Pulsar Star Classification

Machine Learning

Cain Farnam

University of Central Arkansas

Outline

1. Introduction
2. Data
3. Random Forest Classifier
4. Model Selection
5. Results
6. Conclusion

Section 1

Introduction

Introduction

- Pulsars are a rare type of Neutron star that produce radio emissions detectable here on Earth.
- They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter.
- Each pulsar produces a slightly different emission pattern, which varies slightly with each rotation.
- A potential signal detection is averaged over many rotations of the pulsar, as determined by the length of an observation.
- In practice, almost all detections are caused by radio frequency interference (RFI) and noise, making legitimate signals hard to find.

Section 2

Data

Data Features

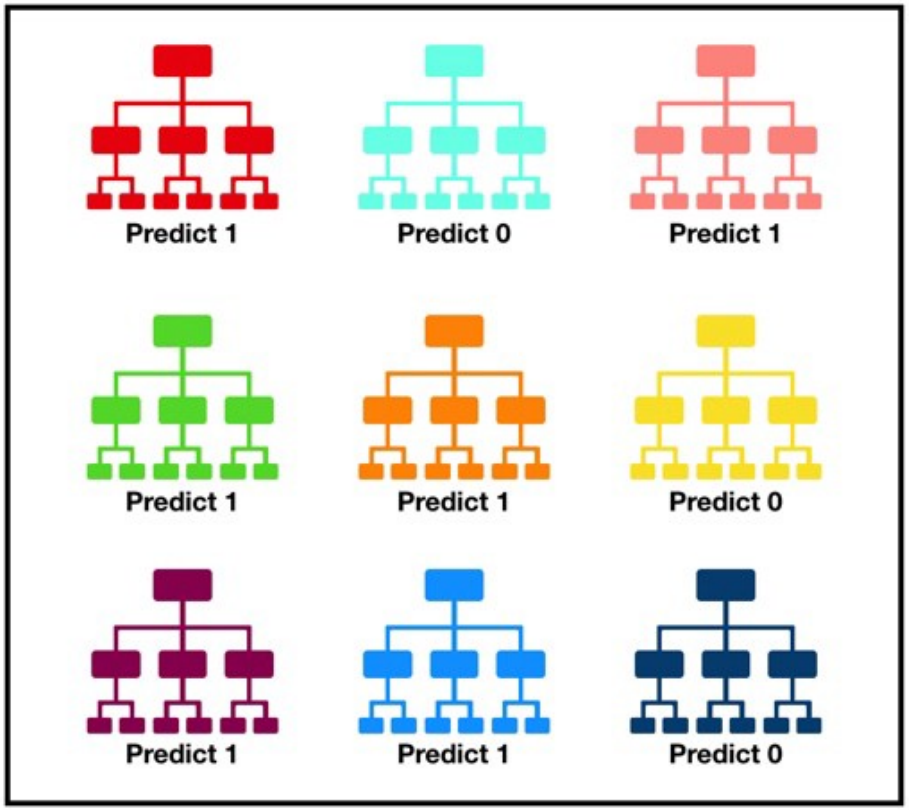
- HTRU2 is a data set which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey.
- The data set contains 16,259 spurious examples caused by RFI/noise, and 1,639 real pulsar examples.
- The features in the data set include:
  - Mean of the integrated profile and DM-SNR curve
  - Standard deviation of the integrated profile and DM-SNR curve
  - Excess kurtosis of the integrated profile and DM-SNR curve
  - Skewness of the integrated profile and DM-SNR curve

Section 3

Random Forest Classifier

Understanding Random Forest Classifier

- Random forests consist of a large number of individual decision trees that operate as an ensemble.



