GENERAL ASSEMBLY: DATA SCIENCE REMOTE

EXOPLANET HUNTING

THE PROBLEM

▸ In order to answer the question of if there's other life out there in the universe we first need to find stars that have planets orbiting them.

▸ Since there are millions of stars in the sky, but a limited amount of telescope time to investigate them all, we first start with one set of observation data, identify likely candidates for stars with planets and those candidates are then investigated further to confirm the existence of a planet.

HOW DO YOU IDENTIFY AN EXOPLANET STAR?

▸ When planets orbit a star, it passes in between the star and us, thus blocking some of the stars light. By looking for stars that regularly dim in a pattern, we can assume that they have a planet or multiple planets orbiting them.

t = 0

light

time

t = 1

light

time

t = 2

light

time
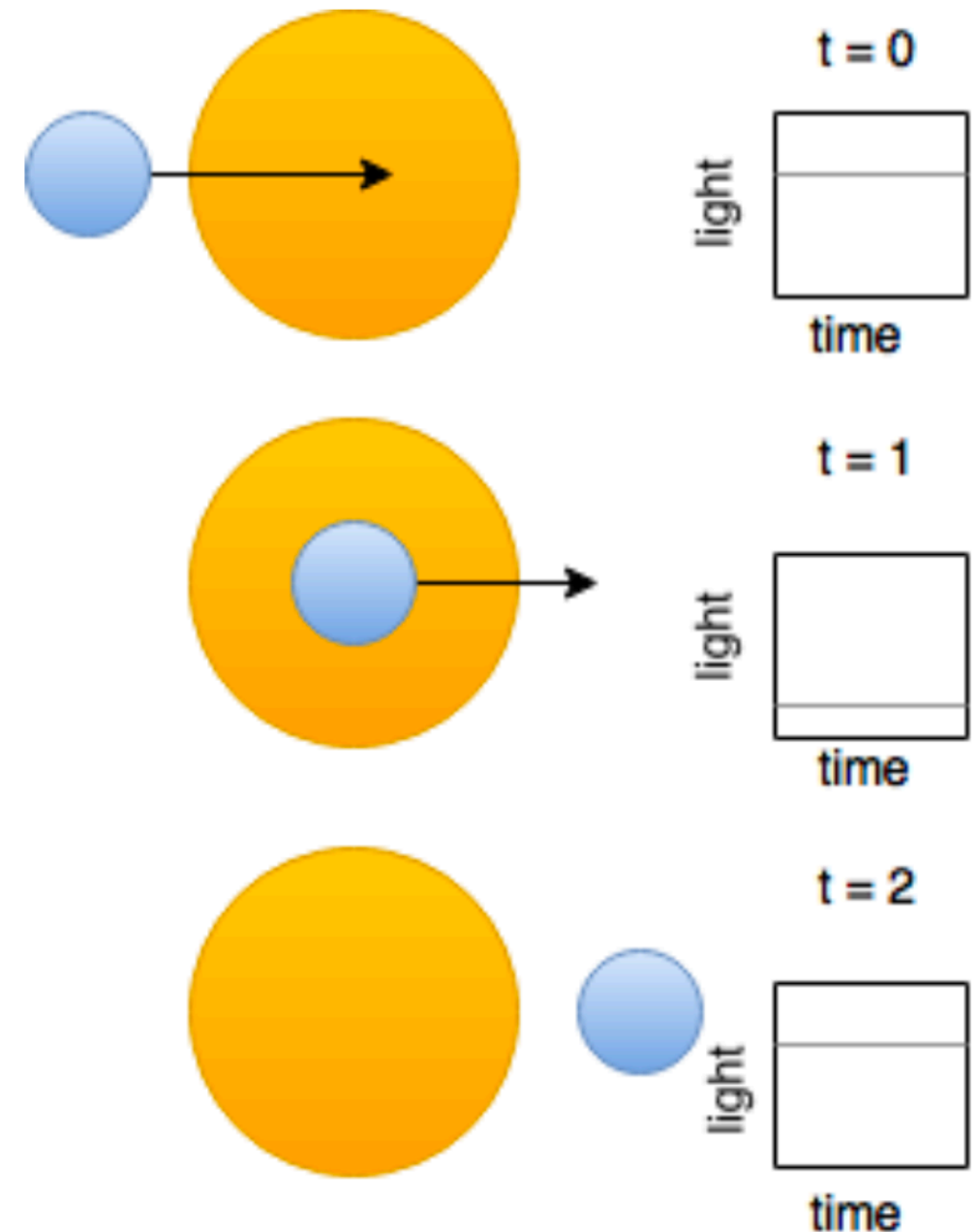
HYPOTHESIS

▸ Using data from the Kepler space telescope we can train an algorithm that can correctly identify those stars with already confirmed planets, so it can then be used to find candidates for additional investigation where planets are yet to be confirmed.

▸ As we are only looking to find stars for additional investigation, the algorithm should cast a decently wide net and not just identify those stars where planets are already confirmed, but instead look to reduce the dataset by at least 85% focused on all stars showing anomalies in light intensity.

▸ This reduces the number of stars needing additional investigation (aka telescope time) significantly.

THE DATA

▸ The data source contains both a Train and Test data set.
The Train set has 5,087 rows/observations with 37
confirmed exoplanet-stars and the Test set has 570 rows/
observations with 5 confirmed exoplanet-stars. This data is
cleaned and derived from observations made by the
NASA Kepler space telescope.

