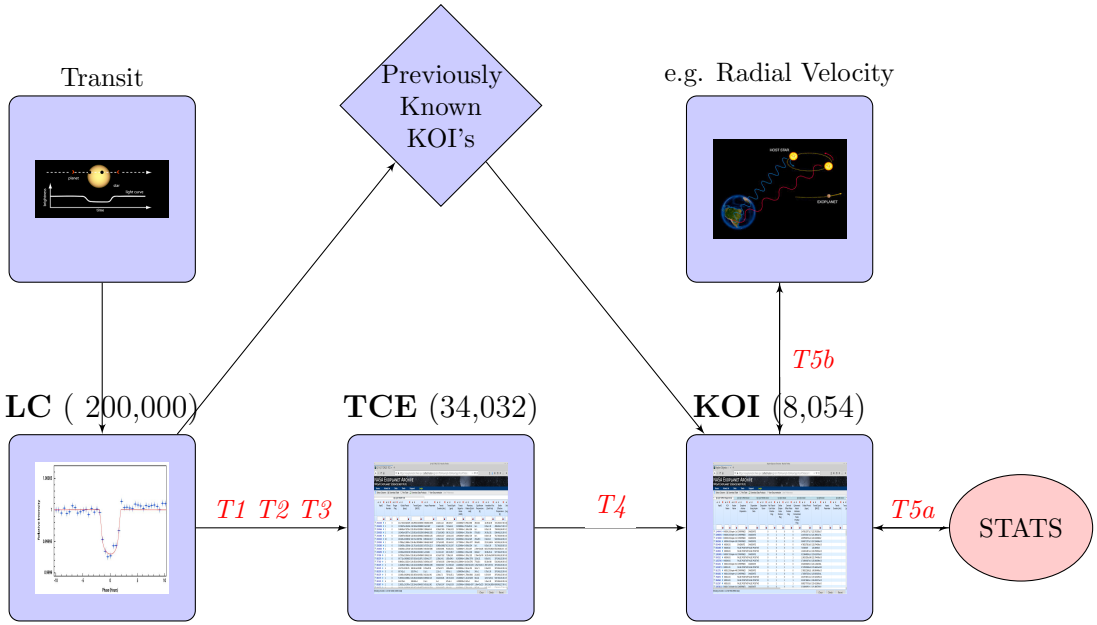
- RQ: It was possible to predict new planets with the help of automated machine learning code TPOT using the latest kepler data.
- RQ: It was possible to order planetary candidates in order of probability.
- Sub-RQ: The pipeline methodology could be applied to other projects such as K2 and TESS

Expert Systems or Machine Learning

Two types of system are used to derive the list of Planetary Candidates from Kepler's 200,000 light curves.

