Report

Conner McCloney

# Background

In 2011, a survey known as the High Time Resolution Survey was conducted, which was a digital, all-sky survey designed specifically to scan the night sky for potential pulsars, or pulsar candidates, and determine their validity. A pulsar is a rare type of rotating Neutron star that emits electromagnetic radiation, and can only be measured when this beam is aimed directly at Earth. These pulsars were searched for using large radio telescopes looking for periodic radio signals that a pulsar would produce. Many measurements taken that could be a pulsar, known as a candidate, are recorded, however in practice most observations recorded are caused by radio frequency interference (RFI) and noise. This data set records the mean, standard deviation, skewness, and excess kurtosis of the integrated profile, which is an array of variables that describe the signal recorded, and of the DM-SNR curve. The DM-SNR curve, which stands for Dispersion Measure - Signal-to-Noise Ratio, describes the relationship between the two for the observed signal, where a curve whose SNR peaks at a DM of zero is likely RFI, or if it is a legitimate signal it should peak at a DM greater than zero. Lastly, there is the target\_class variable, which is a binary classification variable of whether the given signal was truly a pulsar or not. I am interested in investigating which, if any, of the characteristics of the DM-SNR curve and/or the integrated profile are useful in accurately classifying a pulsar candidate, and how accurate a model of this sort is.

# Model Determination

From this data set, data for eight variables in total were collected for every observation: the mean, standard deviation, skewness, and excess kurtosis of both the DM-SNR curve and the integrated profile. From here, I decided to begin with a ‘full’ additive model that included each of these variables as a predictor variable. To answer the research question defined above, I chose to determine which of these variables had sufficient evidence to keep as predictor variables, and which could be dropped to simplify the model. My technique was to combine a p-value and VIF-based approach, by first determining which of the predictors in the model had large p-values, then dropping these and re-fitting this reduced model. Or, if all p-values were sufficiently small, looking at the VIFs for each of the predictors and removing any explanatory variables with large VIFs, refitting this reduced model, and repeating this process. Because these variables are recorded measurements on the same metrics, I would expect to see some form of relationship present between many of the variables, and examining VIFs helps us to find any multicollinearity present in the model. Following the procedure outlined, we end up with a model that includes the mean and skewness of the integrated profile, and the mean and standard deviation of the DM-SNR curve as our predictors. Examining the binned residual plot for this model, and for the predictors present in this model however, it is clear that a pattern is present in the residuals. Including interactions, transformations, and polynomial terms in the models did not fix this issue, therefore I have chosen to proceed with testing using just these simpler additive models. The models and the explanatory variables associated with each is displayed below. The binned residual plots can be found in the Appendix.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Mean | SD | Excess Kurtosis | Skewness | Mean | SD | Excess Kurtosis | Skewness |
| Full Model | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Reduced Model 1 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |  |  |
| Reduced Model 2 | ✓ | ✓ |  |  | ✓ | ✓ |  |  |
| Reduced Model 3 | ✓ |  |  |  | ✓ | ✓ |  |  |

*Table 1*

|  |  |
| --- | --- |
| Integrated Profile | DM-SNR Curve |

# Model Testing

Starting from the ‘full’ model, and refining it down to these four predictors using my technique, generated two more intermediately complex models. I tested these four models, and the reduced model with a squared transformation for predictive ability. The data was split into a training and test set, and the predictive ability of each model was measured and displayed using a confusion matrix. A row defines the true value of an observation, where ‘0’ represents a False Positive reading of a Pulsar candidate, and ‘1’ a true Pulsar reading. A column represents the predicted value of an observation using that model at a decision rule of 0.5.

## Predicted Label - Full Model  
## True Label 0 1  
## 0 15013 267  
## 1 80 1288

## Predicted Label - Reduced Model #1  
## True Label 0 1  
## 0 15014 270  
## 1 79 1285

## Predicted Label - Reduced Model #2  
## True Label 0 1  
## 0 14990 370  
## 1 103 1185

## Predicted Label - Reduced Model #3  
## True Label 0 1  
## 0 14987 373  
## 1 106 1182

From these results, we can see that the very best results come from the full model with a 97.9% accuracy rate, and the highest error rate belonging to the final reduced model with a 97.1% accuracy rate. However, this is only a difference of 0.8%, indicating that the predictor variables present only in the full model were not necessary for accurate classification. Further, all four models have a very small test error rate, indicating that any of these would be very useful and accurate in determining the validity of a pulsar candidate.

For comparison, the k-fold Cross Validation technique was also used at a value of =10. This technique will split the data set into non-overlapping ‘folds’, with each fold acting as a validation set, with the remainder of the data acting as a training set. analyses will be run, and the overall test error rate is the average of all rates. Mathematically, this is represented as,

, where . The misclassification rates for each model using this technique are shown below.

## Full Model: 0.01725574

## Reduced Model 1: 0.01736615

## Reduced Model 2: 0.0233515

## Reduced Model 3: 0.02335574

Using k-fold Cross Validation, we see misclassification rates that are around 2% for all four models. This, again, is an extremely low test error rate, and seems to be just slightly better than the error rates the logistic regression models tested above generated.

# Conclusion