Classifying Pulsar Stars

Cain Farnam

Abstract

Pulsars are highly dense neutron stars that are studied by astrophysicists and astronomers. Finding pulsars is a difficult task due to the large amount of radio frequency interference (RFI) that sometimes passes as pulsar detections. This paper aims to classify pulsar candidates by analysing emissions using various machine learning techniques. Six different models are trained using a data set containing pulsar candidates with eight features. The models are tested for performance and an optimal model is chosen to use for application.

---

1 Introduction

1.1 What is a Pulsar?

Pulsars are a rare type of neutron star that produce radio emissions detectable here on Earth. These emissions occur in short, periodic bursts that can only be recorded when the emission is aimed directly at Earth from the pulsar's poles. These emissions make the star "pulse", giving them their unique name. Dubbed the lighthouses of the cosmos, each pulsar produces a slightly different emission pattern. The emission pattern serves as a kind of fingerprint used to identify each star.

Having only been discovered in 1967, the first pulsar detected was thought to be contact from an alien race. The discoverers aptly named the star "LGM-1" for "little green men". Since that time, there have been over 2500 pulsars discovered. Though we have come a long way in our understanding of astronomy and cosmology, pulsars remain of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter. The periodic accuracy of some pulsars has even exceeded that of atomic clocks in keeping time.

Detecting pulsars is somewhat of a challenge. A potential signal detection is averaged over many rotations of the pulsar (integrated profile), as determined by the length of an observation. In practice, almost all detections are caused by RFI and noise, making legitimate signals hard to find.

1.2 Pulsar Data

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 data set is highly imbalanced, meaning there are much more observations that are not pulsars than there are pulsars (there are 10x as many observations that are not pulsars than there are actual pulsars). 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

The integrated profile is the average of all the pulsar emissions over a given amount of time. The DM-SNR curve is a dispersion measurement-signal to noise ratio plot. Dispersion is caused by the interstellar medium, and is different for every pulsar, depending on its distance and the number of electrons in the interstellar medium in the direction of the pulsar. Dispersion causes the lower frequencies of the signal to arrive later than the higher frequencies. This smears out, or disperses, the pulse. This smearing will completely obliterate the pulse if the signal is not de-dispersed before folding. Excess kurtosis is a statistical term describing that a probability has a kurtosis coefficient that is larger than the coefficient associated with a normal distribution, which is around 3.

2 Analysis

2.1 Objective

The objective is to create a model that correctly classifies the pulsar candidates. Due to the heavy imbalance of the data, the model performance will be measured using not only accuracy, but also recall and precision. Recall is the model's ability to identify relevant instances (pulsars). If recall is 100%, then all pulsars have been correctly classified. Precision, however, is the model's ability to identify only relevant instances. So by increasing recall, precision will decrease due to the non-relevant instances that get misclassified.

It is important to consider these measures because we are more interested in correctly identifying pulsars, not the observations that are caused by RFI. The trade-off between recall and precision can be calculated with another performance measure, the F1 score. The F1 score is the harmonic mean between recall and precision, calculated by the following:

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

2.2 Models

To classify the data, we will be using six different machine learning models:

- Logistic Regression

- K-Nearest Neighbors
- Linear Discriminant Analysis
- Support Vector Machine
- Decision Tree
- Random Forest

Each model will be tuned using their respective hyperparameters, and their model performances will be compared to find the optimal model.

2.3 Random Forest Classifier

The random forest model consists of a large number of individual decision trees that operate as an ensemble. This ensemble model is created in the hopes to increase model performance. The core idea behind an ensemble is a large number of relatively uncorrelated models operating as a committee will outperform any of the individual constituent models. In Figure 1, a representation of a random forest is shown where each individual decision tree classifies an observation. The class that is predicted by more decision trees then becomes the prediction of the forest.

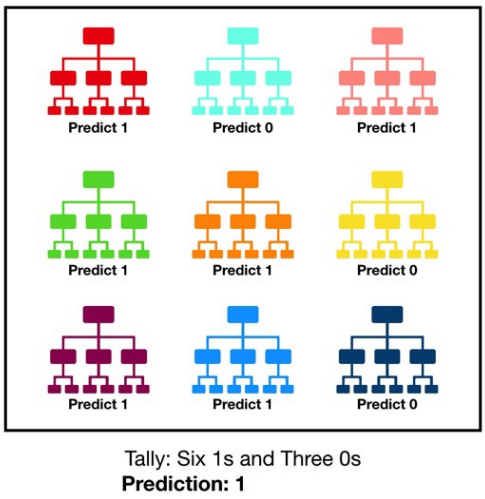

Figure 1: Random forest classifier

Each decision tree needs to be as uncorrelated as possible to the rest of the forest. Decisions trees happen to be 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. This technique will decrease correlation between each tree.

To further reduce correlation, each tree in a random forest only uses a random subset of the features. Figure 2 shows how a simple random forest may create trees using random feature selection.