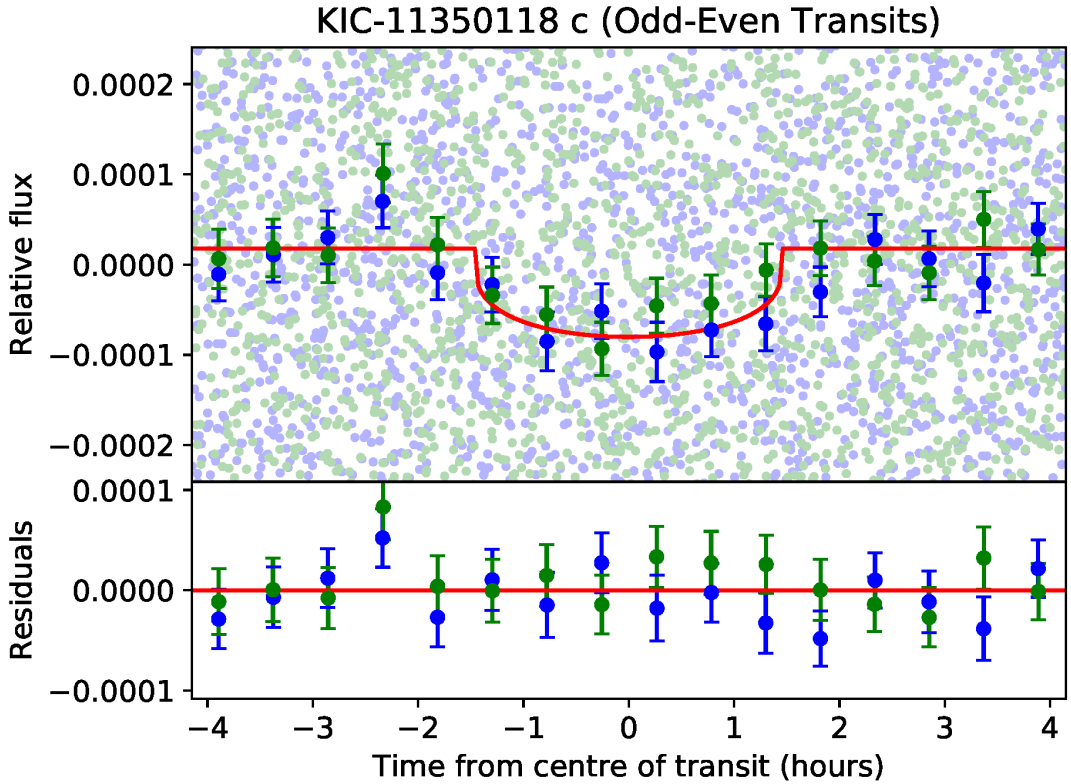
- **Expert systems** are rule based systems where the full knowledge of the expert is digitized, and is used in the decision making.
- **machine learning (ML)** are based on statistical modelling of data.

Expert System 1 (NASA) [Jenkins et al. 2015]



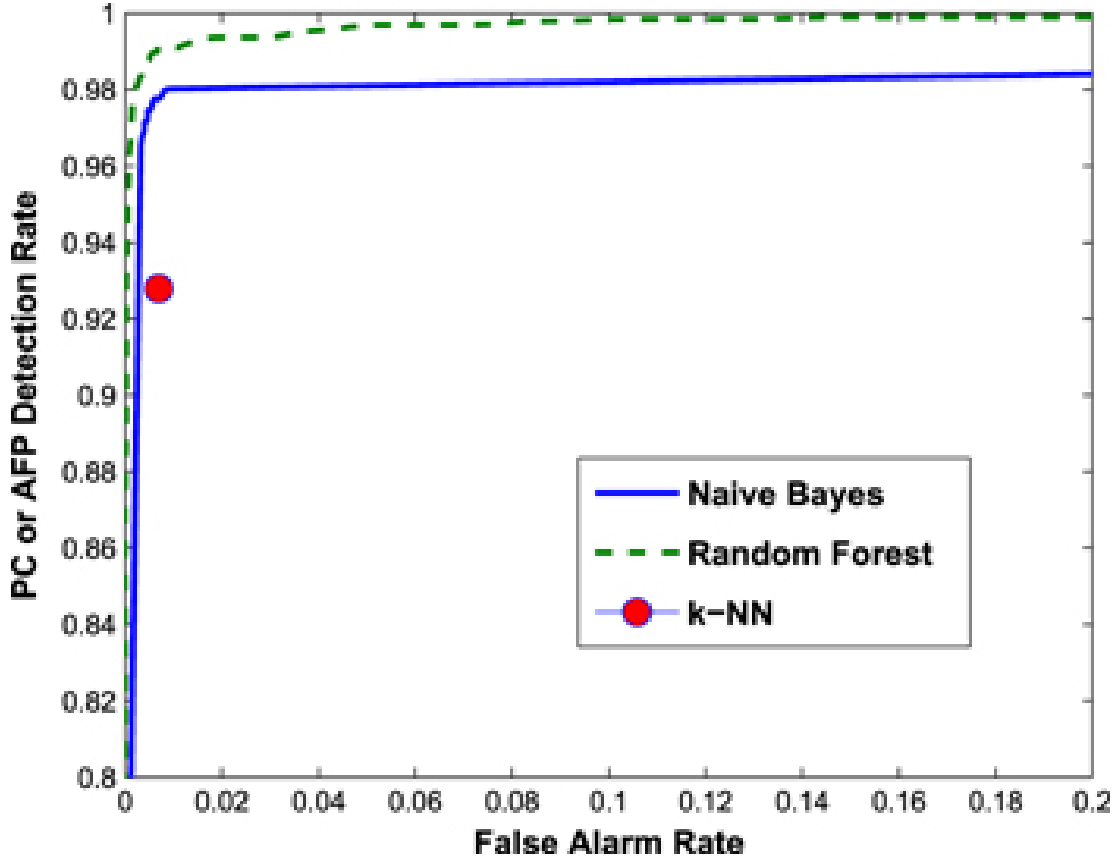
- Expert machines were used to identify Kepler Objects of Interest
- Confirmation by statistics or follow up observations