Tally: Six 1s and Three 0s

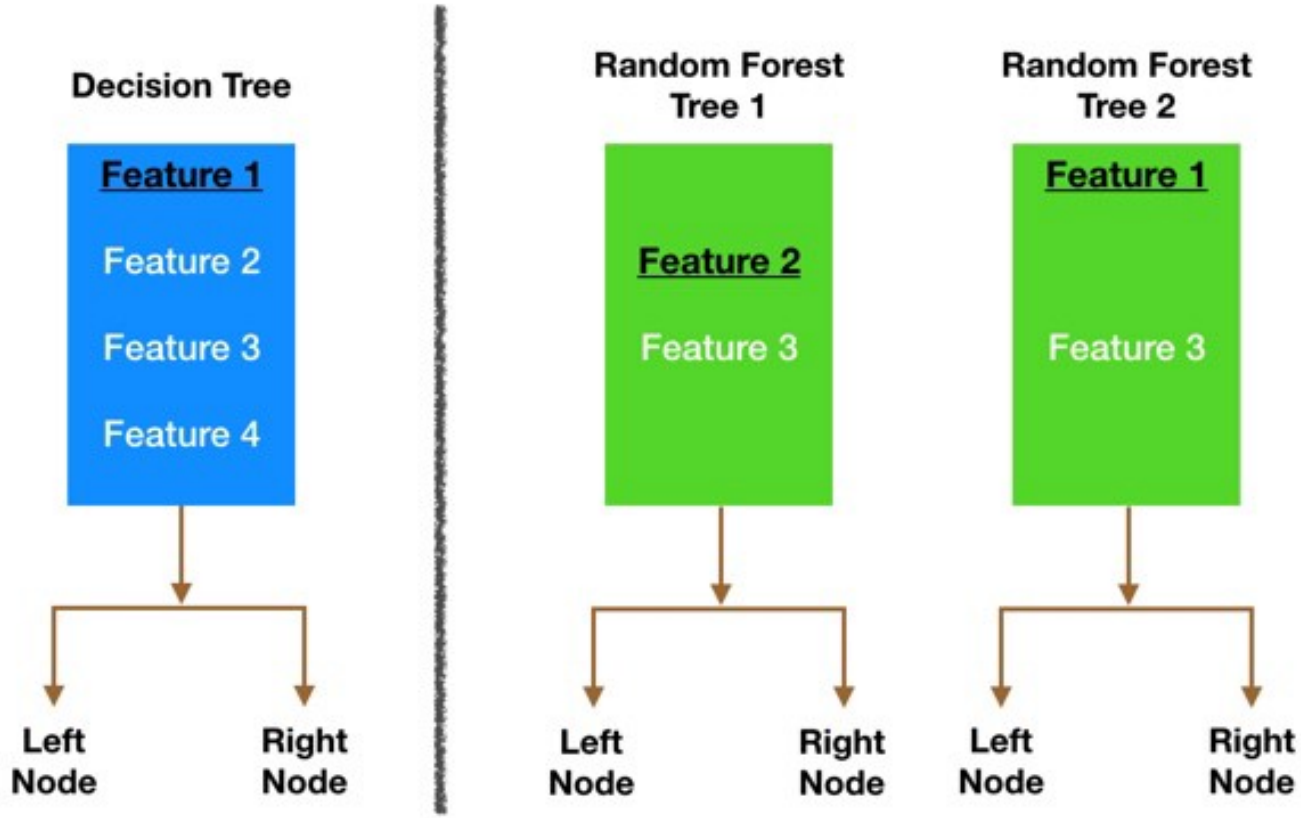
Prediction: 1

Building the Forest

- A large number of relatively uncorrelated models (decision trees) operating as a committee will outperform any of the individual constituent models.
- Decisions trees are very sensitive to the data they are trained on — small changes to the training set can result in significantly different tree structures.
- Random forests take advantage of this by using bootstrap aggregating (bagging), allowing each individual tree to randomly sample from the data set with replacement.

Randomizing the Forest

- Each tree in a random forest only uses a random subset of the features.
- This forces even more variation amongst the trees in the model and ultimately results in lower correlation across trees and more diversification.



Section 4

Model Selection

Models

- The objective is to create a model that correctly classifies the pulsar candidates.
- The models that were considered were:
  - Logistic Regression
  - K-Nearest Neighbors
  - Linear Discriminant Analysis
  - Support Vector Machine
  - Decision Tree
  - Random Forest

Performance Measures

- The model performance will be measured by not only accuracy, but also recall and precision.
- Recall is the model's ability to identify relevant instances (pulsars).
- Precision is the model's ability to identify only relevant instances.
- The F1 score will be used to measure recall and precision.

