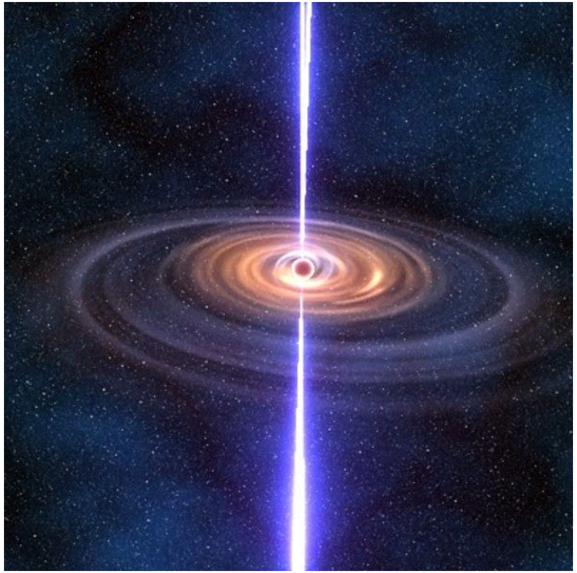
Searching for Pulsars

How Machine Learning Impacts Astronomy

By CJ Johanson

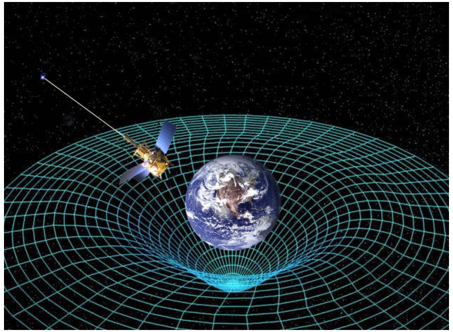
What is a pulsar star?

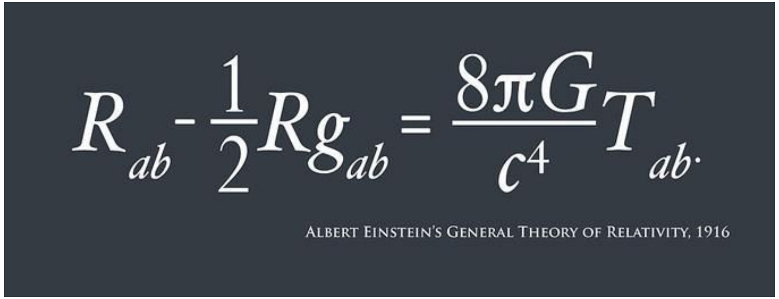
- Type of neutron star that spins rapidly, produces radio emissions in the form of beams which are ejected from its magnetic poles
- Second most dense object in universe, next to black holes.
- 1 tablespoon of neutron star material weighs > 1,000,000,000 tons (more than Mt. Everest)



What are the benefits of finding pulsars?

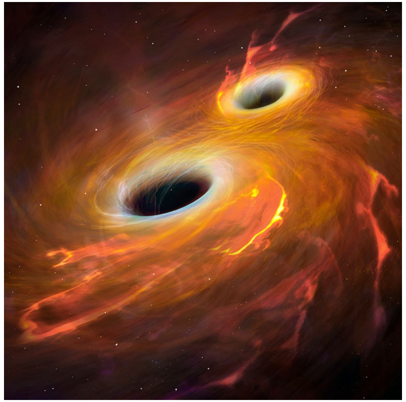
- Studying pulsars has implications for many different aspects of astronomy
- Pulsars play a role in studying gravitational waves that were generated by supermassive black holes merging in the early stages of the universe
- Pulsars also aid in the continued study of Einstein’s theory of general relativity





