ME315: Machine Learning in Practice

London School of Economics (LSE): Summer School

Predicting Pulsars with Machine Learning: A Cost-Sensitive, Cluster-then-Classify Approach

Nicholas Roy

07/07/2022

Abstract

The classification of pulsars comes with many challenges. Proper radio signal metrics are needed to support the classification pipelines that flow from them. This analysis discusses some of these challenges to contextualize supervised (binary classification) and unsupervised (clustering) machine learning approaches to the Pulsar classification problem, as well as several metrics that can be used to evaluate them. It emphasizes a cost-sensitive approach, weighing sensitivity and specificity with the relative costs of false negatives and false positives. Finally, it combines the highest-performing clustering and classification models into two Cluster-then-Classify models for different cost contexts. Both combined models have higher accuracy than their stand-alone counterparts, due to synergies between features and cluster classes born from unknown patterns. The analysis end with a motivation for ongoing pulsar classification.

Table of Contents

"The isolation spins its mysterious cocoon, focusing the mind on one place, one time, one rhythm - the turning of the light."

-M.L. Stedman, The Light Between Oceans

The Hum of Dead Stars

There is no celestial object yet discovered that spin with greater ferocity than a pulsar. In fact, these mysterious "lighthouses" have been discovered to spin up to 716 times per second—nearly twice as fast as the blades of a household blender (Chauta, 2017). Scientists believe these astronomical speeds are the result of centrifugal acceleration post-Supernova; a red giant collapses, its matter becomes concentrated closer to the axis of rotation, and the rotational velocity of the resulting neutron star increases. Once a magnetized neutron star reaches speeds of a rotation between every 20 seconds to every 1-10 milliseconds, its whirring magnetic poles emit large amounts of energy in a range of waveforms, from radio to gamma (Dunbar, 2019). If a radio telescope or probe can detect this sweeping emission beam, it does so as pulses with smooth phases.

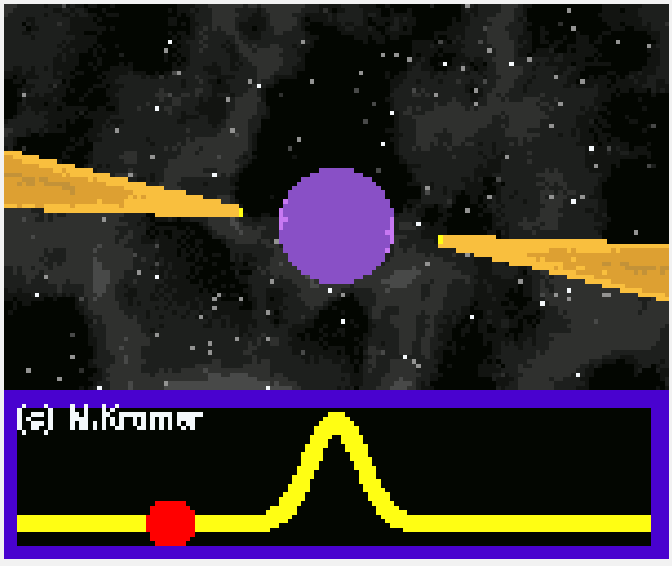


Fig. 1.1: ("Pulsar", Wikipedia)

It is a neutron star's rapid rotation and electromagnetic profile that drives its discovery and classification as a pulsar. Nevertheless, there are many challenges along the paths of their detection and classification.

Flickers From Beyond?

If a pulsar's radio signal has survived its long journey to Earth and is detected by a radio telescope, separating the "wheat from the chaff" - the signal from the noise -presents no easy task.

Large numbers of false, but convincing pulsar-like signals emanate from man-made and cosmic origins. They distort the classification of legitimate pulsar emission beams, increasing the risk of false positives (Lyon, 2016). Looking to other radio frequencies may offer clues surrounding the signals' origin, but the underlying waves' varying speeds introduces uncertainty via delays across channels. On its voyage, each pulse meets different cosmic and atmospheric conditions; consequently, individual pulses or phases from the same source can vary significantly in intensity and duration over the detection period (as9736, 2019). Metrics such as the DM-SNR ("dispersion-measure-signal-to-noise-ratio", Fig. 2.1 & 2.2) are employed to average the signal across radio frequency and time, and the Integrated Profile (Fig. 2.3) is another attempt to average varying phase signals within a detection event (HTRU, Planck). Powerful classification procedures and pipelines proliferate from the formalization of pulsar metrics (Lyon, 2016).