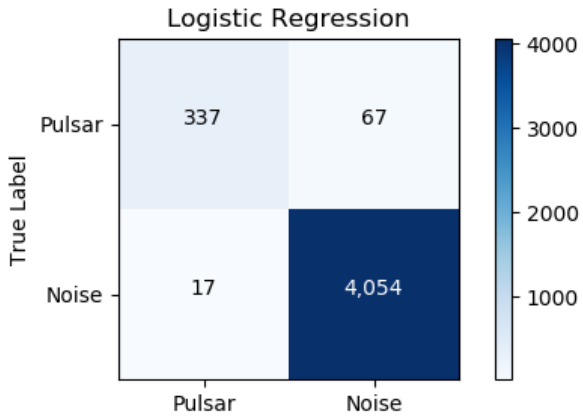
$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Section 5

Results

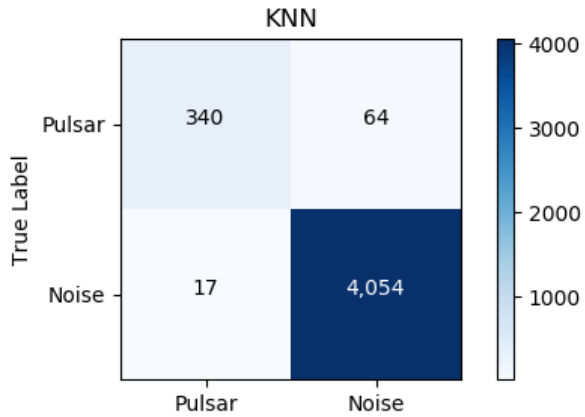
Logistic Regression

- The logistic regression model was tuned using 5-fold cross validation to find the optimal hyperparameters.
- Hyperparameters
  - Penalty: L1
  - C (Inverse Regularization Parameter): 7.75
- Performance
  - Accuracy: 98.1%
  - F1: .889



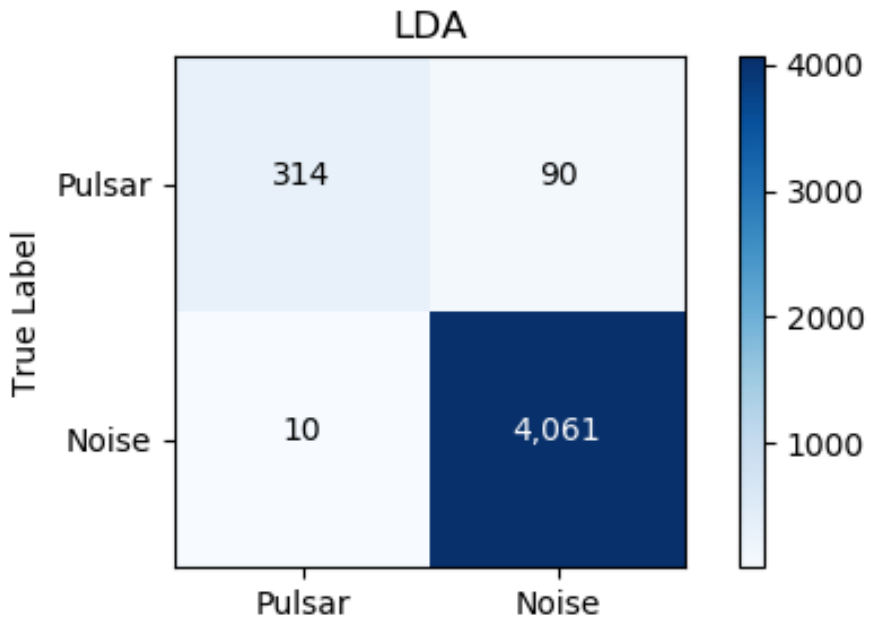
K Nearest Neighbors

- The KNN model was tuned using 5-fold cross validation to find the optimal hyperparameters.
- Hyperparameters
  - K: 9
  - Metric: Manhattan
- Performance
  - Accuracy: 98.2%
  - F1: .894