Expert System 2 (Independent researcher): [Kunimoto, Matthews, and Ngo 2020]



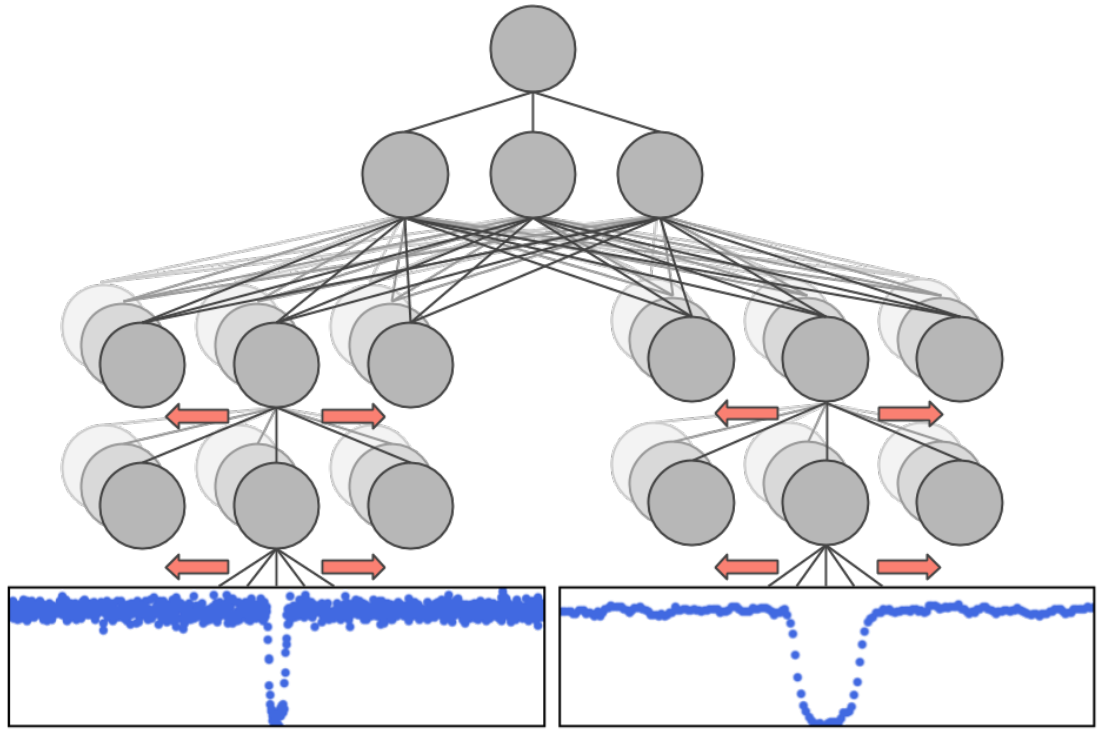
- Kunimoto used box least squares and found 17 new planets
- Entire dataset 200,000 stars searched.