$$R_{ab} - \frac{1}{2} R g_{ab} = \frac{8\pi G}{c^4} T_{ab}.$$

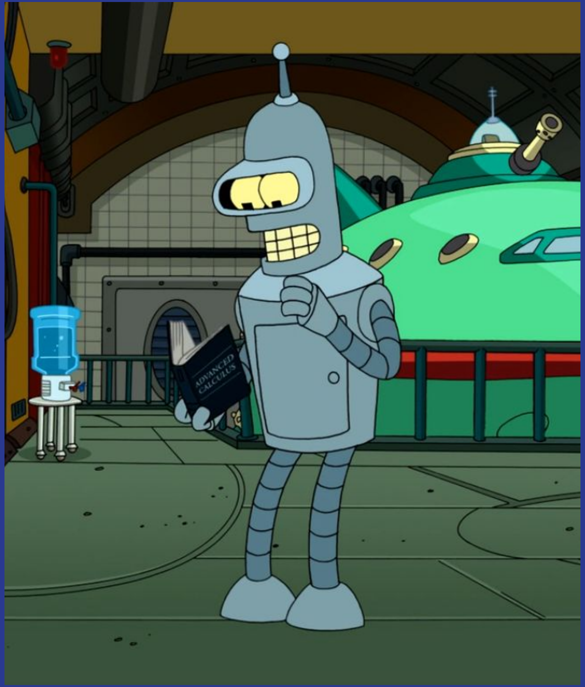
Albert Einstein’s General Theory of Relativity, 1916



...so, how do you find them?

- Radio telescopes play a critical role in gathering the necessary astronomical data used to find pulsar stars
- Contrary to common belief, space isn't exactly "empty"
- Radio telescopes collect data, but included in that data are problems like "background noise," making it difficult to distinguish whether or not an observation is a pulsar, or not.
- You then need an effective method of sifting through an enormous amount of data (too much for humans to work with!)

Machine Learning

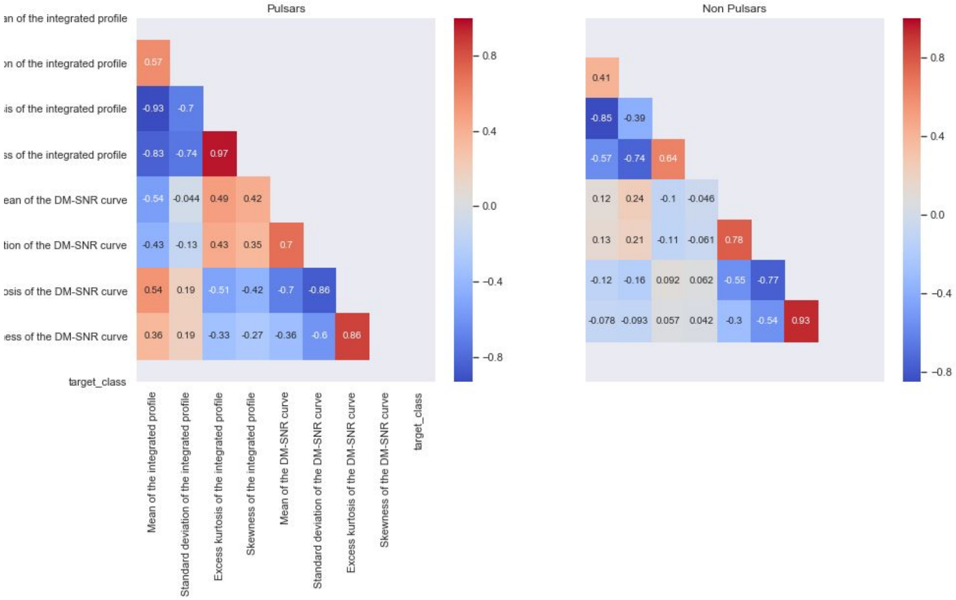


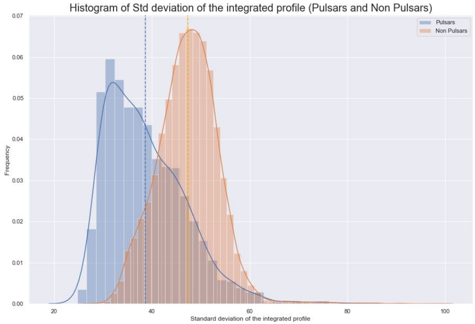
Machine learning for pulsar detection

- The goal is to look at a large data set that consists of various characteristics of individual astronomical observations (via radio astronomy)
- Observations have two possible outcomes: an observation is a pulsar star or an observation is <u>not</u> a pulsar star
- In machine learning terms, this is a classification problem, more specifically a binary classification problem