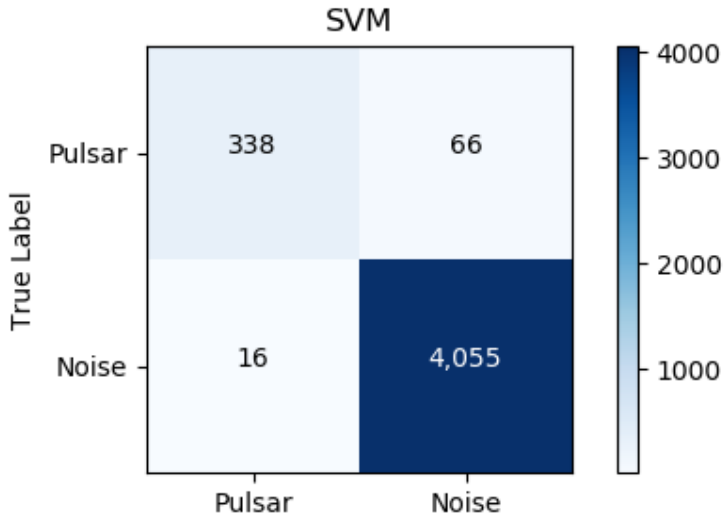
Linear Discriminant Analysis

- The LDA model was trained using 5-fold cross validation.
- Performance
  - Accuracy: 97.8%
  - F1: .863



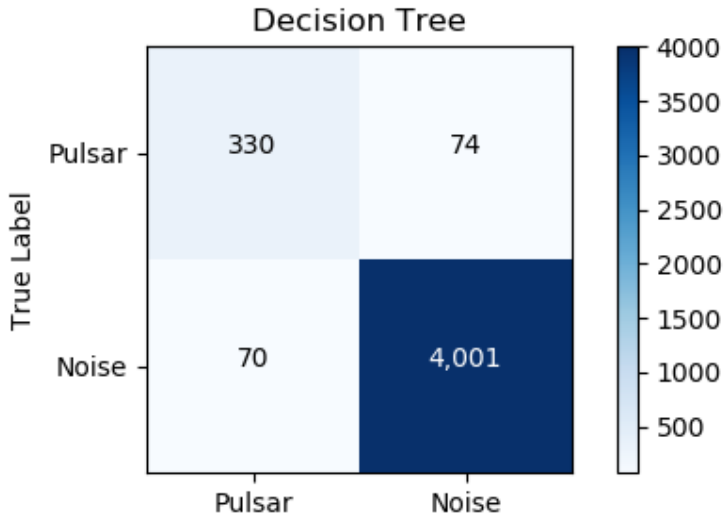
Support Vector Machine

- The SVM model was tuned using 5-fold cross validation to find the optimal kernel.
- Vernal: Linear
- Performance
  - Accuracy: 98.2%
  - F1: .892



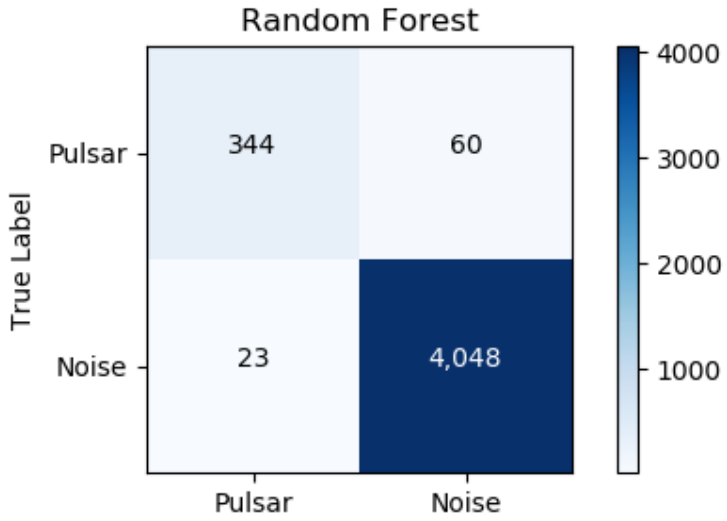
Decision Tree

- The decision tree model was tuned using 5-fold cross validation to find the optimal class weights.
- Class weights: 1:100
- Performance
  - Accuracy: 96.7%
  - F1: .821

