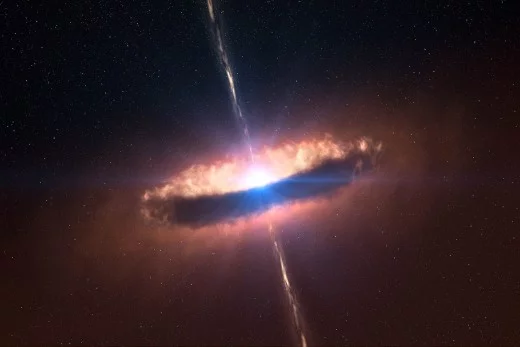
**Executive Summary: Studying the Relationship between the Moments of Integrated Pulse Profile and the Mean of the Dispersion Measure of the Signal to Noise Ratio Curve for Spuriously Recorded Non-Pulsar Neutron Stars**

By Mukund Raghav Sharma [Moko Sharma]



**Introduction**

A **Neutron Star** is essentially a collapsed core of a giant star which before collapse had a total mass of between 10 and 29 solar masses. A **Pulsar** is a rare type of Neutron Star that produces radio emission of such high intensity that they are detectable from millions of light years away. As Pulsars rotate, their emission beams sweep across the sky and produces a periodic detectable pattern of broadband radio emission. Pulsars are of paramount importance to scientific interest as they are probes of space-time, the interstellar medium and various states of matter as per Lorimer et al. (2005).

The methodology to detect a Pulsar is based on the statistics of close recording of the radio frequencies emitted by them over a definite period of time. Unfortunately, since there is high variance in the way these celestial bodies emit their radiation, sometimes certain patterns of emissions of a non-Pulsar Neutron Stars are falsely recorded and can be spuriously be classified as Pulsars; this false positive is usually a result of high Radio Frequency Interference and noise as per Lyon (2016).

The stakeholders or target audience here are hence, astrophysicists, celestial researchers and astronomers in companies such as NASA as it is extremely important for astrophysicists and celestial researchers to ensure the false positive non-Pulsar results aren’t peppered into the set of true Pulsars as it mars the integrity of the data used for further studies to reveal newer and more significant secrets of the universe via the study of Pulsar stars. Additionally, there is a clearly a cost to collecting, sanitizing and analyzing the data and therefore, it is prudent to understand the relationship between the fundamental observed quantities before jumping to collect and analyze more data points. Defining and describing this aforementioned relationship via a Supervised Statistical Learning technique is one of the goals of this project. Since the response variable here is a numeric variable, it is clear that some form of Regression will be used to conduct the relationship analysis.

**Details of the Data Set**

The data at hand from the famous HTRU2 Dataset by Keith et al. (2010) contains samples of Pulsar candidates collected during the High Resolution Universe Survey (South) consisting of statistics based on the **Integrated Pulse Profile, Dispersion Measure Signal to Noise Curve (DM-SNR Curve)** and a **label** highlighting if the candidate is truly a Pulsar Star or a Spuriously recorded plain vanilla Neutron Star.

**Integrated Pulse Profiles** are aggregations of single pulses of each of the candidates i.e. aggregated pulse emissions from each individual rotation; note: we don’t even consider pulse emissions from individual rotations because of their highly variable nature unlike integrated pulse profiles that are more consistent and stable over the passage of time. Computing the first through 4th statistical moments [technically standard deviation is not a moment but variance is] of these integrated pulse profiles gives us insight about the distribution of integrated pulses.