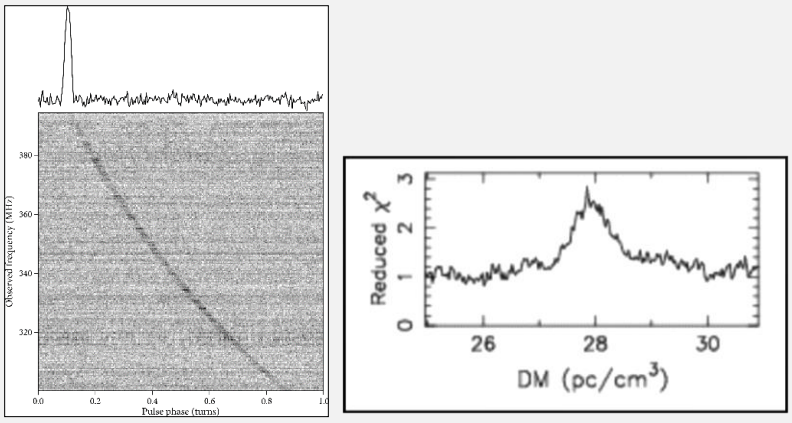


Fig. 2.1 & 2.2: DM-SNR (Chapter 6 Pulsars, NRAO)

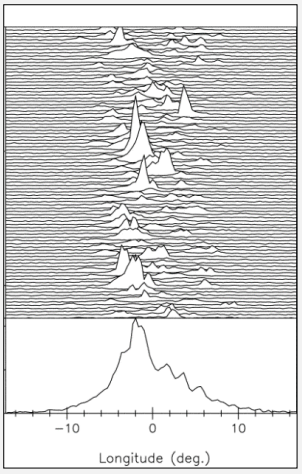


Fig. 2.3: Integrated Profile (as9736, 2019)

The most successful classification campaigns result from creative combinations of powerful radio telescopes, signal processing algorithms, and human ingenuity (Lyon, 2016). In more recent years, Machine Learning techniques ranging from the Supervised to Unsupervised have proven to be particularly useful (R.J. Lyon et al). After all, there are many candidate signals, and the question of causality is of little importance. The ultimate performance of these correlation-based predictions depends on the peculiarities of and interactions between the methodologies (of the chosen evaluation metric and prediction approach) and the data.

The Tool and the Substrate

The following analysis applies a variety of Machine Learning tools to a pulsar classification dataset sourced from the UCI repository (HTRU2, UCI Rep.) and originating from the second HTRU ("High Time Resolution Universe") survey conducted to detect and classify pulsars (HTRU, Planck). Each pulsar candidate is described according to 8 real-valued characteristics: 1) Mean of the integrated profile 2) Standard deviation of the integrated profile 3) Excess kurtosis of the integrated profile 4) Skewness of the integrated profile 5) Mean of the DM-SNR curve 6) Standard deviation of the DM-SNR curve 7) Excess kurtosis of the DM-SNR curve 8) Skewness of the DM-SNR curve 9) as well as, a binary "class" variable (1=pulsar; 0=not pulsar). The data set is highly imbalanced, with 16,259 spurious examples caused by noise, and only 1,639 confirmed pulsar examples, totaling 17,898 candidate instances. As this between-class imbalance is intrinsic to the dataset, truncating and resampling instances diminishes the utility of predictions for real pulsar classifications. Simple classification accuracy is not meaningful for pulsar classification (Lyon, 2016), as variance in the minority set will tend to be larger due to the fewer datapoints, and the number of false positives will be relatively high. If the cost of classifying a background or noise event as a pulsar is greater than that of classifying a true pulsar signal as noise, then a more balanced approach is needed. Asymmetric costs may take the form of additional resources needed to verify the signal or setbacks to the scientific body of knowledge surrounding pulsars. Conversely, there may be large opportunity costs associated with failing to classify a true pulsar signal, although they are much harder to quantify or estimate with certainty.

As such, the analysis employs a cost-sensitive accuracy metric to compare the efficacy of several models: LDA, QDA, Logistic Regression, and KNN, at various assumed relative costs of false positive to false negative classification. Then, it compares the results of these simple classifications with those of a "cluster-then-classify" approach, with the goal of capturing the interactions of each of the attributes with more holistic binary cluster classifications.