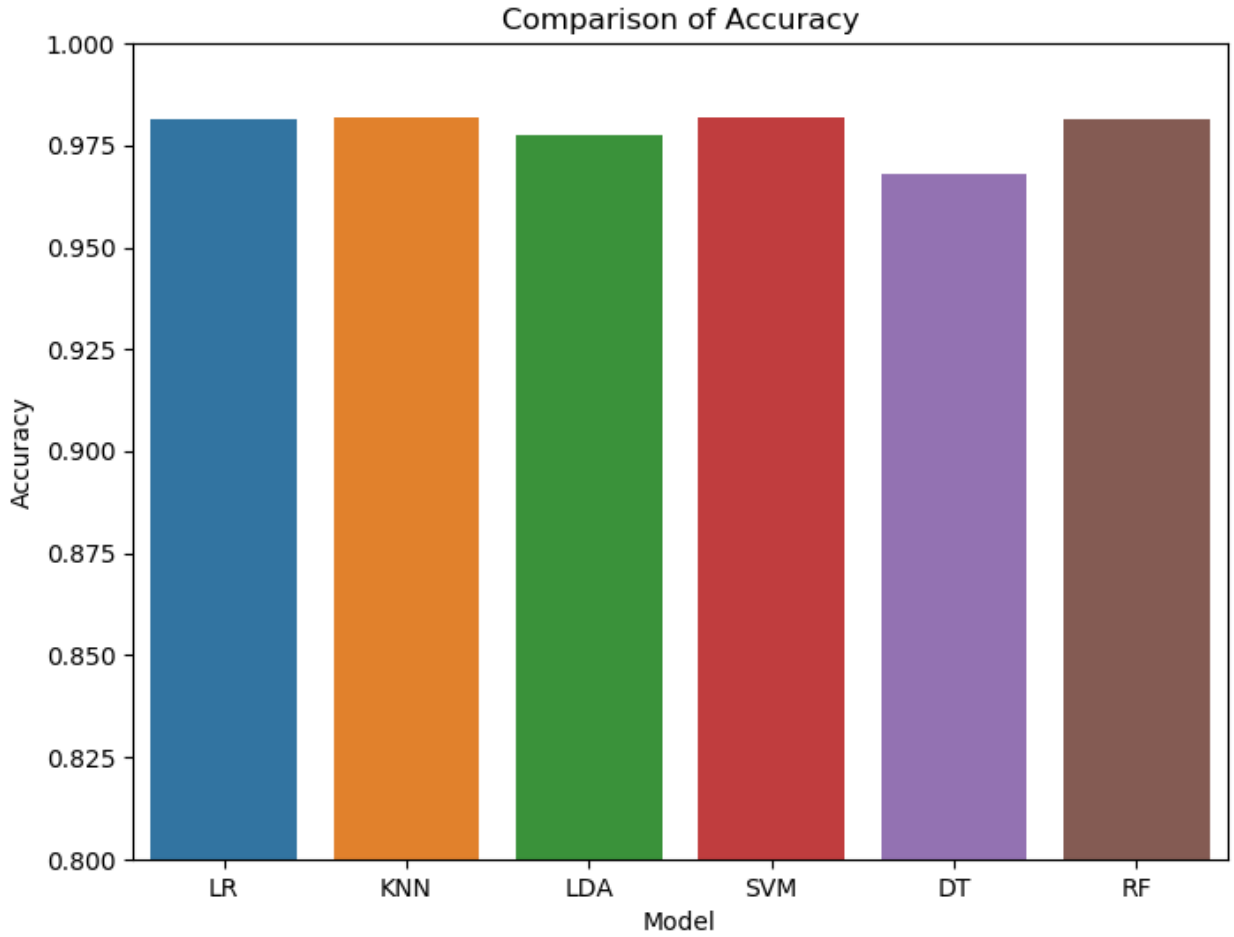


Random Forest

- The random forest model was tuned using 5-fold cross validation to find the optimal class weights.
- Class weights: 1:100
- Performance
  - Accuracy: 98.1%
  - F1: .892



Comparison of Accuracy



Comparison of F1 Score

Section 6

Conclusion

Conclusion

- The HTRU2 data set was used to train six different models to classify pulsar candidates.
- All models were highly accurate at classifying candidates, but due to the imbalance of the data, the F1 score was used to measure performance.
- KNN, SVM, and the Random Forest models had the highest F1 score (Random Forest had the highest recall)

Future Work

- Collect more pulsar data for a more balanced data set.
- Consider other ensemble classification models (gradient boosting, Ada-Boost, XGBoost).

References

An Introduction to Statistical Learning. 2013. Springer.

Rhys, Hefin I. 2020. *Machine Learning with R, the Tidyverse, and Mlr*. Manning.

n.d. https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

n.d. https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c.