Classifying pulsars from radio signals HTRU2 Dataset.

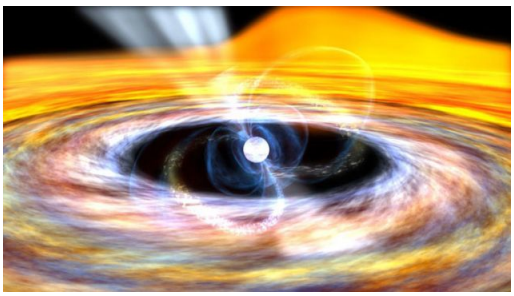
By George Bennett

Classifying pulsars from radio signals, HTRU2 Dataset. By George Bennett.

Image source right:
(http://www.astronomy.com/news/2018/03/all-galaxies-rotate-once-every-billion-years)
Image source left:
(https://dissolve.com/stock-photo/Satellite-dishes-field-near-mountain-royalty-free-image/101-D145-244-773)

Problem Statement

My task is a classification task to distinguish radio signals that are from pulsar stars from radio signals that are not from pulsar stars.

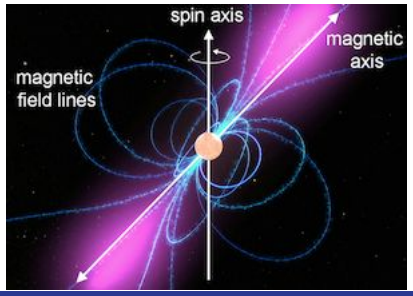


My task is a classification task to distinguish radio signals that are from pulsar stars from radio signals that are not from pulsar stars.

Image source: (https://physicsworld.com/a/magnetic-fields-put-the-brakes-on-millisecond-pulsars/)

Scientific Value

Scientists study pulsars for a variety of reasons. They are used to study space-time, the interstellar medium, and exotic states of matter. (Source: https://www.kaggle.com/pavanraj159/predicting-a-pulsar-star)



Scientists study pulsars for a variety of reasons. They are used to study space-time, the interstellar medium, exotic states of matter. (Source: https://www.kaggle.com/pavanraj159/predicting-a-pulsar-star)

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

Methodology

- I created several 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.

- I created several 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.

Performance

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

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

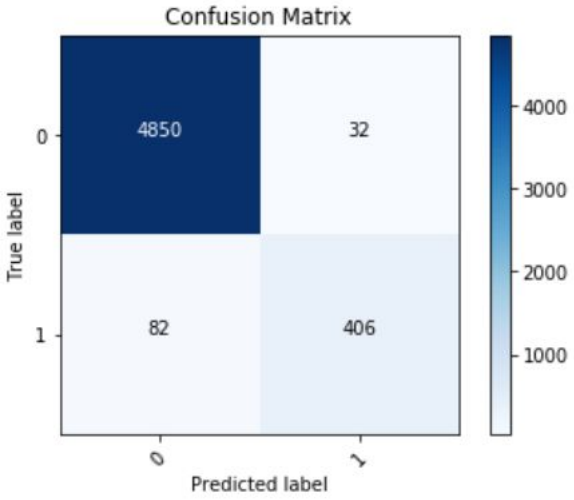
I chose Precision and f1-score as my metrics of success because I wanted to avoid false positives and false negatives, but I believe false positives are more detrimental to the astronomers then false negatives.

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.

Random forest confusion matrix



Here is a confusion matrix for the random forest model.

- 4850 true negatives
- 406 true positives
- 32 false positives
- 82 false negatives

Future Work

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.

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)