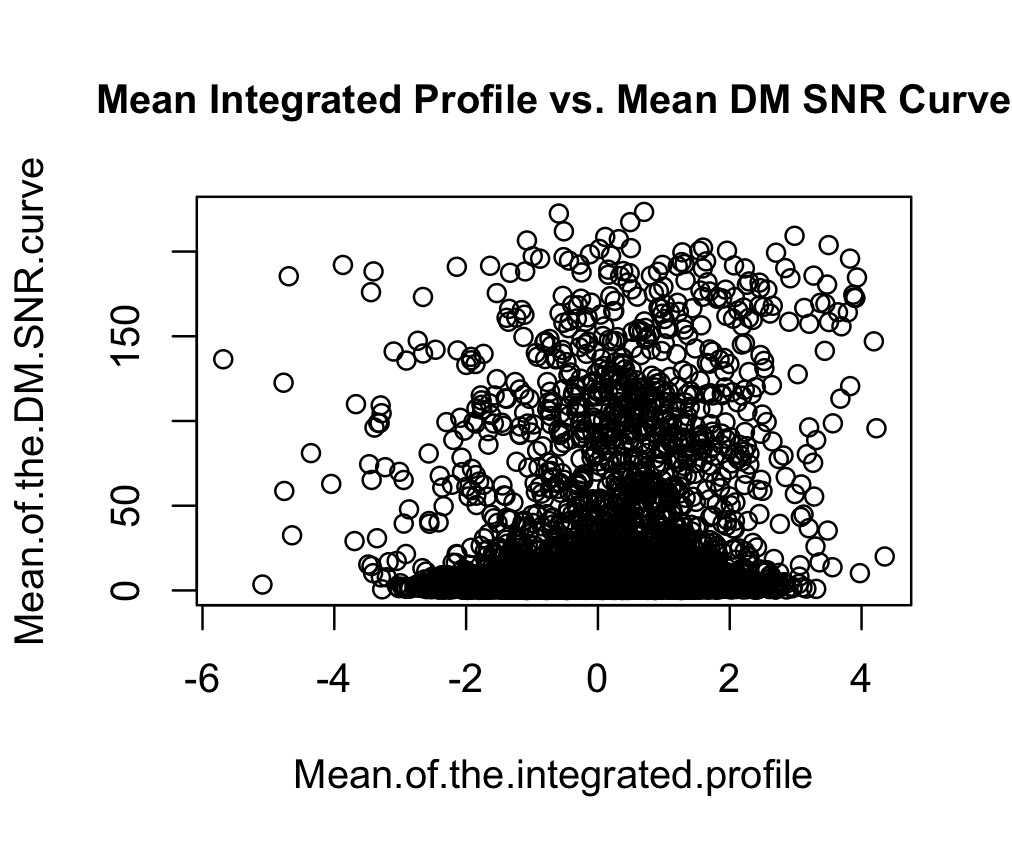
**Dispersion Measure Signal to Noise Curve (DM-SNR Curve)** elucidates on a unique blue print of the emission of the particular candidate and helps us distinguish between Pulsar emissions and other radar-microwave generating noise in the universe. For our analysis, we will only consider the mean of DM-SNR curve vs. the other moments as it serves as a reasonable proxy of the DM-SNR Curve behavior.

As stated before, the goal of this project is to make use of the data from the spuriously identified non-pulsar stars and to try to find a relationship between the Integrated Pulse Profiles and the mean DM-SNR Curve per candidate. The dataset aids the goal as it contains a majority of the Non-Pulsars that were spuriously considered (16,259 vs. 1,639) highlighting another reason to study the relationship.

**Exploratory Data Analysis**

The exploratory data analysis portion is the first step in the model creation phase as it is extremely important to visualize the relationships between the predictors and response variables. After transforming any right skewed variables and scaling, correlation analysis and plots between the predictors and response are created. It is clear from the correlation plot below that there is high correlation, both positive and negative, amongst the predictor variables and not necessarily between the predictors and response. Additionally, based on observing the plots of predictor vs. the response, it is clear that there exists an extremely non-linear relationship between them.



