Pulsar Star Prediction

MGS 8040, Data Mining

Brian, Sheu

Tsao-Chin, Yu

Tsung-Wei, Chen

Ya-Han, Chang

30 April 2019

Introduction

This report develops a model trying to predict the likelihood of one observation being Pulsar star. The dataset was collected during the High Time Resolution Universe Survey and describes characteristics of a sample of pulsar star candidates. We build a regression model to identify potential pulsar star in the universe. Each candidate could be potentially identified as a real pulsar without additional information. However, according to the publisher, in practice, "almost all detections are caused by radio frequency interference (RFI) and noise, making legitimate signals hard to find" and rendering outcomes of True Negative classification. Therefore, this report targets at assisting in accurately predicting real pulsar stars using analytical constructed model. Besides, this would be able to help facilitate rapid analysis significantly by automatically labeling pulsar candidates.

Data

The dataset was obtained from Kaggle. (https://www.kaggle.com/) It is a community to hold data science competitions, where publishers provide their data in an accessible format. The particular dataset we use is designed for "Predicting a Pulsar Star" project last updated on 2018-05-09. (https://www.kaggle.com/pavanraj159/predicting-a-pulsar-star) It describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey. (see [2] for sources). Pulsars are a rare type of Neutron star that produce radio emission detectable on Earth.

Overall, this dataset contains 16,259 spurious examples caused by RFI/noise (negative examples), and 1,639 real pulsar examples(positive examples). The observations would be 17,898 in total.

There are 9 variables within, including eight continuous variables and one categorial variable. The first four continuous variable are simple statistics drawn from integrated pulse profile. This describe a longitude-resolved version of the signal that has been averaged in both time and frequency, which is used to detect a real pulsar star. (see [1] for more its reference)

The remaining four continuous variables were obtained from the DM-SNR curve. (DM represents the "dispersion measure" and increases with distance and electron density between Earth and pulsar; SNR stands for the signal-to-noise ratio, increasing with integration time. Both are good indicators to distinguish pulsar from other kinds of stars.) The dependent variable here we use is the categorial variable "target_class," denoting whether the observation is predicted as a pulsar star, with (1) for pulsar star, and (0) for not a star. There are no missing values observed in this dataset.

(Further information please refer to appendix A and B for data dictionary and frequency table.)

Methodology

The process undertaken follows the traditional steps for logistic regression analysis.

1. Import and Examine the Data

The raw data was imported into R and Excel in the CSV format to ensure that each column was labeled with the appropriate variable names. Firstly, we understood the definition of each variable and identify their data types. (Appendix A. Data Dictionary)

A univariate analysis was then performed to find any missing, negative and unusual values. As a result, there are no missing values in all of the variables in this dataset, and no obvious outliers are found by looking at the box plot we drawn in R. We also constructed covariance matrix to examine overall data structure between variables. After that, we decided to include all eight variables in our initial version of model to start predicting the possibility of one being pulsar star.

2. Define dummies

After familiarizing ourselves with the data we did a 70/30 split of the data into a training and validation dataset. From there we created our crosstabs (frequency of each variable against the “target_class” variable), determined the bins, in which we decided to contain about 5% of the observations of each variable, and used those cutoffs to create a new format in SAS.

Figure 1. Format Creation Example

```
VALUE _Mean_of_the_DM_SNR_curve
0.213210702 - 0.994983278= “0.213210702 TO 0.994983278”
0.995819398 - 1.29264214= “0.995819398 TO 1.29264214”
1.293478261 - 1.530936455 = “1.293478261 TO 1.530936455”
1.531772575 - 1.740802676= “1.531772575 TO 1.740802676”
1.741638796 - 1.913879599= “1.741638796 TO 1.913879599”
```

However, instead of calculating the good-to-bad ratio, we computed bad-to-good ratio to make numbers bigger. Therefore, the trend would be more obvious and easier to identify when making breakpoints decisions. Finally, the neutral (baseline) groups and dummy breakpoints are picked by hand for each variable.