Machine learning process: first step

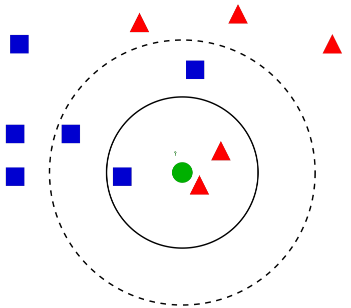
- The first step in any analysis is exploratory data analysis
- Get an idea of what the data looks like, such as:
    - Shape of the dataset
    - Distributions of the variables
    - Independent and dependent variables
    - Ideas about cleaning the data

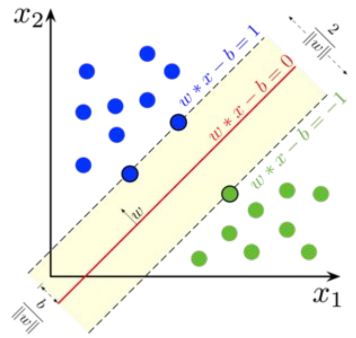


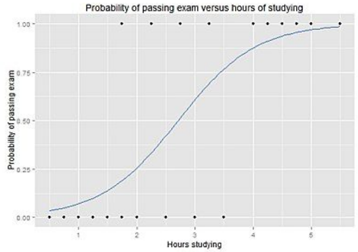


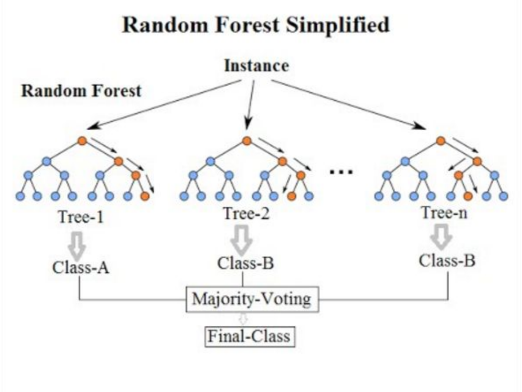
Machine learning process: model selection

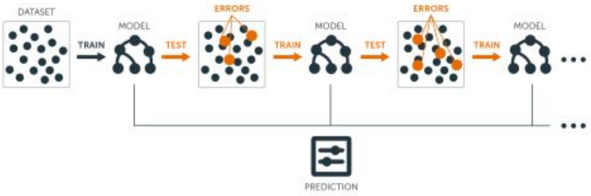
- There are several different classification algorithms to choose from for machine learning
- The following algorithms were chosen and implemented with the Python library Scikit-learn
  - K-Nearest Neighbors
  - Support Vector Machine
  - Logistic Regression
  - Random Forest
  - Gradient Boosting Decision Trees











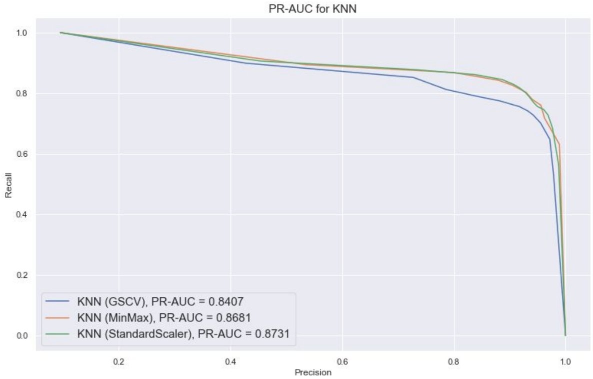
Machine learning process: training and testing

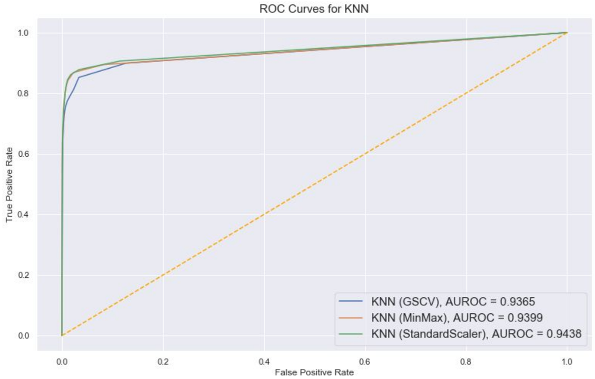
- Once the initial choices for models have been made, the data is split between a training set and a testing set (normally 75/25 split)
- Some models perform better when they data is preprocessed, so the training and test features (the dependent variables) are preprocessed
  - For example, in this analysis K-Nearest Neighbors, Support Vector Machine, and Logistic Regression used preprocessed data, utilizing with two different scaling methods (MinMax and Standard scaler)