****

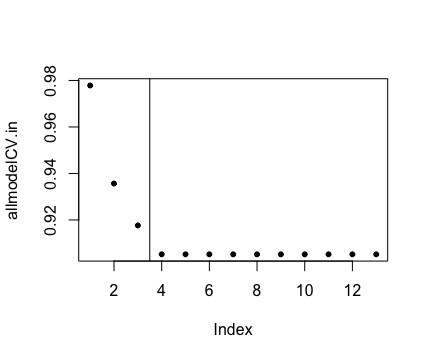
**Setting up the Experiment: Model Considerations and Hyperparameter Tuning**

Using the insight gained in the exploratory data analysis phase, the conclusions are that there exists a highly non-linear relationship amongst the predictors vs. response and high correlation amongst the predictors. Ruling out any type of model that has some semblance to a linear based regression approach and considering models that can handle high correlation amongst predictor variables, the choices considered are **Random Forest** and **Artifical Neural Networks**. These two choices were a result of the need to choose a sophisticated model that can also handle correlation amongst the predictors to best represent the data at hand.

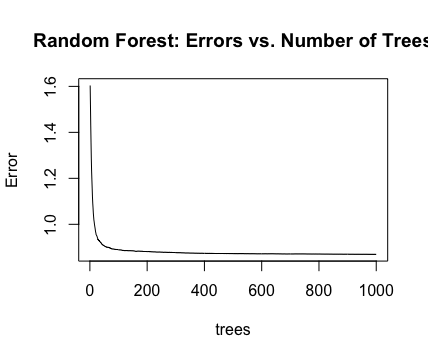
To state explicitly, models based on **Random Forest** and **Artificial Neural Networks** are used to establish the relationship between the statistical moments of the integrated pulse profiles and mean of the DM-SNR curve for all the candidates. A 10 fold double cross validation will be used to get the model, for each fold that has the lowest cross validation.

The hyperparameters tuned for the Random Forest will be the number of variables sampled as candidates at each split or “**mtry**” where the number of trees grown will be set to 500 to ensure that there are trees that are averaged to reduce overall variance. The number of variables sampled as candidates at each split was taken to be between 1 and 3 based on the recommendation given for regression i.e. n / 3 where n is the number of predictors. Similarly, the hyperparameter tuned for Artificial Neural Network is the number of hidden layers that’ll range from 1 to 10.

**Model Selection, Validation and Analysis**

The best performing model as a majority from the double 10 fold cross-validation turned out to be the Random Forest with the number of variables randomly sampled as candidates at each split equal to 1 as per the plot indicating the Cross Validation Error rate across all 13 models considered.

The overall best models Double Cross Validation Error Rate of 0.8867948 however, the R-Squared value of 0.1131507 highlights a low predictive power of even the best model in each of the folds with the lowest Cross Validation Error Rate.

For the sake of experimentation, isolating the best performing Random Forest Model with the number of variables randomly sampled as candidates at each split equal to 1, it is observed there is no significant decrease in the error rate implying diminishing returns if we were to increase the number of trees produced as a means of further optimizing the model and 13.05% of variance is explained by the model. Additionally, performing 10-fold single cross validation on the best performing model gives a mean of squared residuals of 0.8694308 and 13.22% of variance is explained.

**Conclusion**

The final model chosen is that of a Random Forest one with the number of variables randomly sampled as candidates at each split equal to 1. The overall predictive power is low based on the 10-fold single cross validation mean of squared residuals of 0.8694308 and 13.22% of variance explained and therefore, it is probably not prudent and safe to not overstate the inference that there is isn’t strong relationship between the moments of the integrated pulse profile and the mean of the DM-SNR curve for non-pulsar neutron stars that have been spuriously considered as Pulsars.

**References**

D. R. Lorimer and M. Kramer, 'Handbook of Pulsar Astronomy', Cambridge University Press, 2005.

M. J. Keith et al., 'The High Time Resolution Universe Pulsar Survey - I. System Configuration and Initial Discoveries',2010, Monthly Notices of the Royal Astronomical Society, vol. 409, pp. 619-627. DOI: 10.1111/j.1365-2966.2010.17325.x

R. J. Lyon, 'Why Are Pulsars Hard To Find?', PhD Thesis, University of Manchester, 2016