Pulsar Classification

Anil Kumar P

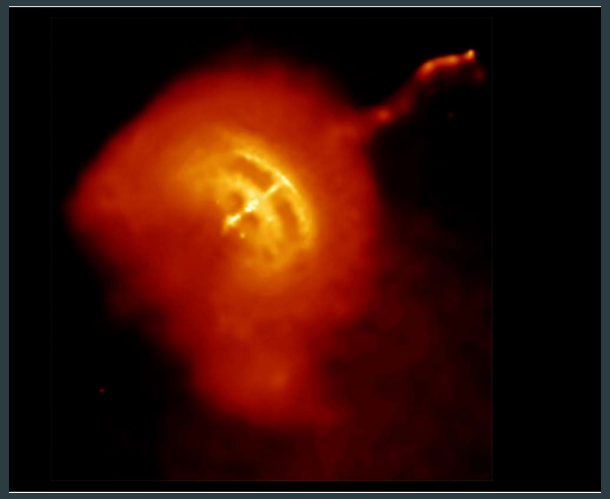
Harrisburg University

Agenda:

- Introduction
- Data Description
- Objective
- Data Prep & Exploratory Analysis
- Prediction
- Modeling
  - Model I
  - Model II
- Performance Analysis
- Conclusion
- Q & A
- My Learnings
- Citations

Introduction:

Crab Nebula



Vela Pulsar

- Pulsar is a neutron star which spins and it's emotions are equal to x-ray and gamma wavelengths.
- It provides patterns which can be identified from the Earth.
- These patterns can be observed using large radio telescopes.

Data Description:

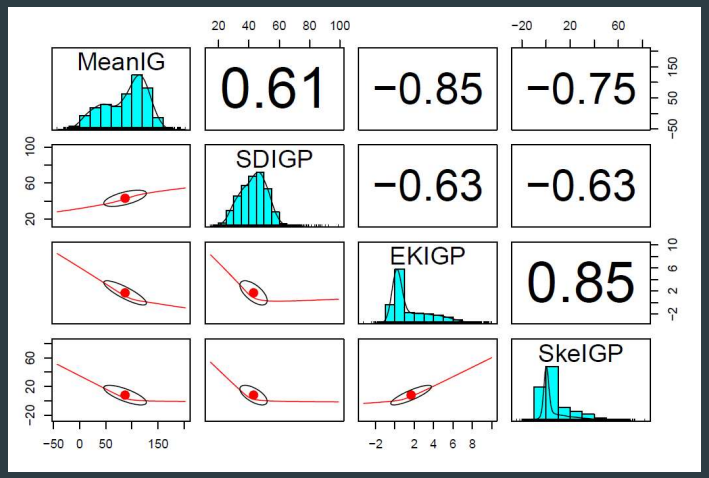
- HTRU2 (High Time Resolution Universe Survey) consists of 17,898 observations with 9 variables.
- This data set has labeled data.
- 16,259 observation are classified as noise and rest 1,639 are classified as pulsars.
- This data set has cleaned data but is very much unbalanced.
- Out of 9 variables, first 4 variables are related to integrated profile measures, next 4 are related to DM-SNR curve measures and last one is the type of class.
- This data set is annotated by humans.

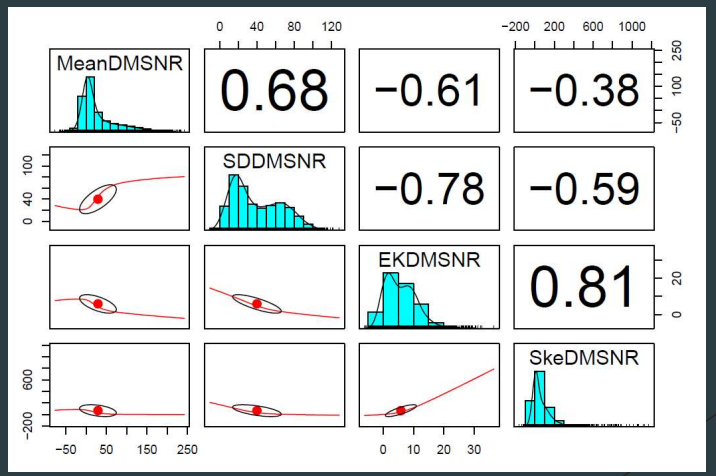
Objective:

The objective of this modeling is to help astronomic survey department with machine learning algorithm to classify the candidate profile data into pulsar and noise groups with accuracy rate of more than 90%. To accomplish this task, as a part of modeling, I am going to use Decision Tree classifier and Naïve Bayes classifier algorithms. The balanced data is spit into train set and test set. Train set data is used to train the Decision Tree Classifier and test set is used to validate the classifier.