Figure 2. Dummy Creation Example

Table of _Standard_deviation_of_the_integ by target_class

_Standard_deviation_of_the_integ	target_class					
	0	1	Total	Ratio G/B	Ratio B/G	
24.79161196 TO 34.78722907	218	405	623	1.857798	0.538272	
34.79057654 TO 37.45911753	471	152	623	0.322718	3.098684	D1
37.45973017 TO 39.42365236	515	108	623	0.209709	4.768519	
39.42615724 TO 41.07531594	550	72	622	0.130909	7.638889	
41.07598758 TO 42.35697945	578	45	623	0.077855	12.84444	
42.35793985 TO 43.48870352	584	39	623	0.066781	14.97436	
43.49005083TO 44.42877189	588	35	623	0.059524	16.8	Neutral
44.4317309 TO 45.32689427	584	39	623	0.066781	14.97436	
45.3275938 TO 46.13574821	597	25	622	0.041876	23.88	
46.13667427TO 46.95866427	605	18	623	0.029752	33.61111	
46.96049495 TO 47.78708922	591	32	623	0.054146	18.46875	D2
47.78743152 TO 48.5422431	606	17	623	0.028053	35.64706	
48.54230597 TO 49.37093474	603	20	623	0.033167	30.15	
49.37232427 TO 50.14624547	605	17	622	0.028099	35.58824	
50.14686378 TO 51.04274918	613	10	623	0.016313	61.3	
51.04310791 TO 51.91867925	609	14	623	0.022989	43.5	
51.92064774 TO 52.93790584	610	13	623	0.021311	46.92308	
52.93858781 TO 54.39266449	617	6	623	0.009724	102.8333	D3
54.39281317 TO 56.42645068	613	9	622	0.014682	68.11111	
143.2578125 TO 98.77891067	596	27	623	0.045302	22.07407	Neutral
Total	11353	1103	12456	0.097155		

3. Build regression model

Once the dummy variables were created, the regression model was ready to be built. All the dummies for all variables are run together. We started with including all of the independent variables, eight in total. And by looking at the regression results, some of the p-values are so high that we decided to eliminate 3 variables in our model after a few times of iterations. Once the final model was determined, we continued to evaluate the parameter estimators. This step was to ensure that the estimators matched the behavior and logical sense that we had initially expected. All coefficients of the final model were seen to be meaningful and making sense in our case.

Besides, we also combined a few dummy variables with parameter estimates similar to neighboring ones to simplify the model. The regression output is displayed in Appendix C.

4. Score the model

To score all observations, we used scores formatted in a range of 0 to 1000. The scoring program is run for the training dataset in SAS. And the output was saved in a new file called “scrtrain.” After that, we conducted initial version of KS test and drew diagrams first, to ensure our results were reasonable.

Once the initial KS test had validated the model’s acceptable performance, these steps (creating dummies, applying model and scoring) were repeated for the validation data.

5. Complete Kolmogorov–Smirnov Test (KS Test)

The initial version of KS test table produced in step 4 was then completed in this step. By looking at the result of KS test, we found that the optimal cutoff score would be 100. And the validation data's KS test is completed afterwards, with the same results of a 100 optimal cutoff score.

The final KS test results are shown in the "Results" section.

6. Create the scorecard

Once the model was finalized the scorecard was created to make interpretation easier. Also, the trends are checked so that the estimators are ensured to be logical and making sense. (The sense of this particular profession is according to research paper in the "Reference" section)

The final scorecard is shown in the "Results" section.