Machine Learning 1 (NASA) [McCauliff et al. 2015]



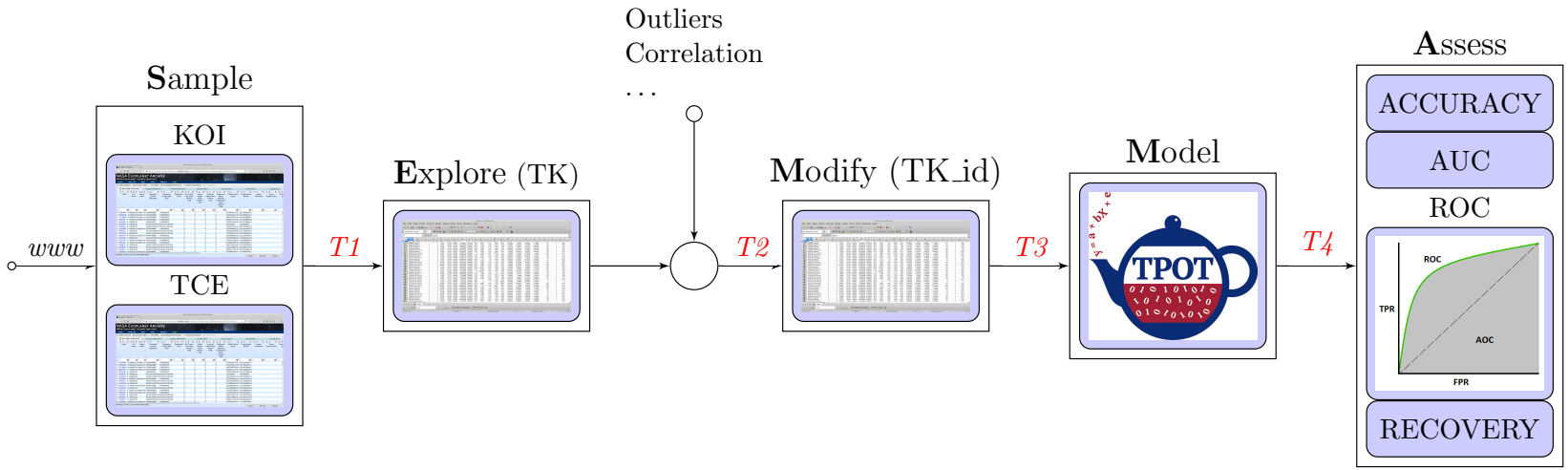
- Random Forest vs Naïve Bayes vs k-nn
- Random Forest was the most accurate.
- Older data used approx 18,000 cases. Newer **DR25** data has 34,000 cases and many corrections were performed. `https://exoplanetarchive.ipac.caltech.edu/docs/KSCI-19114-001.pdf`

Machine Learning 2 with CNN (Independent Researcher) [Shallue and Vanderburg 2018]



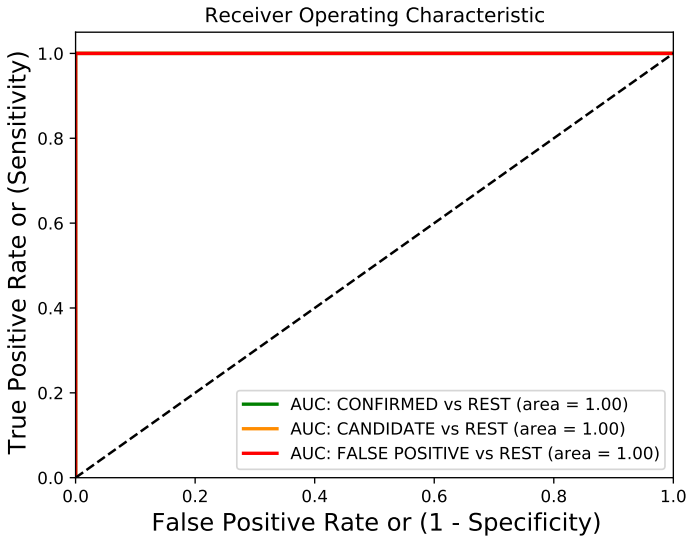
- Shallue used deep learning to discover 2 new planets.
- Only 670 multi-planetary systems were searched.