Data Prep & Exploratory Analysis:

- As I have cleaned and labeled data, I do not have much tasks in data cleaning.
- I had started exploring the dataset with finding correlations and distributions of the data.
- Scatter plot matrix has been used to identify correlation and distribution of the data in a single chart.





Prediction:

- When I used Decision Trees classification algorithm to classify the given data, confusion matrix results were.

	Reference	
Prediction	0	1
0	3225	55
1	20	281

Accuracy : 0.9791

- Area under the curve (AUC) value for pulsar predictors was 0.915.
- Confusion matrix says that data is not balanced.
- Train and test set proportions of given data set.
- To over come this issue, we should balance the data.
- After balancing the data set, proportions of train and test values were:

```
> prop.table(table(trainset$Class))*100

       0        1
90.899567  9.100433
> prop.table(table(testset$Class))*100

       0        1
90.617146  9.382854
>
```

```
> prop.table(table(trainRose$Class))*100

       0        1
50.32826 49.67174
> prop.table(table(testRose$Class))*100

       0        1
51.43815 48.56185
```

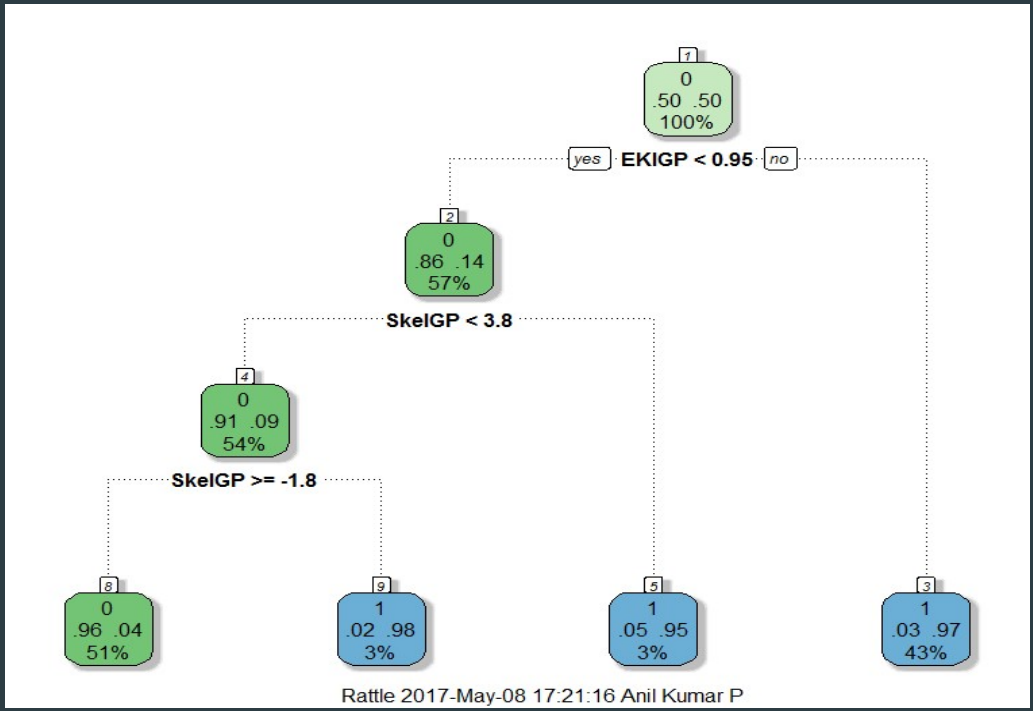
Model - I:

I have used Decision Tree classification algorithm to generate model to classify pulsar or noise from the modified balanced data set. I have used '*rpart*' package to perform the modeling. To predict the test values, I have used '*rpart*' object with '*predict*' function which provides array of classified values where we can check the performance with actual test set.

To understand the performance of the model in a better way, I have used confusion matrix and Area Under Curve (AUC) to check the classified values.

Model - I Tree Structure:

The tree structure of rpart object looks as illustrated below.