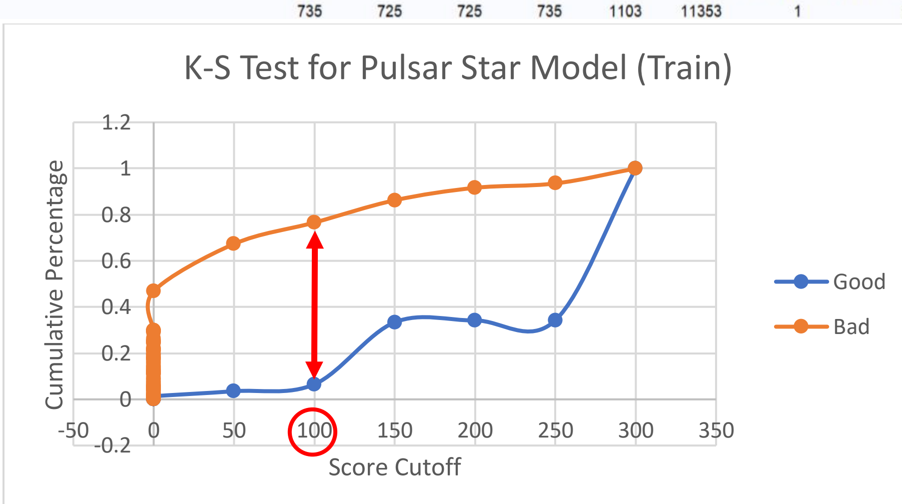
Results

1. KS Test Result – Training Data

The following is the KS Test for training dataset. We found that the optimal cutoff would be 100, by choosing the score with largest different value of 70.05% between good and bad. By adopting this score, 6.5% of the good candidates as well as 76.6% of the bad candidates would be included. We also observed that 0.58% of total candidates are predicted to be a pulsar star by the model.

Table of bgscore by target_class

bgscore	target_class 0	1	Total			0	1 Good	Bad	Cumulative Good	Cumulative Bad		Cumulative Percent Good	Cumulative Percent Bad	Difference
-0.16042707	1	0	1	-0.160427	-0.160427	1	0	0	1	0	1	0	8.80824E-05	-8.80824E-05
-0.15954805	1	0	1	-0.159548	-0.159548	1	0	0	1	0	2	0	0.000176165	-0.000176165
-0.00458049	3	0	3	-0.00458	-0.00458	3	0	0	3	9	3379	0.00816	0.297630582	-0.289471017
-0.00366563	2	0	2	-0.003666	-0.003666	2	0	0	2	9	3381	0.00816	0.297806747	-0.289647182
...														
0						1934	6	6	1934	15	5315	0.013599	0.468158196	-0.454558921
50						2327	24	24	2327	39	7642	0.035358	0.673126046	-0.637767932
100						1052	33	33	1052	72	8694	0.065277	0.765788778	-0.70051226
150						1095	296	296	1095	368	9789	0.333636	0.862239056	-0.528603516
200						614	9	9	614	377	10403	0.341795	0.916321677	-0.574526573
250						215	1	1	215	378	10618	0.342702	0.935259403	-0.59255768
300						735	725	725	735	1103	11353	1	1	0

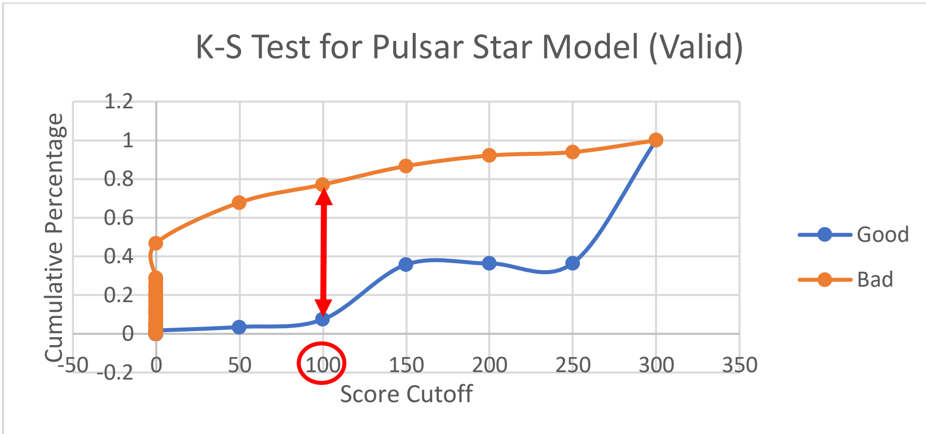


2. KS Test Result – Validation Data

The KS Test shown below is for validation dataset. We found that the optimal point here is also the cutoff score 100, based on the largest KS difference value of 69.66%. By adopting this score, 7.4% of the good candidates as well as 77.1% of the bad candidates would be included. Also, we observed that 0.7% of total candidates are predicted to be a pulsar star by the model.