Exploratory Data Analysis: HTRU2

An exploration of this data set begins with combining the preset training and test split (43-57) and reshuffling the candidates to mitigate sampling bias when a custom split is later made. A glance at the distribution of NA values shows a large number in the class feature "is_pulsar" (Fig. 3.1). This is likely due to the candidates' failure to meet a certainty threshold. All are removed from the dataset, introducing the potential for some sampling bias if NA values are systematically produced, in exchange for greater predictive performance.

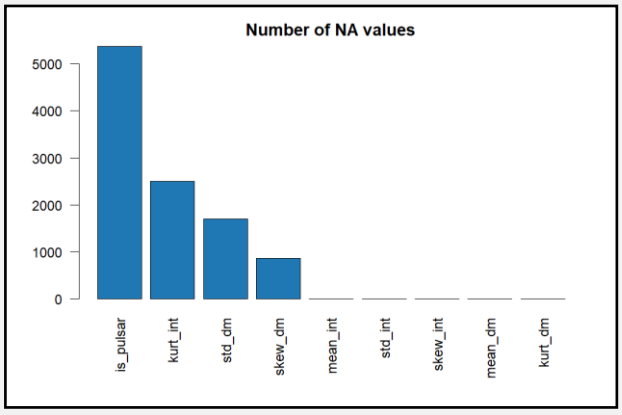


Fig. 3.1

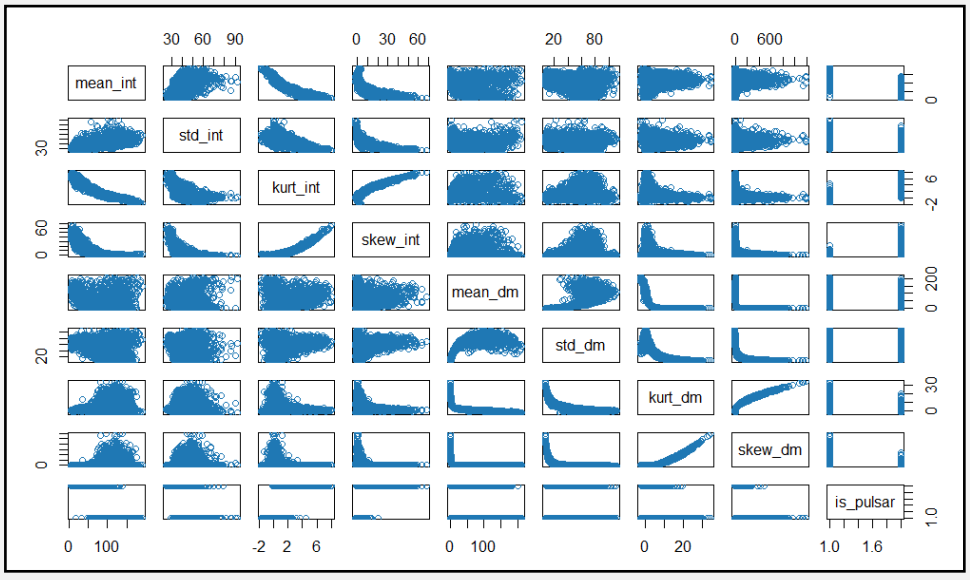


Fig. 3.2

Scatterplots of all variable combinations (Fig. 3.2) reveal a high degree of nonlinear multicollinearity, particularly amongst kurt_dm, skew_dm, skew_int, and kurt_int. Variables such as the kurtosis and skew of the integrated profile or DM-SNR curve are positively correlated—one is more likely to witness a positive skewed candidate that also has a light tail. Correlations of variables across the two radio signal metrics appear to have weaker nonlinear correlations, as the Integrated Profile and DM-SNR curves characterize pulsar signals from slightly different angles— across time only, and across radio frequency and time. Nevertheless, all features have an absolute correlation with the class greater than 0.35. They will be retained in the dataset to aid in this omnidirectional prediction task. Individual histograms of the features (Fig. 3.4) also reveal significant nonnormality of the kurt_int, skew_int, mean_dm, std_dm, and skew_dm features. Taking the logarithm of these variables significantly improves their normality (Fig 3.5) according to the Shapiro-Wilk test and likely enhances the predictive power of LDA and QDA. Kurt_int, skew_int, and skew_dm take on negative values for some candidates, so their respective minimums are added alongside a small constant (a=0.001) to ensure no NA values are produced in the log transformation.