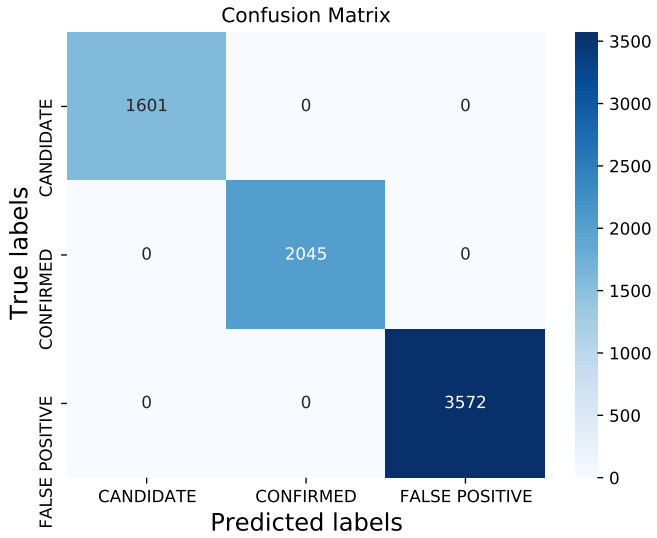
Automated Machine Learning pipeline with TPOT [Olson and Moore 2019]

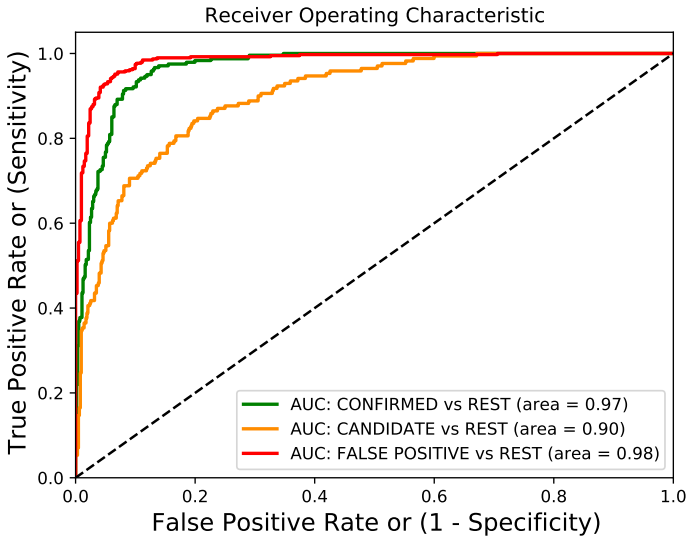


- Treatment *T1, T2, T3, T4, T5*
- Candidates are ordered by probability
- New candidates identified from TCE
- Follow up work for individual observations needed.