	LABEL	FLUX.1	FLUX.2	FLUX.3	FLUX.4	FLUX.5	FLUX.6	FLUX.7	FLUX.8	FLUX.9	...	FLUX.3188	FLUX.3189	FLUX.3190	FLUX.3191	FLUX.3192
0	2	93.85	83.81	20.10	-26.98	-39.56	-124.71	-135.18	-96.27	-79.89	...	-78.07	-102.15	-102.15	25.13	48.57
1	2	-38.88	-33.83	-58.54	-40.09	-79.31	-72.81	-86.55	-85.33	-83.97	...	-3.28	-32.21	-32.21	-24.89	-4.86
2	2	532.64	535.92	513.73	496.92	456.45	466.00	464.50	486.39	436.56	...	-71.69	13.31	13.31	-29.89	-20.88
3	2	326.52	347.39	302.35	298.13	317.74	312.70	322.33	311.31	312.42	...	5.71	-3.73	-3.73	30.05	20.03
4	2	-1107.21	-1112.59	-1118.95	-1095.10	-1057.55	-1034.48	-998.34	-1022.71	-989.57	...	-594.37	-401.66	-401.66	-357.24	-443.76

5 rows × 3198 columns

APPROACH

▸ In order to more effectively train my models, I limited the size of my training set so that the models would have more exoplanet stars to investigate versus those without (by percentage)

▸ Tried out a number of different models, including:

  ▸ Logistic Regression

  ▸ K-Nearest Neighbors

  ▸ Decision Tree Classifier

  ▸ Random Forest

SCORING THE MODELS

▸ There are many ways to score a model and I looked at both RMSE and accuracy scores

▸ Most important to this particular problem was to ensure all stars with possible planets are captured for further investigation, so looking at the confusion matrix to see how many False Negatives resulted was key in picking the model

	Predicted: NO	Predicted: YES
Actual: NO	True Negative	False Positive
Actual: YES	False Negative	True Positive

RESULTS

▸ Logistic Regression

    ▸ Accuracy Score: 68.4%

    ▸ Confusion Matrix: ([[23, 9],

                        [3, 3]])

▸ K-Nearest Neighbors

    ▸ Accuracy Score: 80.4%

    ▸ Confusion Matrix: ([[34, 0],

                        [9, 3]])

▸ Decision Tree

    ▸ Accuracy Score: 89.1%

    ▸ Confusion Matrix: ([[33, 1],

                        [4, 8]])

▸ Random Forest

    ▸ Mean Accuracy Score: 90.9%

CURRENT STATISTICS ON THE HUNT