Model - II:

I have used Naïve Bayes classification algorithm to generate model to classify pulsar or noise from the modified balanced data set. I have used '*e1071*' package to perform the modeling. To predict the test values, I have used *Naïve Bayes* object with *predict* function which provides array of classified values to check the performance with actual test set.

To understand the performance of model in a better way, I have used Area Under Curve (AUC) and confusion matrix to check the classified values.

Performance Analysis:

Confusion Matrix and Statistics

Prediction	Reference 0	1
0	1779	50
1	63	1689

Accuracy : 0.9684
95% CI : (0.9622, 0.9739)
No Information Rate : 0.5144
P-Value [Acc > NIR] : <2e-16
Kappa : 0.9369
Sensitivity : 0.9658
Specificity : 0.9712
Pos Pred Value : 0.9727
Neg Pred Value : 0.9640
Prevalence : 0.5144
Detection Rate : 0.4968
Detection Prevalence : 0.5108
Balanced Accuracy : 0.9685
Area under the curve (AUC): 0.969

Confusion Matrix and Statistics

Prediction	Reference 0	1
0	1750	62
1	92	1677

Accuracy : 0.957
95% CI : (0.9498, 0.9634)
No Information Rate : 0.5144
P-Value [Acc > NIR] : < 2e-16
Kappa : 0.914
Sensitivity : 0.9501
Specificity : 0.9643
Pos Pred Value : 0.9658
Neg Pred Value : 0.9480
Prevalence : 0.5144
Detection Rate : 0.4887
Detection Prevalence : 0.5060
Balanced Accuracy : 0.9572
Area under the curve (AUC): 0.957

Conclusion:

- Both models are performing well in terms of accuracy because both model's accuracy level is more than 90%.
- Decision Tree classifier performance is higher when compared with Naïve Bayes classifier.
- Decision Tree will be added to the pulsar identifying process pipeline to classify the pulsar and noise from observed patterns.

My Learnings:

Sampling methods:

- Under Sampling: Works with majority class
  - Random
  - Informative
    - *EasyEnsemble*
    - *BalanceCascade*
  - Chance of loosing the data
- Over Sampling: Works with minority class
  - Random
  - Informative
  - No Information loss but leads to overfitting
- Synthetic Data Generation: uses synthetic minority oversampling technique
  - Uses SMOTE algorithm.
    - *Bootstrapping*
    - *K-nearest neighbours*
- Cost Sensitive Learning: Likely provides alternative for sampling and provides cost associated with misclassifying observations.

Citations:

- 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
- D. R. Lorimer and M. Kramer, 'Handbook of Pulsar Astronomy', Cambridge University Press, 2005.
- R. J. Lyon, 'Why Are Pulsars Hard To Find?', PhD Thesis, University of Manchester, 2016.
- R. J. Lyon et al., 'Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach', Monthly Notices of the Royal Astronomical Society 459 (1), 1104-1123, DOI: 10.1093/mnras/stw656
- R. P. Eatough et al., 'Selection of radio pulsar candidates using artificial neural networks', Monthly Notices of the Royal Astronomical Society, vol. 407, no. 4, pp. 2443-2450, 2010.
- S. D. Bates et al., 'The high time resolution universe pulsar survey vi. an artificial neural network and timing of 75 pulsars', Monthly Notices of the Royal Astronomical Society, vol. 427, no. 2, pp. 1052-1065, 2012.
- D. Thornton, 'The High Time Resolution Radio Sky', PhD thesis, University of Manchester, Jodrell Bank Centre for Astrophysics School of Physics and Astronomy, 2013.
- K. J. Lee et al., 'PEACE: pulsar evaluation algorithm for candidate extraction a software package for post-analysis processing of pulsar survey candidates', Monthly Notices of the Royal Astronomical Society, vol. 433, no. 1, pp. 688-694, 2013.
- V. Morello et al., 'SPINN: a straightforward machine learning solution to the pulsar candidate selection problem', Monthly Notices of the Royal Astronomical Society, vol. 443, no. 2, pp. 1651-1662, 2014.
- R. J. Lyon, 'PulsarFeatureLab', 2015, [Web Link].
- R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, J. D. Knowles, Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach, Monthly Notices of the Royal Astronomical Society 459 (1), 1104-1123, DOI: 10.1093/mnras/stw656.
- R. J. Lyon, HTRU2, DOI: 10.6084/m9.figshare.3080389.v1.
- Analytic Vidhya Content Team, ' Practical Guide to deal with Imbalanced Classification Problems in R', Analytics Vidhya, 2016.

The End