Pulsars

Dylon Hussain

5/3/2022

Introduction

Here I describe an algorithm that identifies pulsars from metrics of their observed signals. Pulsars are difficult to identify because they have a periodic signal but the observed signal from stars that are not pulsars are often periodic as the result of noise. Considering the large amount of stars that are observed and the scarcity of pulsars, there is great interest in utilizing machine learning to identify pulsars. Here we examine the suitability of multiple popular machine learning algorithms for this task. The dataset used for this analysis contains 17,898 observations; 1639 of which were examined by humans and were determined to have come from pulsars (referred to as '1') and the remaining observations did not (referred to as '0'), but have similar profiles due to noise. This dataset contains 8 predictors, listed below in Table 1, which are functionals on the observed signals.

Table 1: Predictors

Predictors	Predictors_names
Mean of the integrated profile.	pmean
Standard deviation of the integrated profile.	psd
Excess kurtosis of the integrated profile.	pkurt
Skewness of the integrated profile.	pskew
Mean of the DM-SNR curve.	cmean
Standard deviation of the DM-SNR curve.	csd
Excess kurtosis of the DM-SNR curve.	ckurt
Skewness of the DM-SNR curve.	cskew

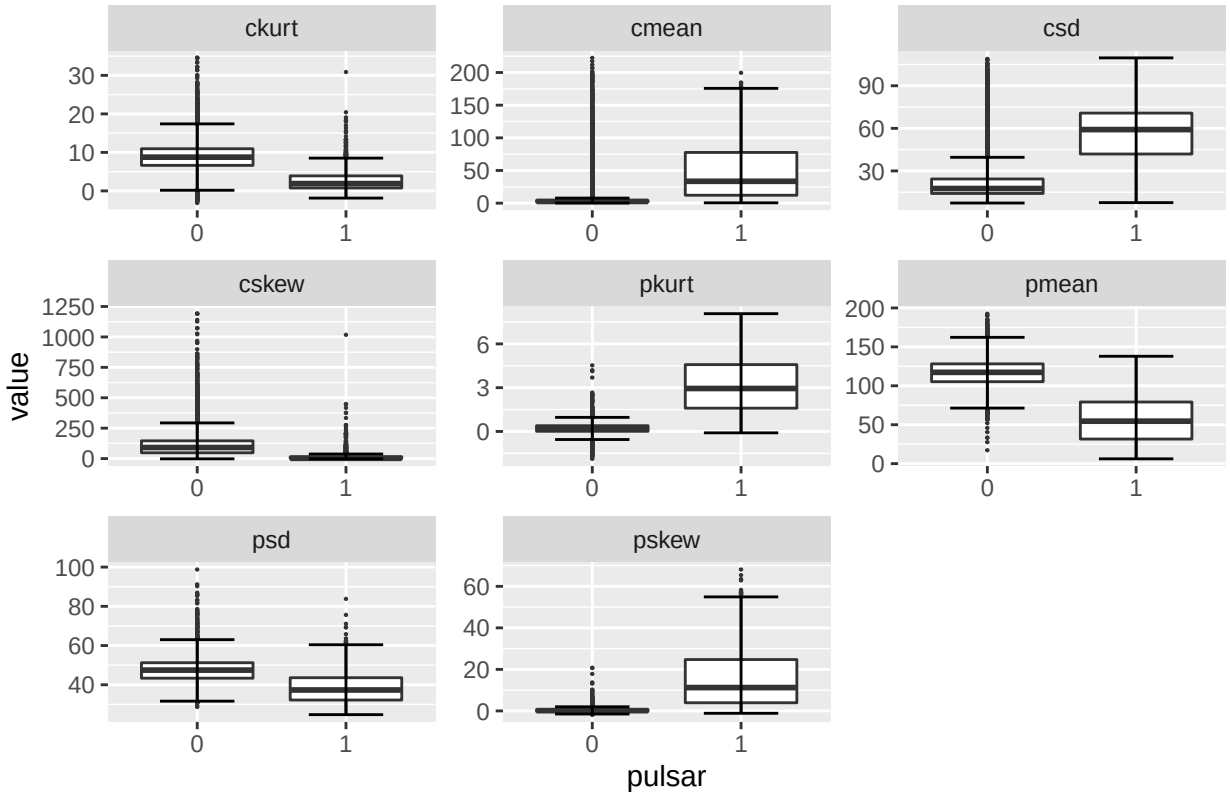
Analysis

The dataset was split into a train set to train the algorithm and a test set to test its performance. A 70:30 split was chosen for this analysis because it resulted in an adequate number of pulsars to train the algorithm, more than 1000. This split also retained a sufficient number of pulsars for an accurate estimate of its performance in future applications, whereas a 90:10 split would have resulted in less than 200 pulsars in the test set and may not have provided an accurate estimate.

Predictor Selection

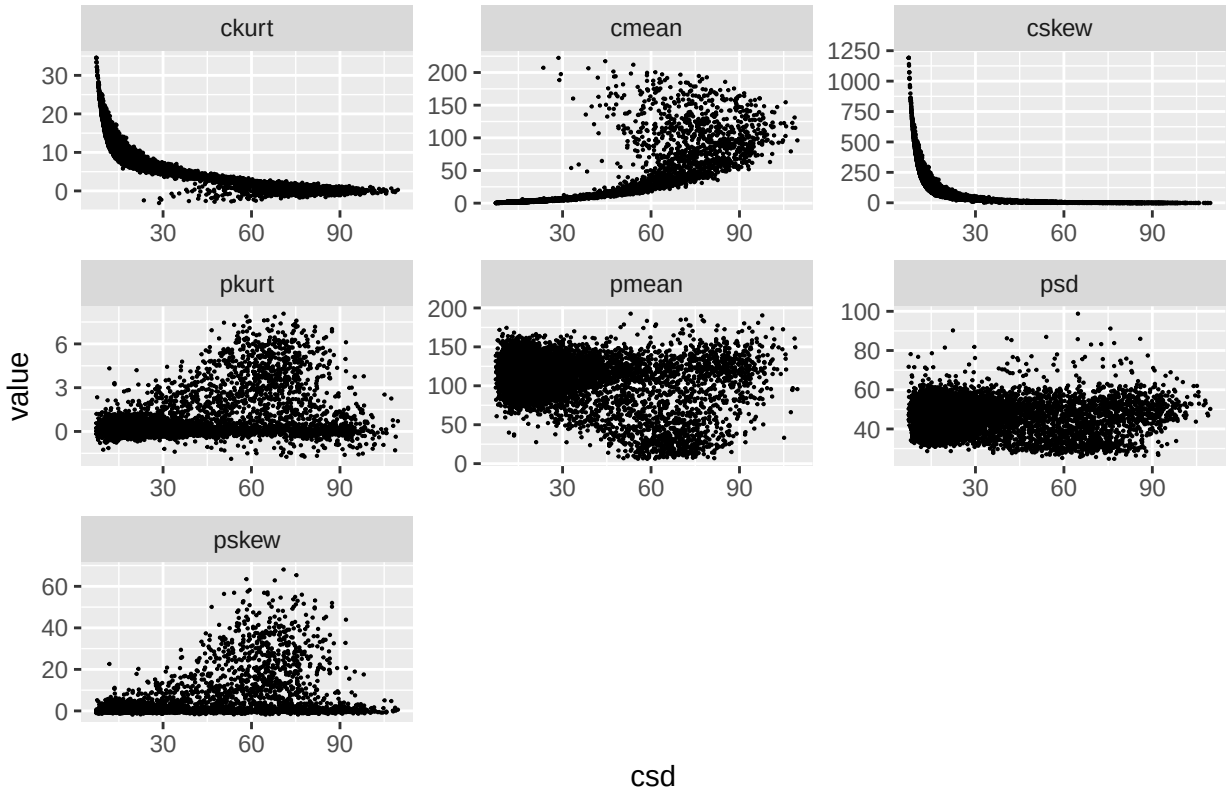
In consideration of the large amount of stars that will be observed in future studies, it was desirable to improve the computational efficiency of this algorithm by reducing the number of predictors. Predictors that are (possibly noisy) functions of other predictors, $x_j = f(x_i) + \epsilon$, are redundant and can be excluded without being detrimental to the performance of the algorithm. Visual methods were used to ascertain which predictors were functions of other predictors, and could be discarded.