Performance Measurement. For Gradient Boost







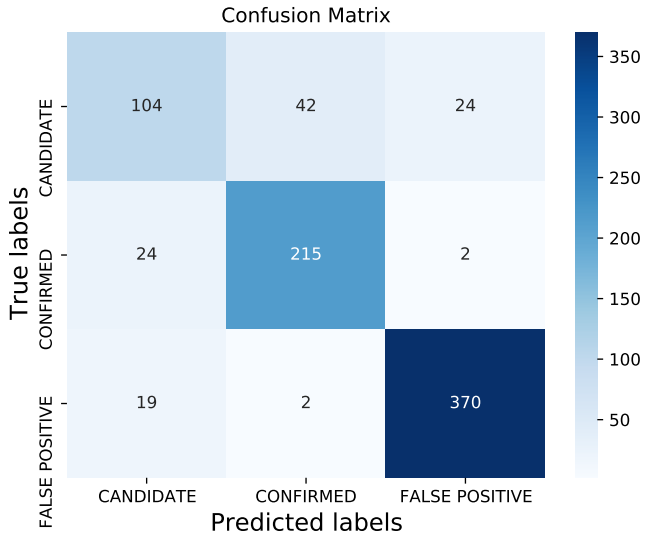
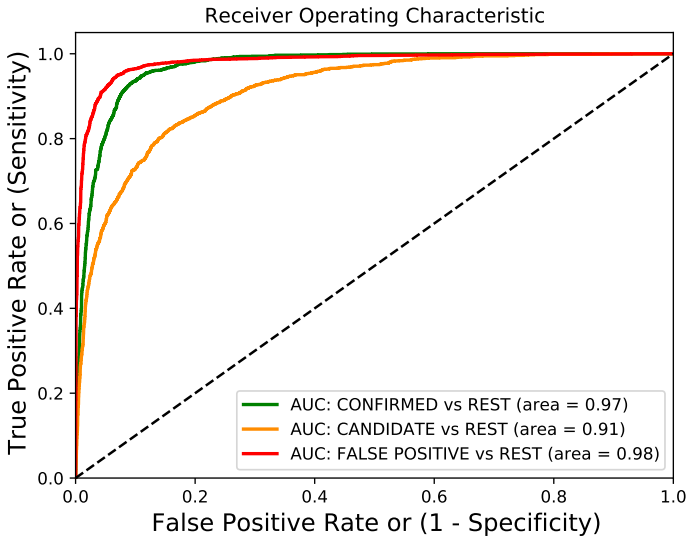
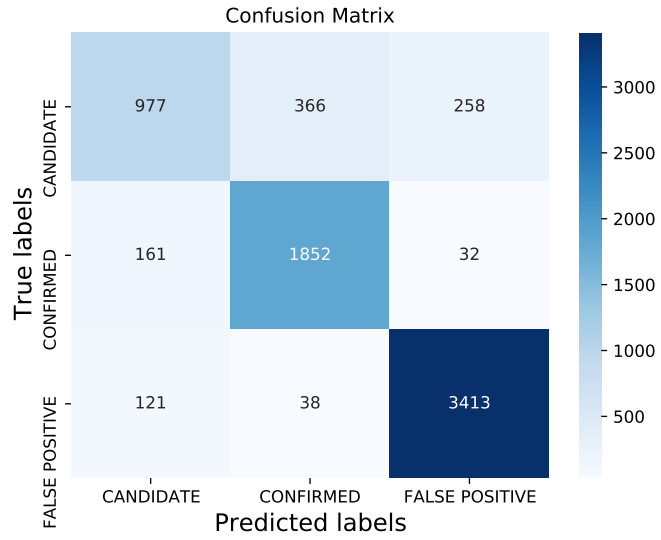
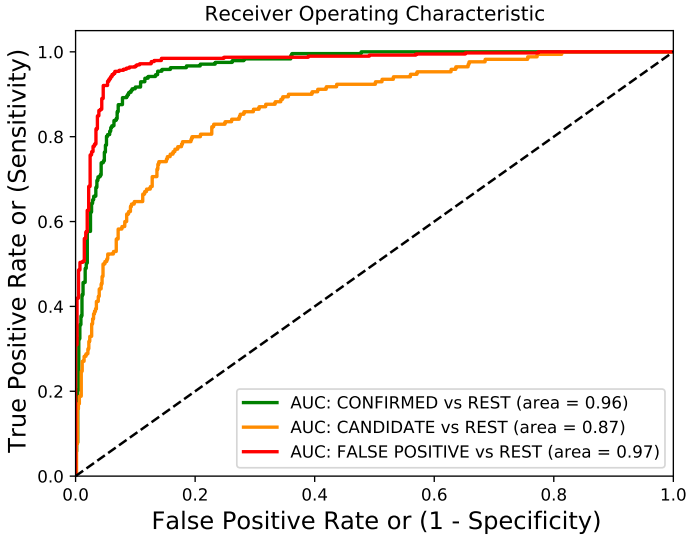