Fig. 3.4

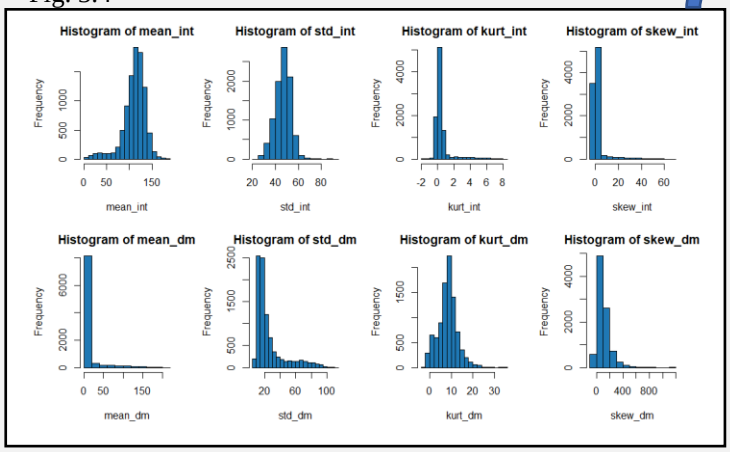
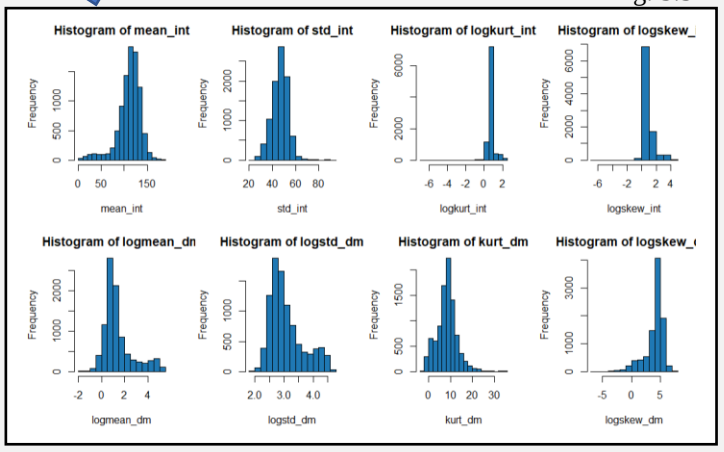


Fig. 3.5



Box plots of the features 8 features with respect to the class (Fig. 3.6) confirm the impact of the data set's imbalance: higher variance for positive classifications. Additionally, the relative positions of the medians, lower quartiles, and upper quartiles for the pulsar class are mirrored across the variables derived from the Integrated Profile and DM-SNR, respectively. For instance, the median of the standard deviation of the Integrated Profile is lower for positive pulsar

classifications, while that of the DM-SNR curve is higher— an indicator that true pulsars have on average more noise across radio frequency and time than time, alone, perhaps due to their broad emission spectra.

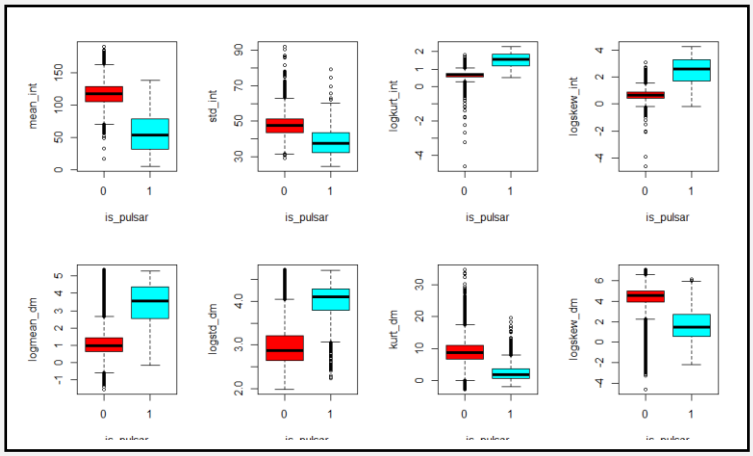


Fig. 3.6

Binary Classification

After cleaning, exploration, and transformation, the data is split into training and testing subsets. A 60:40 (training:test) split is employed, slightly higher than the preset 57:43. The reasoning behind this minor adjustment being, simply, to round to a cleaner split and reduce the number of false positives, due to smaller variance of the positive class in the training set relative to the test set, at the cost of a slight increase in the number of false negatives. Several models are applied to the training dataset and evaluated with the test dataset: LDA, QDA, Logistic, and KNN. For the Logistic model, an optimal threshold is found with a "for" loop, and the optimal K for the KNN model is found similarly.