Table of bgscore by target_class

bgscore	target_class									Cumulative		Cumulative Percent		
	0	1	Total			0	1	Good	Bad	Good	Bad	Good	Bad	Difference
-0.15088491	5	0	5	-0.150885	-0.150885	5	0	0	5	0	5	0	0.001019	-0.001019
-0.15000589	10	0	10	-0.150006	-0.150006	10	0	0	10	0	15	0	0.003057	-0.003057
-0.00756512	9	0	9	-0.007565	-0.007565	9	0	0	9	6	1409	0.011194	0.287199	-0.276005
-0.00639739	1	0	1	-0.006397	-0.006397	1	0	0	1	6	1410	0.011194	0.287403	-0.276209
...														
0					0	891	3	3	891	9	2301	0.016791	0.469018	-0.452226
50						1029	10	10	1029	19	3330	0.035448	0.678761	-0.643313
100						454	21	21	454	40	3784	0.074627	0.7713	-0.696674
150						470	152	152	470	192	4254	0.358209	0.867102	-0.508893
200						266	3	3	266	195	4520	0.363806	0.921321	-0.557515
250						88	1	1	88	196	4608	0.365672	0.939258	-0.573586
300						298	340	340	298	536	4906	1	1	0



3. Scorecard

Below is the explained results (estimators) of our final regression. The only variable that is positively impacting the possibility of being a pulsar star is “Profile Standard Deviation.” Others, including “Profile Mean,” “Profile Excess Kurtosis,” “Profile Skewness” and “DMSN Standard Deviation,” are affecting the result in a negative manner.

However, within those of negative impact, Profile Mean, Profile Excess Kurtosis and DMSN Standard Deviation are increasingly negative while Profile Skewness is decreasingly negative. This should be noticed when interpreting the result.

Figure 3. Scorecard

Variable	Range	Points
Intercept		174
Profile Mean	< 117	0
	117 to 127	-70
	> 127	-94
Profile Standard Deviation	<41	172

	41 to 45	0
	45 to 50	0.8
	>50	4.8
Profile Excess Kurtosis	<0.06	0
	0.06 to 0.07	-32
	0.07 to 0.30	-71
	>0.30	0
Profile Skewness	<0.18	-46
	0.18 to 0.20	-55
	>0.20	0
DMSN Standard Deviation	<13.69	-112
	13.69 to 14.38	0
	14.38 to 16.56	-114

4. Discussion of improvement

There are multiple techniques to identify pulsar star; however, we barely have the chance to have a confident model that only uses a few factors of a more complicated phenomenon in the universe.

Our client will be able to identify the real pulsars that are of significant scientific interest as probes of space-time, the inter-stellar medium, and states of matter. Our model can also be used to automatically label pulsar candidates to facilitate rapid analysis.

However, we would like to mention that we have run an additional regression (see appendix D for additional regression output) to build another version of model without using dummies. The output of the additional regression appears to be more perfect than our original one, which separates variables into dummies. And it does not require us to drop variables by looking at the p-values. (The KS test result for regression without dummies is also shown in Appendix D.)

This can be taken into consideration that the model may actually be able to run without using dummies in this case.

Implementation

Conclusion

In conclusion, we recommend clients to score each pulsar star data with scored built from our model.

As mention earlier, we have dropped some of the insufficient variables due to the large p-value. As result, we identified 5 variables from 8 original variables.

1) _Mean_of_the_integrated_profile
2) _Standard_deviation_of_the_integ
3) Excess_kurtosis_of_the_integrat
4) Skewness_of_the_integrated_prof
5) _Standard_deviation_of_the_DM_SN

* See appendix A for variable descriptions

Score Cutoff

According to our model, the best cutoff is score point 100 , with a KS difference of -69.67%. Hence, the score points of over 100 should be classified as a pulsar star.

Cost

Since we do not know the cost, so the best way to use the strategies is Global Classification Rate:

$$\text{Global Classification Rate} = \frac{(True\ Positive + True\ Negative)}{Total\ Oberservation}$$

But we still recommend the clients know the cost of misclassification for a non-pulsar star, since the error is tremendous.

Monitoring Reports

The performance of the model has to be monitored to ensure it remains effective. Each pulsar produces a slightly different emission pattern, which varies slightly with each rotation. Thus, a potential signal detection known as a 'candidate', is averaged over many rotations of the pulsar, as determined by the length of an observation. In the absence of additional information, each candidate could potentially describe a real pulsar. However, in practice almost all detections are caused by radio frequency interference (RFI) and noise, making legitimate signals hard to find.

Therefore, it is important to be able to separate the non-pulsars from the pulsars. We can do so by examining the differences between the Expected Score Distribution, as predicted by our model, and the Actual Score Distribution, as observed in the future. Given that stars usually survive much longer than the human time frame, their characteristics probably will not change by much within the foreseeable future. Therefore, adjustment to the algorithm for the existing variables will not be necessary.