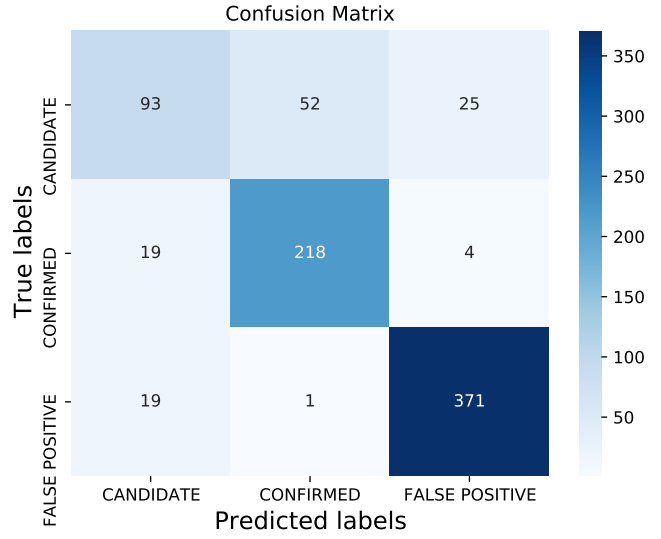
Figure: GBvif: Top Row: overfit test. Middle and bottom row test data