The models are evaluated with various accuracy metrics: simple accuracy, the proportion of all classifications that the model correctly predicted, and balanced accuracies according to various relative costs of false positives. These balanced accuracies are computed by taking the weighted average of the model's sensitivity and specificity with respect to relative costs of FP to FN of three different orders of magnitude (1:1, 10:1, 100:1). For instance, if the relative cost of FP is 10 times that of FN (10:1), then sensitivity is underweighted, and specificity is overweighted. In other words, there is a larger emphasis on the model's ability to avoid false positives in the balanced accuracy metric. In the context of these relative costs, models that classify many false signals as pulsars will have a low accuracy, in percent. The balanced accuracy metric is formalized, as follows:

$$Cost := \boldsymbol{c} = \begin{bmatrix} Cost_{FP} \\ Cost_{FN} \end{bmatrix}$$

$$CostWeights := \boldsymbol{w} = \begin{bmatrix} \dfrac{Cost_{FP}}{Cost_{FP} + Cost_{FN}} \\ \dfrac{Cost_{FN}}{Cost_{FP} + Cost_{FN}} \end{bmatrix}$$

$$Accuracy := \boldsymbol{a} = \begin{bmatrix} Specificity \\ Sensitivity \end{bmatrix} = \begin{bmatrix} \dfrac{TN}{FP + TN} \\ \dfrac{TP}{TP + FN} \end{bmatrix}$$

$$BalancedAccuracy := \boldsymbol{b} = \boldsymbol{w} \cdot \boldsymbol{a}$$
$$= \frac{Cost_{FP}}{Cost_{FP} + Cost_{FN}} \left(\frac{TN}{FP + TN} \right) + \frac{Cost_{FN}}{Cost_{FP} + Cost_{FN}} \left(\frac{TP}{TP + FN} \right)$$

Finally, the Area Under ROC curve is also computed to show the model which optimizes the trade-off between sensitivity and specificity, albeit with little consideration for their absolute levels. Here, a score of 100% indicates a perfect discrimination ability, maximizing sensitivity and specificity at every possible decision threshold (or their interpolated connecting lines). Scores of 50% are akin to random guessing. Models with generally high sensitivity and low specificity may be deemed equally fit as those with low sensitivity and high specificity. The underlying cost structure of FP and FN may make the former much less feasible, in practice, and may select for the latter according to the balanced accuracy metric. To evaluate a model using information from both Balanced Accuracy and AUC, one might consider taking their average.

Binary Classification Model Accuracy

Model/Accuracy	**Simple Accuracy (%)**	**1:1 (FP:FN cost)(%)**	**10:1 (FP:FN cost)(%)**	**100:1 (FP:FN cost)(%)**	**Area Under ROC (AUC)(%)**
LDA	97.898	90.620	83.310	81.863	90.62
QDA	97.897	94.927	91.942	91.351	94.93
Logistic	98.329	94.381	90.416	89.631	94.38
KNN	98.005	92.376	86.722	85.602	92.38