However, scientific advancement may allow us to collect more sophisticated evidences that describe a real pulsar. Therefore, addition of new variables will be the primary modification to our model. Based on the afore-mentioned assumptions, we recommend a not-so-often evaluation of the existing model. A significant number of misclassifications of non-pulsars or real pulsars should also trigger the use of this report.

Figure 4. Monitoring Report - Actual vs. Expected Score Distribution (Partial)

Score Range	Expected Score Distribution	Actual Score Distribution	Difference
>0	1.36%		
>50	3.54%		
>100	6.53%		
>150	33.36%		
>200	34.18%		
>250	34.27%		
>300	100.00%		
…	…		

Once the expected versus observed differences are calculated for each score range, then it should be determined if they are statistically significant. The minimum required difference at a 95% confidence level has to be determined. If all of the differences are below this number, then the fluctuations are among what is expected and are insignificant.

Project Flow Diagram

Figure 5. Process flow Diagram

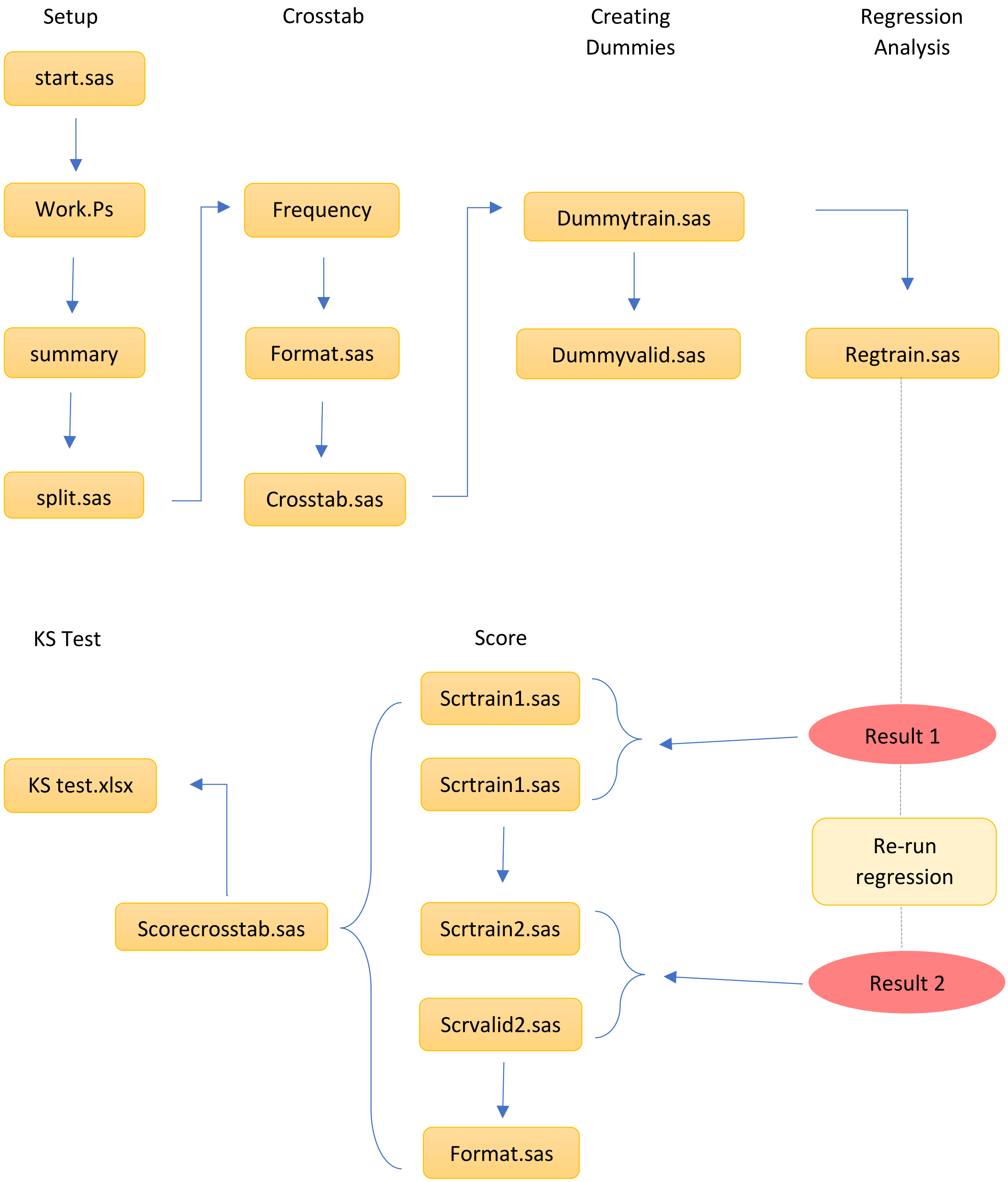


Figure 6. Process Flow Chart

	Step	Description	Input Files	Out Files
1	Start.sas	Assign SAS library	N/A	N/A
2	Work.PS	Input pulsar_ star dataset	pulsar_star.csv	work.Ps
3	summary	Check dataset	work.ps	N/A
4	split.sas	Spilt dataset into training(30%) and validation(70%)	work.ps	train.sas valid.sas
5	Frequency	Create frequency table for each variable	train.sas	FrequencyTable.xlsx
6	Format.sas	Create format library for variables (5% separation)	N/A	N/A
7	Crosstab.sas	create crosstab using format against variable: target_class	train.sas	FrequencyTable2.xlsx
8	Dummytrain.sas	Create dummy in training dataset	Train.sas	Train2.sas
9	Dummyvalid.sas	Apply same dummy from train dataset	valid.sas	Valid2.sas
10	Regtrain.sas	Run regression on all dummy variables for training dataset	train2.sas	regression1.xlsx
11	Re-run regression	Drop any variables until accepting the model	train2.sas	regression2.xlsx