Performance Measurement. For Logistic Regression









Recovered Planets Overfitted

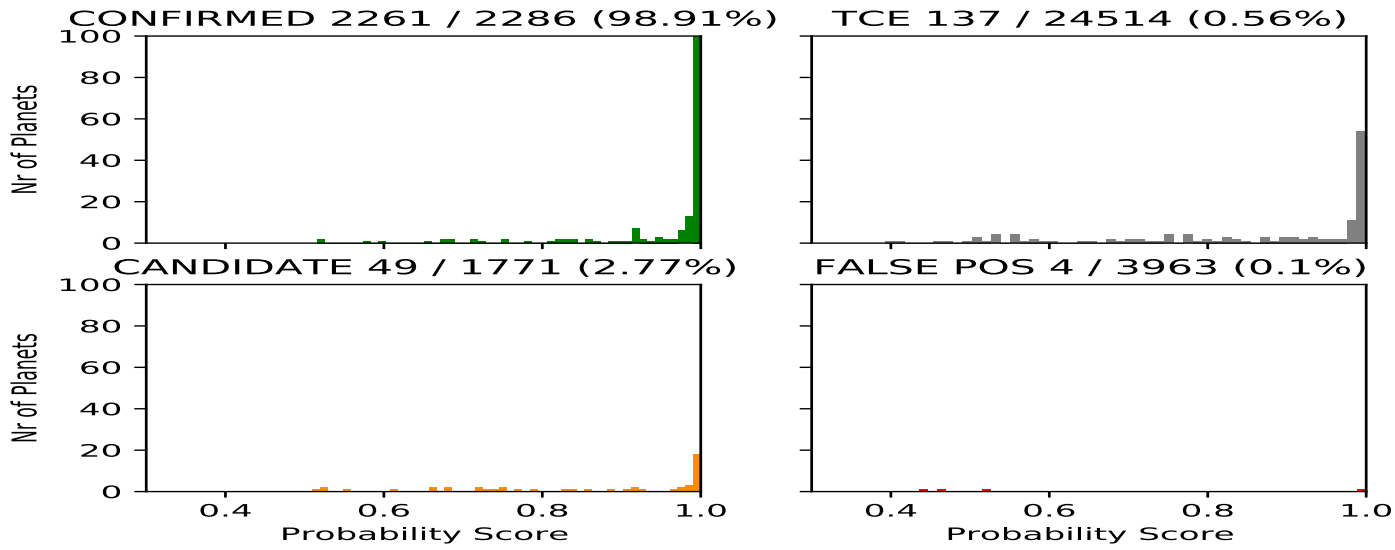


Figure: GB_vif: Planets predicted as CONFIRMED

Recovered Planets

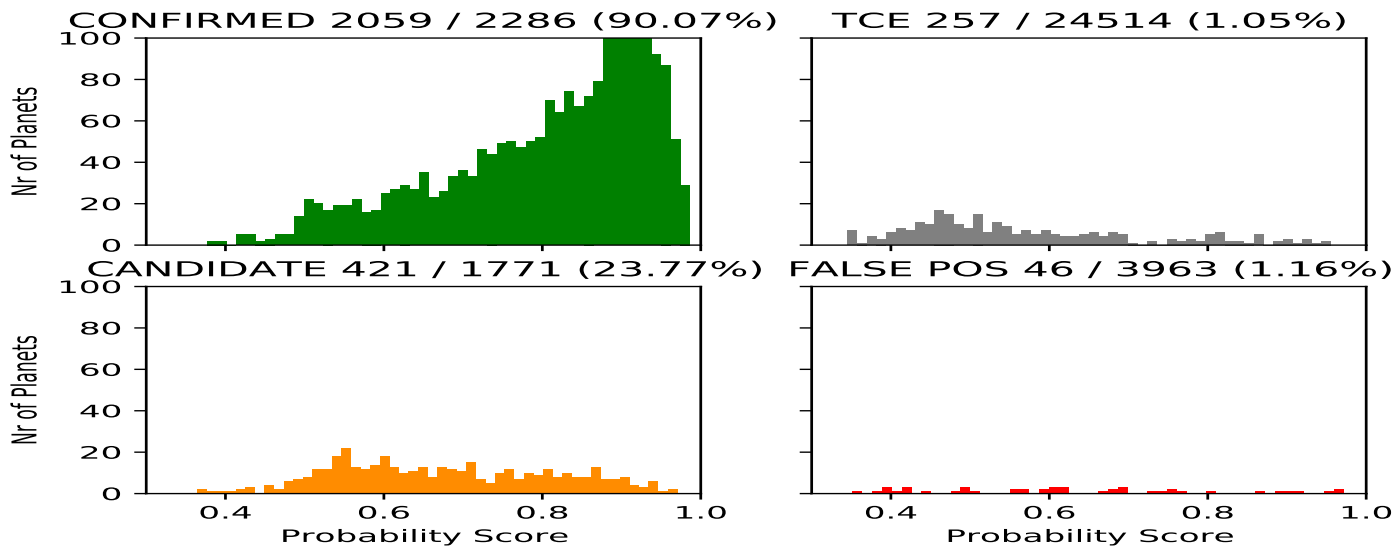


Figure: LR: Planets predicted as CONFIRMED

Result Table. Gradient Boost, Random Forest, Logistic Regression, Decision Trees - Sorted by logistic regression

Table: Top probabilities of being a confirmed planet sorted by Logistic Regression

kepid	**plnt num**	**kname**	**dispos**	**GB vif**	**RF vif**	**LR**	**DT**
8456679	2	K00102.02	FP	-	-	0.972	0.951
8480285	1	K00691.01	CAND	-	0.503	0.971	0.948
8804455	2	K02159.02	FP	-	-	0.963	0.986
10788461	1	K03925.01	CAND	-	-	0.959	0.571
8644365	1	K03384.01	CAND	-	-	0.948	0.943
8804845	1	K02039.01	CAND	-	-	0.945	0.919
3831053	1	K00388.01	CAND	-	0.642	0.944	0.966
12505076	1	K02154.01	CAND	-	-	0.944	0.971
4149450	1	K01864.01	CAND	-	-	0.943	0.992
2581316	2	**K03681.02**	CAND	1.0	0.952	0.942	0.932
5709725	2	K00555.02	CAND	-	0.516	0.937	0.919

References

Jenkins et al. (2010)

OVERVIEW OF THE KEPLER SCIENCE PROCESSING PIPELINE

The Astrophysical Journal 713, 2

D. McCauliff et al. (2015)

Automatic Classification of Transit Candidates

The Astrophysical Journal 801, 1

C. Shallue (2018)

Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90

The Astronomical Journal 155, 2

Kunimoto (2020)

Searching the Entirety of Kepler Data. I. 17 New Planet Candidates Including 1 Habitable Zone World

The Astronomical Journal 159, 3

Olson and Moore (2020)

Automated Machine Learning: Methods, Systems, Challenges, (The Springer Series on Challenges in Machine Learning)

Chapter 8:TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning

The End