Figure 1



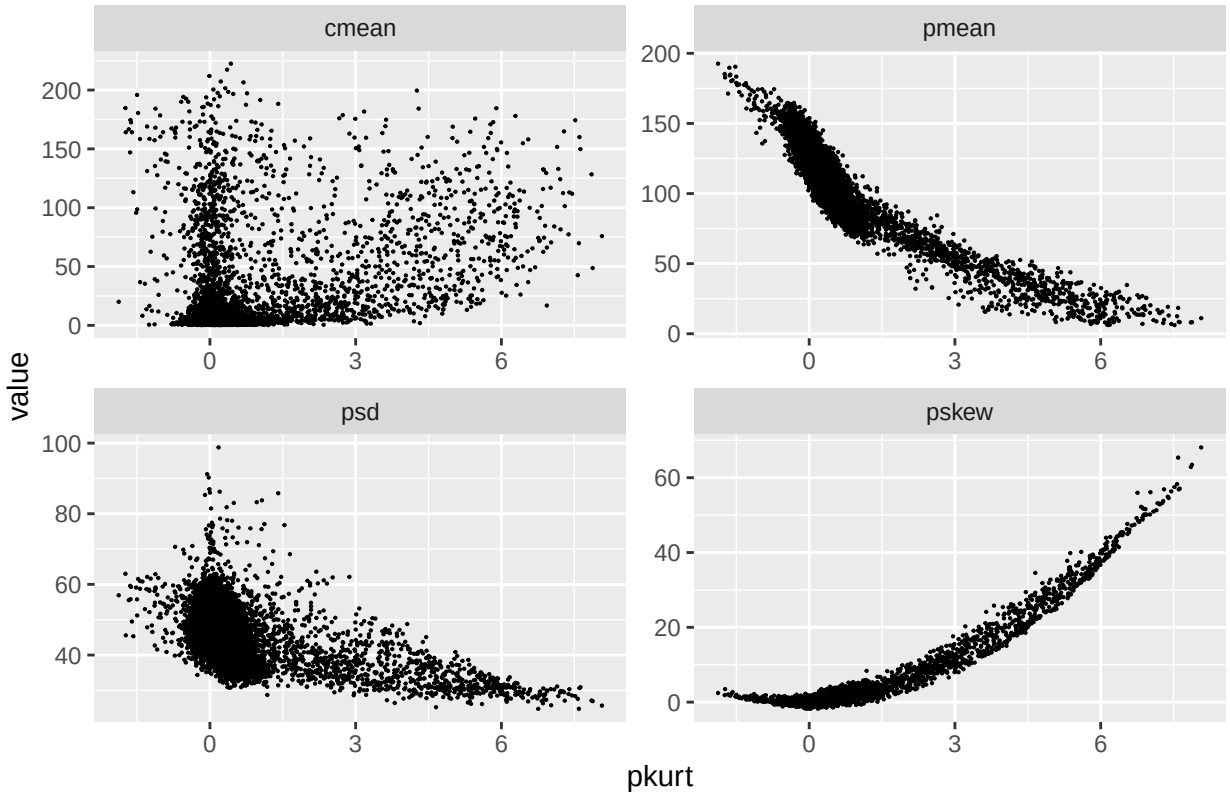
Boxplots were generated for each of the predictors and are shown above in Figure 1. Csd was the first predictor that was chosen to retain because, as shown above, Q_1 of 1 is greater than Q_3 of 0 . The data was then examined to determine which predictors were functions of csd and thus, could be excluded. Each of the predictors were plotted against csd and, as shown in Figure 2, cskew and ckurt both appear to be functions of csd and were removed.

Figure 2, Functions of CSD



Pkurt was then selected to retain because it also has the property that Q_1 of 1 is greater than Q_3 of 0. Each remaining predictors were then plotted against pkurt, shown in Figure 3 below, to identify predictors that were functions of pkurt. Pmean and pskew appeared to be functions of pkurt so they were also removed.