12	Scrtrain1.sas	Create frequency table for original scores of each variable using result 10 (step 10) on training dataset	train2.sas	scrtrain1.html
13	Scrvalid1.sas	Create frequency table for original scores of each variable using result 10 (step 10) on valid dataset	valid2.sas	scrtrain1.html
14	Scrtrain2.sas	Create frequency table for final scores of each variable using result11 (step11) on training dataset	train2.sas	scrtrain1.html
15	Scrvalid2.sas	Create frequency table for final scores of each variable using result11 (step11) on valid dataset	valid2.sas	scrtrain1.html
16	Format.sas	create format library for bgscore	N/A	N/A
17	Scorecrosstab.sas	Create crosstab for each score set against target_class using bgscrore (comparison of scores crosstab before and after dropping variables)	N/A	scrtrain1.html scrvalid1.html scrtrain2.html scrtvalid2.html
18	KStest.xlsx	Conduct KS test using the result from step 17 in excel	scrtrain1.html scrvalid1.html scrtrain2.html scrvalid2.html	KStest.xlsx

Appendix A.

Data Dictionary

Alphabetic List of Variables and Attributes		
Variable	**Type**	**Description**
_Excess_kurtosis_of_the_DM_SNR_c	Num	Excess kurtosis of the DM SNR curve
_Excess_kurtosis_of_the_integrat	Num	Excess kurtosis of candidate’s profile
_Mean_of_the_DM_SNR_curve	Num	Mean of the DM-SNR curve
_Mean_of_the_integrated_profile	Num	Mean of candidate’s profile
_Skewness_of_the_DM_SNR_curve	Num	Skewness of the DM-SNR curve
_Skewness_of_the_integrated_prof	Num	Skewness of candidate’s profile
_Standard_deviation_of_the_DM_SN	Num	Standard deviation of the DM-SNR curve
_Standard_deviation_of_the_integ	Num	Standard deviation of candidate’s profile
target_class	Categorial	Class of pulsar star: 1 for pulsar star,0 for not a star

Note:

For the DM_SNR curve, DM represents the "dispersion measure" and increases with distance and electron density between Earth and pulsar; SNR stands for the signal-to-noise ratio, increasing with integration time. Both are good indicators to distinguish pulsar from other kinds of stars. (see [3] for sources)

Appendix B.