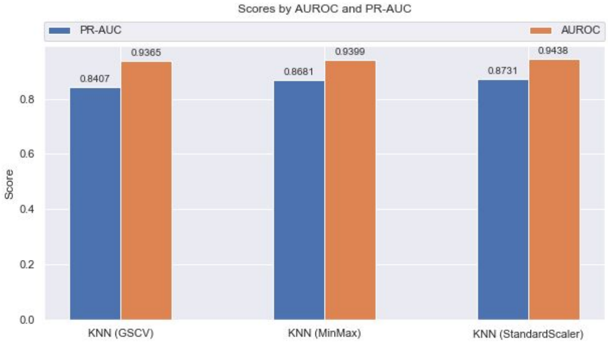
Machine learning process: evaluation

- The machine learning models in this analysis were area under ROC curve (AUROC), area under Precision-Recall curve (PR-AUC), ROC-AUC score, and average precision scored
- For each type of classifier/algorithm, the scores were calculated and each individual classifier/algorithm was compared to others in same group
- Then, each the top performer from each classifier/algorithm was compared to the top performers for all groups

Example: K-Nearest Neighbors

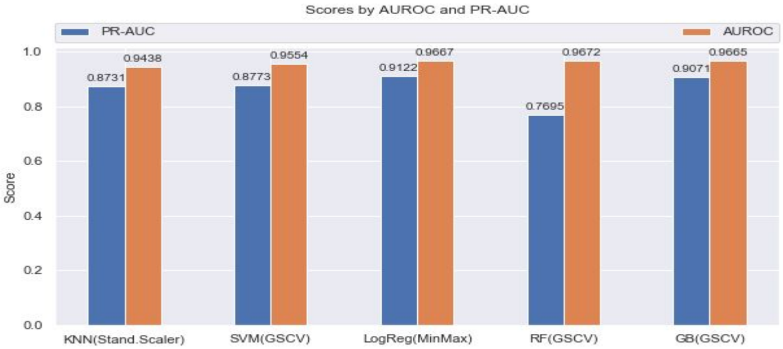


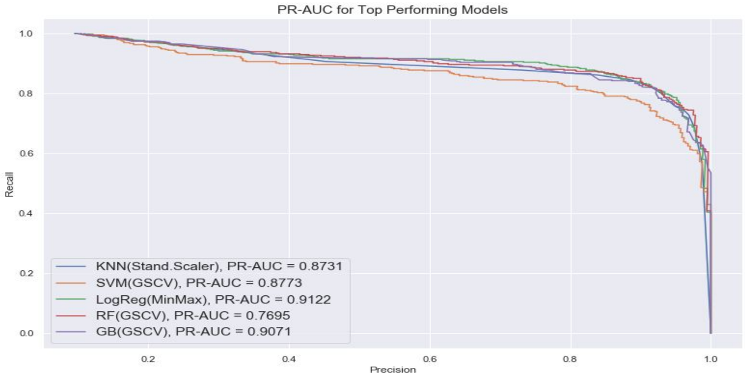


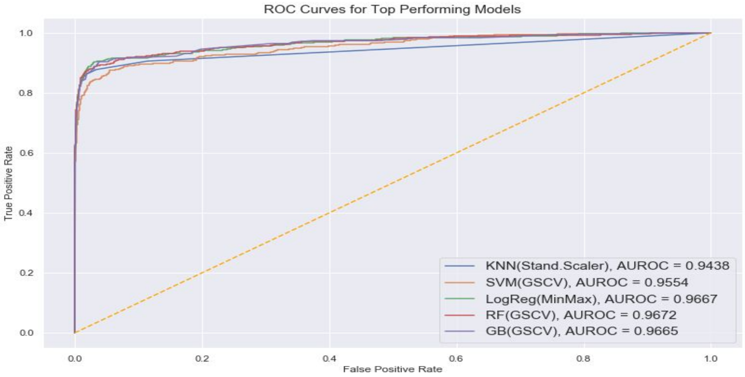


Results: final comparison of top performers

- The machine learning analysis concluded that the logistic regression algorithm was the best performer
    - ○ Gradient Boosting was the second best







Thank you!