Figure 3



Finally, the remaining two predictors, cmean and psd, were plotted against each other. As shown in Figure 4 below, psd is not a function of cmean so they were both retained. The final predictors are summarized in Table 2 below.

Figure 4

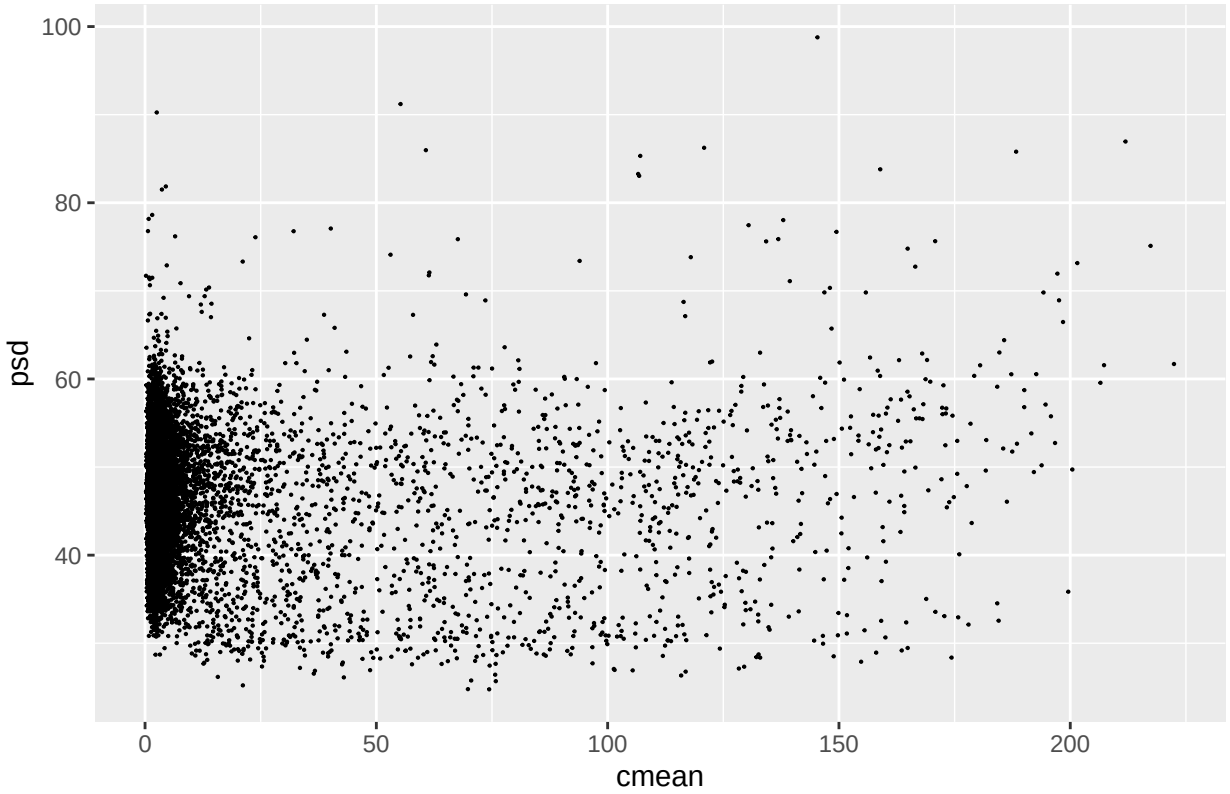


Table 2: Final Predictors

Predictors_final	Predictors_names_final
Standard deviation of the integrated profile.	psd
Excess kurtosis of the integrated profile.	pkurt
Mean of the DM-SNR curve.	cmean
Standard deviation of the DM-SNR curve.	csd

Algorithm Selection

Four popular machine learning algorithms were trained with the "train" partition and with different summary functions. To determine the optimal algorithm, bootstrap samples were created from the "train" partition, the algorithms were then applied to these bootstrap samples, and the relevant average metrics were recorded (see table 3 below). The algorithms that were tested were Random Forest, KNN, Naive Bayes, and Negative Binomial Generalized Linear Mode labeled Rborist, knn, nb, and glm respectively. Note that the algorithms labeled "no metric" were trained with the default metric, accuracy. Sensitivity was the metric that was chosen to select the algorithm because stars that are identified as pulsars by the algorithm will be further investigated so it is more important for the algorithm to identify potential pulsars than it is for it to reject stars that are not pulsars. The algorithm that had the highest predicted sensitivity was Rborist trained with beta = 0.8, this algorithm also had the highest predicted specificity.