Frequency Table Example

The FREQ Procedure

_Mean_of_the_integrated_profile	Frequency	Percent	Cumulative Frequency	Cumulative Percent
5.8125	1	0.01	1	0.01
6.1875	1	0.01	2	0.02
6.265625	1	0.01	3	0.02
6.5	1	0.01	4	0.03
6.9375	1	0.01	5	0.04
6.984375	1	0.01	6	0.05
7.0625	1	0.01	7	0.06
7.4609375	1	0.01	8	0.06
7.6328125	1	0.01	9	0.07
7.796875	1	0.01	10	0.08
7.921875	1	0.01	11	0.09
8.1015625	1	0.01	12	0.10
8.109375	1	0.01	13	0.10
8.15625	1	0.01	14	0.11
8.1953125	1	0.01	15	0.12
8.2265625	1	0.01	16	0.13
8.2421875	1	0.01	17	0.14
8.25	1	0.01	18	0.14
8.2734375	1	0.01	19	0.15
8.3515625	1	0.01	20	0.16
8.75	1	0.01	21	0.17
8.84375	1	0.01	22	0.18
8.875	1	0.01	23	0.18
9.046875	1	0.01	24	0.19
9.234375	1	0.01	25	0.20
9.3359375	1	0.01	26	0.21
9.6796875	1	0.01	27	0.22
9.7421875	1	0.01	28	0.22

Appendix C.

1. First Regression Output from SAS

Model: bgscore
Dependent Variable: target_class
Number of Observations: 12456

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	310.89957	17.27220	309.34	<.0001
Error	12437	694.42791	0.05584		
Corrected Total	12455	1005.32747			

Root MSE	0.23630	R-Square	0.3093
Dependent Mean	0.08855	Adj R-Sq	0.3083
Coeff Var	266.84487		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.06713	0.00674	9.96	<.0001
Meanoftheintegratedprofile1	1	-0.06546	0.00599	-10.92	<.0001
Meanoftheintegratedprofile2	1	-0.09486	0.00569	-16.67	<.0001
Standarddeviationoftheinteg1	1	0.15601	0.00673	23.20	<.0001
Standarddeviationoftheinteg2	1	0.00637	0.00609	1.05	0.2953
Standarddeviationoftheinteg3	1	0.00641	0.00712	0.90	0.3678
Excesskurtosisoftheintegrat1	1	-0.04121	0.03021	-1.36	0.1726
Excesskurtosisoftheintegrat2	1	-0.05906	0.00494	-11.96	<.0001
Skewnessoftheintegratedprof1	1	-0.04925	0.00616	-7.99	<.0001
Skewnessoftheintegratedprof2	1	-0.05252	0.02394	-2.19	0.0283
MeanoftheDMSNRcurve1	1	0.00352	0.01075	0.33	0.7431
MeanoftheDMSNRcurve2	1	0.02373	0.01213	1.96	0.0504
Standard_deviationoftheDMSN1	1	-0.00497	0.00660	-0.75	0.4515
Standard_deviationoftheDMSN2	1	-0.00519	0.00715	-0.73	0.4684
ExcesskurtosisoftheDMSNRc1	1	0.20326	0.01406	14.46	<.0001
ExcesskurtosisoftheDMSNRc2	1	-0.00187	0.00958	-0.20	0.8449
ExcesskurtosisoftheDMSNRc3	1	0.00012766	0.00960	0.01	0.9894
SkewnessoftheDMSNRcurve1	1	-0.00176	0.01323	-0.13	0.8943
SkewnessoftheDMSNRcurve2	1	-0.00942	0.00759	-1.24	0.2145

2. *Last Regression output from SAS*

Model: bgscore
Dependent Variable: target_class
Number of Observations: 12456

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	213.39229	19.39930	304.83	<.0001
Error	12444	791.93518	0.06364		
Corrected Total	12455	1005.32747			

Root MSE	0.25227	R-Square	0.2123
Dependent Mean	0.08855	Adj R-Sq	0.2116
Coeff Var	284.88390		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.17417	0.00565	30.85	<.0001
Meanoftheintegratedprofile1	1	-0.07040	0.00639	-11.01	<.0001
Meanoftheintegratedprofile2	1	-0.09394	0.00607	-15.47	<.0001
Standarddeviationoftheinteg1	1	0.17285	0.00717	24.12	<.0001
Standarddeviationoftheinteg2	1	0.00087902	0.00650	0.14	0.8924
Standarddeviationoftheinteg3	1	0.00478	0.00760	0.63	0.5294
Excesskurtosisoftheintegrat1	1	-0.03159	0.03225	-0.98	0.3273
Excesskurtosisoftheintegrat2	1	-0.07132	0.00526	-13.56	<.0001
Skewnessoftheintegratedprof1	1	-0.04577	0.00658	-6.96	<.0001
Skewnessoftheintegratedprof2	1	-0.05531	0.02555	-2.16	0.0304
Standard_deviationoftheDMSN1	1	-0.11221	0.00580	-19.36	<.0001
Standard_deviationoftheDMSN2	1	-0.11403	0.00649	-17.58	<.0001

Appendix D.

