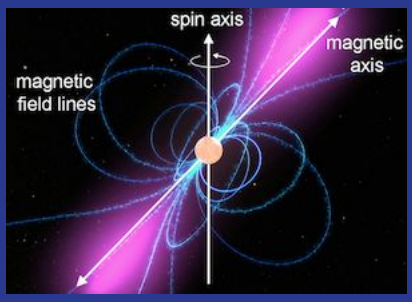
Classifying Radio Signals To Identify Pulsars HTRU2 Dataset

By George Bennett



Classifying radio signals to Identify pulsars. HTRU2 Dataset. By George Bennett.

Image source left :
https://imagine.gsfc.nasa.gov/science/objects/neutron_stars1.html

Image source right:
http://www.astronomy.com/news/2018/03/all-galaxies-rotate-once-every-billion-years

Problem Statement

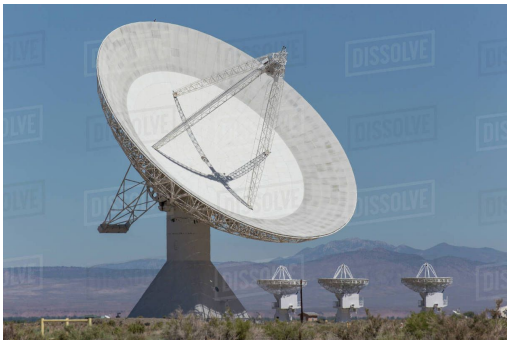
- Classification
- Increase speed and efficiency

- My task is a classification task to distinguish radio signals that are from pulsar stars from radio signals that are not from pulsar stars using machine learning, thus increasing speed of classification and reducing resources needed to perform classification.

Image source:
https://phys.org/news/2018-10-aussie-telescope-mysterious-fast-radio.html

Business value

- Efficient classification.
- More time and resources.



- The machine learning model will increase classification speed thus reducing the need for human classifiers.
- More time and resources can be put into things other than classifying pulsars.

Image source left:
https://dissolve.com/stock-photo/Satellite-dishes-field-near-mountain-royalty-free-image/101-D145-244-773

Methodology

- Several supervised machine learning models
- GridSearchCV
- PCA
- SMOTE

- I created several supervised machine learning models to classify the radio signals. First I used a decision tree. Next I used random forest, and finally I used XGBoost.
- I used GridsearchCV to tune the hyper parameters of the models.
- During the analysis I also used Principal component analysis to reduce the dimensionality of the data and try to see if it made for a better modeling.
- Since the dataset was imbalanced 9:1 I used SMOTE to upsample the data. I did this to lessen the bias and see if it made for better modeling.

Image Source: https://en.wikipedia.org/wiki/Scikit-learn

Findings

- Random Forest
- Precision %93
- F1-score %88
- Accuracy %98

The random forest model without pca or smote proved to be the most accurate.

- Precision %93
- F1-score %88
- Accuracy %98

I chose Precision and f1-score as my metrics of success because I wanted to avoid false positives and false negatives, but I believe false positives are more detrimental to the astronomers then false negatives.

Image source:
https://www.slideshare.net/palinanalytics/random-forest-classifier-in-machine-learning-palin-analytics

Business Recommendations

- The classification model performs well so It will be a good idea to automate the classification process using Machine Learning models instead of previous methods which rely much more on human labor.
- Increase the size of surveys because the data obtained can be processed much more quickly.
- Hire more people to study pulsar behavior because the amount of known pulsars will increase proportionately to survey size.

Future Work

- Stars
- Galaxies
- Other astronomical phenomena

Data Science can be used by astronomers in many ways. Given more time and data I could classify all types of stars, galaxies and other astronomical phenomena using my skills as a data scientist.

Image source:
https://www.businessinsider.com/hubble-telescope-galaxies-photo-legacy-wide-field-deep-universe-2019-5

Thank You