Table 3: Algorithm Performance

sensitivity	specificity	fScore	beta	model
0.9969	0.9974	0.9986	0.8	Rborist
0.9968	0.9974	0.9974	2.0	Rborist Specificity
0.9967	0.8601	0.9908	0.9	Rborist
0.9966	0.8596	0.9913	1.0	Rborist Sensitivity
0.9965	0.9966	0.9988	0.6	Rborist
0.9964	0.9974	0.9995	0.3	Rborist
0.9963	0.9041	0.9933	1.0	Rborist no metric
0.9963	0.9974	0.9997	0.1	Rborist
0.9963	0.9974	0.9996	0.2	Rborist
0.9963	0.9974	0.9993	0.4	Rborist
0.9962	0.9974	0.9986	0.7	Rborist
0.9961	0.9966	0.9989	0.5	Rborist
0.9950	0.8224	0.9886	1.0	glm no metric
0.9950	0.8224	0.9824	0.1	glm
0.9950	0.8224	0.9827	0.2	glm
0.9950	0.8224	0.9833	0.3	glm
0.9950	0.8224	0.9840	0.4	glm
0.9950	0.8224	0.9848	0.5	glm
0.9950	0.8224	0.9856	0.6	glm
0.9950	0.8224	0.9864	0.7	glm
0.9950	0.8224	0.9872	0.8	glm
0.9950	0.8224	0.9879	0.9	glm
0.9950	0.8224	0.9924	2.0	glm Specificity
0.9950	0.8224	0.9886	1.0	glm Sensitivity
0.9946	0.7346	0.9739	0.1	knn
0.9946	0.7346	0.9745	0.2	knn
0.9946	0.7346	0.9754	0.3	knn
0.9946	0.7346	0.9904	2.0	knn Specificity
0.9944	0.6863	0.9815	1.0	knn no metric
0.9944	0.6863	0.9756	0.6	knn
0.9944	0.6863	0.9772	0.7	knn
0.9944	0.6863	0.9788	0.8	knn
0.9944	0.6863	0.9802	0.9	knn
0.9944	0.6863	0.9815	1.0	knn Sensitivity
0.9943	0.6996	0.9736	0.4	knn
0.9943	0.6996	0.9750	0.5	knn
0.9863	0.8435	0.9852	1.0	nb no metric
0.9863	0.8435	0.9842	0.1	nb
0.9863	0.8435	0.9843	0.2	nb
0.9863	0.8435	0.9844	0.3	nb
0.9863	0.8435	0.9845	0.4	nb
0.9863	0.8435	0.9846	0.5	nb
0.9863	0.8435	0.9847	0.6	nb
0.9863	0.8435	0.9849	0.7	nb
0.9863	0.8435	0.9850	0.8	nb
0.9863	0.8435	0.9851	0.9	nb
0.9863	0.8435	0.9859	2.0	nb Specificity
0.9863	0.8435	0.9852	1.0	nb Sensitivity

Results

Using the Random Forest algorithm trained with Beta=0.8 a sensitivity of 0.993 was achieved when it was applied to the test set. As expected, it is slightly lower but still very close to the 0.997 predicted by the bootstrap estimation. This algorithm also had excellent specificity, 0.823 this indicates that the Random Forest algorithm is a suitable method to identify pulsars.

Conclusion

Multiple machine learning algorithms were tested for their suitability to identify pulsars. Redundant predictors were identified and excluded to improve computational efficiency. The optimal algorithm was found to be the Random Forest algorithm trained with a Beta=0.8; this algorithm provides exceptional sensitivity and specificity and uses only four of the original eight predictors. This is a scalable algorithm to identify pulsars because of the few predictors and excellent sensitivity. Future work on this problem may include the addition of different predictors, such as the standard deviation of the period of the profile, and the addition of more observations to train the algorithm.