1. *Additional Regression Output from SAS (without using dummies)*

Model: bgscore
Dependent Variable: target_class

Number of Observations Read	12456
Number of Observations Used	12456

Analysis of Variance

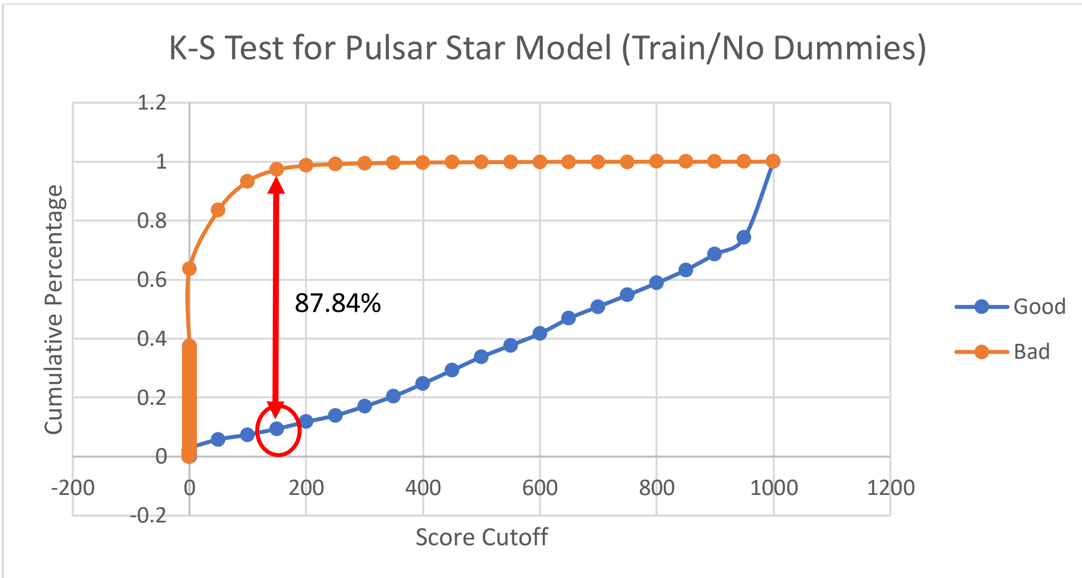
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	684.95443	85.61930	3326.45	<.0001
Error	12447	320.37304	0.02574		
Corrected Total	12455	1005.32747			

Root MSE	0.16043	R-Square	0.6813
Dependent Mean	0.08855	Adj R-Sq	0.6811
Coeff Var	181.17519		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.35620	0.02455	-14.51	<.0001
_Mean_of_the_integrated_profile	1	0.00325	0.00015126	21.50	<.0001
_Standard_deviation_of_the_integ	1	-0.00198	0.00028117	-7.03	<.0001
_Excess_kurtosis_of_the_integrat	1	0.41263	0.00739	55.80	<.0001
_Skewness_of_the_integrated_prof	1	-0.02830	0.00093980	-30.12	<.0001
_Mean_of_the_DM_SNR_curve	1	-0.00096003	0.00009273	-10.35	<.0001
_Standard_deviation_of_the_DM_SN	1	0.00302	0.00020588	14.65	<.0001
_Excess_kurtosis_of_the_DM_SNR_c	1	-0.00798	0.00196	-4.07	<.0001
_Skewness_of_the_DM_SNR_curve	1	0.00028158	0.00005992	4.70	<.0001

2. Additional Regression (without using dummies) – KS Test Result



Reference

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

[2] R. J. Lyon, 'PulsarFeatureLab', 2015

[3] http://www.jb.man.ac.uk/distance/frontiers/pulsars/section4.html