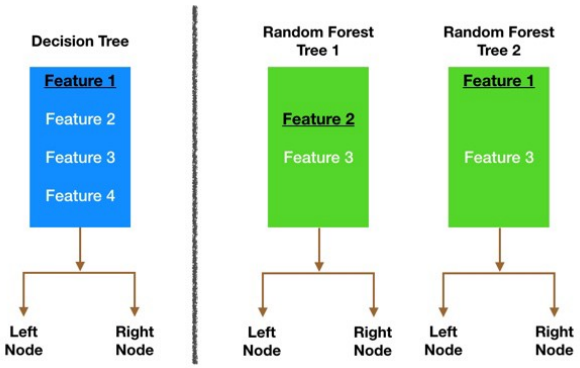

Figure 2: Random feature selection

3 Results

3.1 Logistic Regression

The logistic regression model was tuned using 5-fold cross validation to find the optimal hyperparameters. They hyperparameters considered were the type of penalty (L1 or L2) and the regularization parameter. After testing various combinations of hyperparameters, the optimal logistic regression model was found to have the following hyperparameters:

- Penalty: L1
- C (inverse regularization parameter): 7.75

The model performance measures from this model were as follows:

- Accuracy: 98.1%
- F1: .889

Figure 3 provides the confusion matrix for the model. A high recall would mean that the value in the upper right corner of our matrix should be lower. This model left 67 pulsars misclassified. We'll monitor this value between all models.

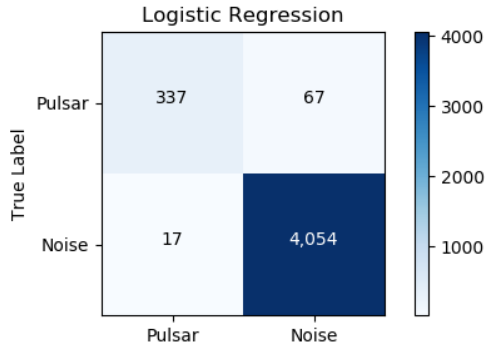

Figure 3: Logistic regression confusion matrix

3.2 K Nearest Neighbors

The KNN model was tuned using 5-fold cross validation to find the optimal hyperparameters. The hyperparameters considered were the value of K and which metric to measure the distance. After testing odd K values ranging from 3-17 and various metrics, the KNN model was found to have the following hyperparameters:

- K: 9
- Metric: Manhattan

The model performance measures from this model were as follows:

- Accuracy: 98.2%
- F1: .894

The confusion matrix in Figure 4 informs us that there were 64 pulsars that the model was unable to classify, slightly better than the logistic model.

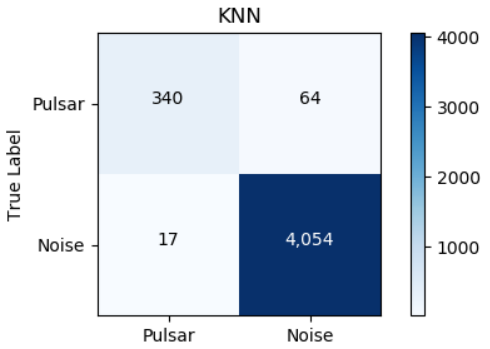

Figure 4: K-nearest neighbors confusion matrix

3.3 Linear Discriminant Analysis

The LDA model was trained using 5-fold cross validation. There we no hyperparameters that were optimized for this model.

The model performance measures from this model were as follows:

- Accuracy: 97.8%
- F1: .863

Figure 5 provides the confusion matrix. This model is failing to classify 90 pulsars, the lowest recall of the three models so far.

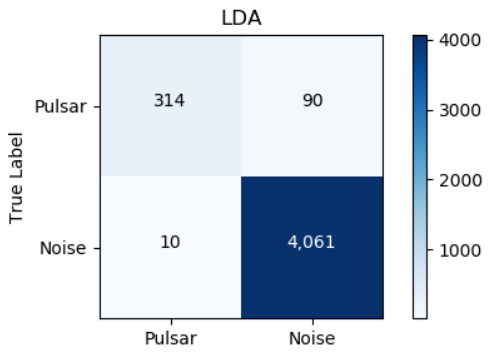

Figure 5: Linear discriminant analysis confusion matrix

3.4 Support Vector Machine

The SVM model was tuned using 5-fold cross validation to find the optimal kernel. The optimal kernel for this model was found to be a linear kernel.

The model performance measures from this model were as follows:

- Accuracy: 98.2%
- F1: .892

Figure 6 displays the confusion matrix for the SVM model. The model misses 66 pulsars.

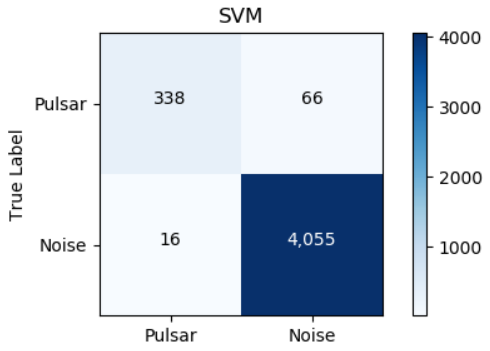

Figure 6: Support vector machine confusion matrix

3.5 Decision Tree

The decision tree model was tuned using 5-fold cross validation to find the optimal class weights. This hyper parameter was optimized due to the data being so imbalanced. Had this hyper parameter been left as the default, the model would be slightly biased to the RFI observations. The optimal weights to use were ones that "balanced" the data. The model uses weights inversely proportional to the class frequencies.

The model performance measures from this model were as follows:

- Accuracy: 96.7%
- F1: .821

Figure 7 displays the confusion matrix for the decision tree model. This model fails to predict 74 pulsars.

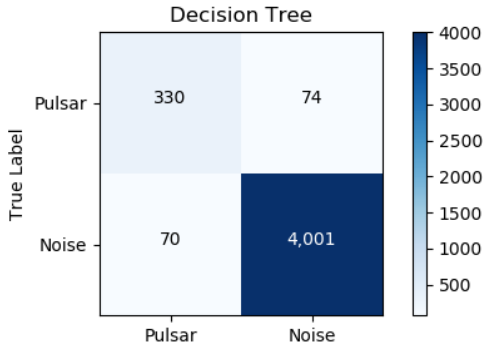

Figure 7: Decision tree confusion matrix

3.6 Random Forest

The random forest model was tuned using 5-fold cross validation to find the optimal class weights. Again, the optimal weights were found to be weights that balanced the data set.

The model performance measures from this model were as follows:

- Accuracy: 98.1%
- F1: .892

The confusion matrix in Figure 8 displays the number of misclassified pulsars to be 60, the lowest of all the models.

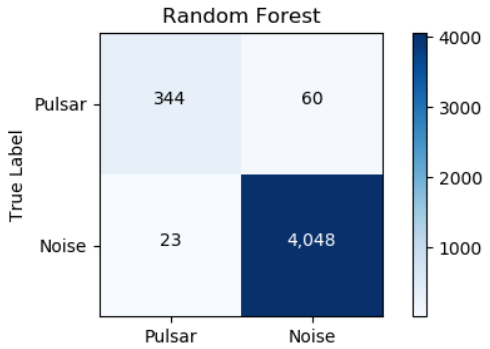

Figure 8: Random forest confusion matrix

3.7 Comparison of Models

The graph in Figure 9 compares all the model accuracies. It is clear to see that they are all relatively close. If this were the only performance measure that was compared, any one of these models could be chosen.

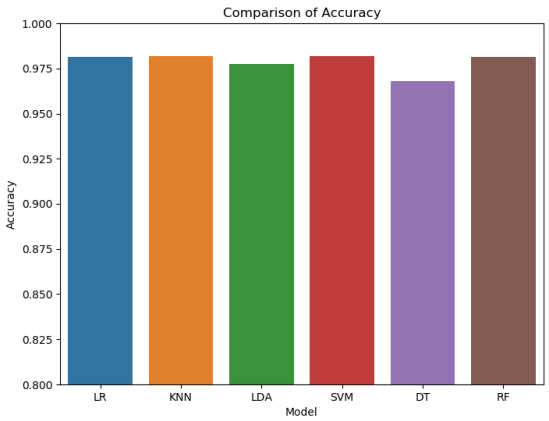

Figure 9: Comparison of model accuracy

However, by looking at Figure 10 we can see that there is quite a big difference in the models' F1 scores. It looks as the KNN, SVM, and the random forest are performing the highest when it comes to F1 score.

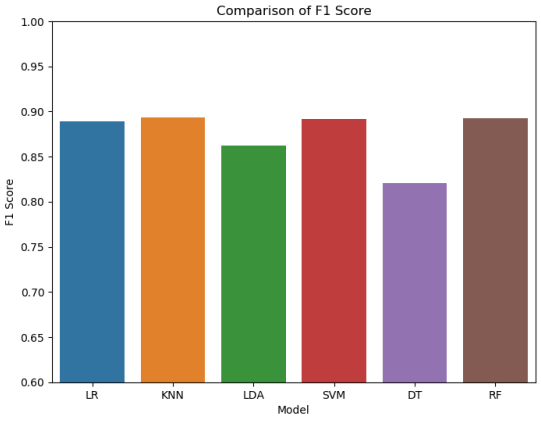

Figure 10: Comparison of model F1 score

4 Discussion

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). Due to random forest having the highest recall, as well as having a lower train time than SVM, it is the proposed model to use. It would also be preferred to KNN, as KNN can be highly computational when it comes to predicting new candidates.

5 Future Work

It is important to always consider new ways to improve the models. With the James Webb telescope being slated to become operational in the next 2-3 years, the number of pulsars discovered is likely to increase tremendously. The models were could be greatly improved with said data. There are also other modeling techniques that were not considered here. Other ensemble classification models (gradient boosting, Ada-Boost, XGBoost) or neural networks may be able to classify the candidates with higher recall. Another way of to increase recall would be to lower the probability threshold for models that calculate probability (this paper used a 50:50 threshold).

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.