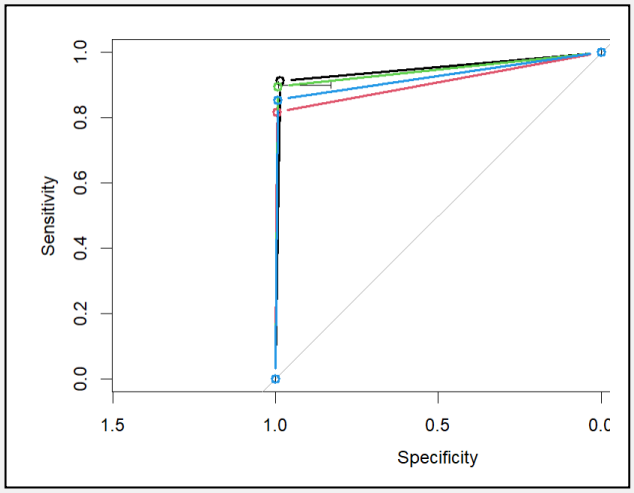


Fig. 4.1

As the FP:FN cost ratio increases, the balanced accuracy of all models tends towards their specificity. Although the Logistic regression with a threshold of 0.4242 has the highest proportion of correct classifications, QDA is slightly more robust in terms of balanced accuracy and AUC, indicating a more efficient specificty-sensitivity trade-off. This could be due to the fact that QDA involves estimating a different covariance matrix for every class, reducing the influence that the higher variance of the positive class has on predictions involving the negative class. Thus, it appears that it is the optimal model when FP costs are larger or equal to FN costs; meanwhile, Logistic is slightly better where these costs are negligable or practically impossible to estimate.

Clustering

It is also possible to classify pulsars solely with unsupervised approaches; namely, Complete, Average, and Single Hierarchical clustering, K-Means clustering, and Gaussian Mixture. To avoid data leakage in the Cluster-then-Classify approach, only training data is used, and clusters are evaluated against the training data labels. For comparison, the same accuracy metrics previously employed in classification are also used.

Cluster Model Accuracy

Model/Accuracy	**Simple Accuracy (%)**	**1:1 (FN:FP cost)(%)**	**1:10 (FN:FP cost)(%)**	**1:100 (FN:FP cost)(%)**	**Area Under ROC (AUC)(%)**
HClust (Comp.)	90.724	49.901	9.073	0.988	50.00
HClust (Avg.)	90.886	49.990	9.089	0.990	50.00
HClust (Single)	90.886	49.990	9.089	0.990	50.00
K-Means	90.688	11.256	13.203	13.589	88.74
Gaussian Mix	85.188	87.051	88.914	89.289	87.05

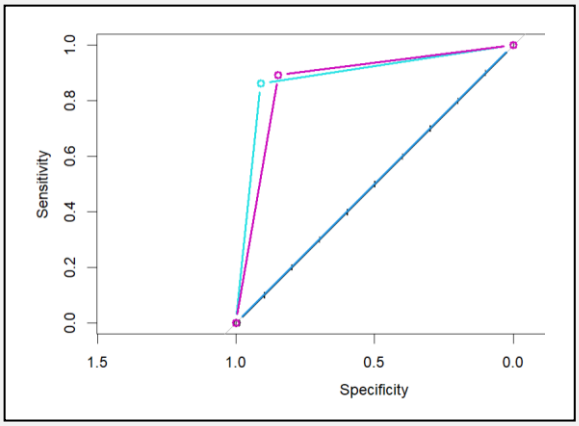


Fig. 4.2

The poor performance of these clustering models relative to the classification models can be explained by the fact that the former search for unknown patterns in unstructured data. As a result, they are likely identifying patterns irrelevant or unknown to the predefined pulsar classification problem. For instance, it could be that some of the pulsar's radio signals interact with an exogenous force or particle in their journey. The unknown patterns may also arise from idiosyncrasies of measurement; that is, the radio telescope's detection or radio signal metrics such as the Integrated Profile or DM-SNR curve. Out of these more holistic classifications, Gaussian mixture underperforms the Hierarchical and K-means clusters in terms of simple accuracy, but far outperforms them in terms of AUC and balanced metrics. For large FP:FN costs, Gaussian Mixture is optimal. Alternatively, K-means could be use when simply maximizing the proportion correct is the goal, with a hedge against the dramatic balanced accuracy and AUC drop-offs if costs are introduced into the model.

Cluster-then-Classify

Even though no new pulsar candidates are added to the data sets, an approach involving 1) Clustering of the training and test data, independently (to avoid data leakage), into K groups 2) Classification based on the original data sets in conjunction with the cluster class, may offer improved predictive performance. Clustering exposes unknown patterns in the data and injects variance—information regarding the feature's interactions with these patterns –into the classification model (Trivedi, 2015). The highest independently performing models from each type of model, clustering and classification, are chosen in accordance with several possible scenarios regarding the choice of evaluation metric. Gaussian Mixture and QDA both proved robust where specificity was emphasized, so they are combined. Similarly, K-means and Logistic Regression proved relatively robust for a wider range of accuracy metrics, but especially simple accuracy and AUC. The classification steps of both models, QDA and Logistic, include interaction terms of every feature with the clustered class.

Cluster-then-Classify Model Accuracy

Model/Accuracy	**Simple Accuracy (%)**	**1:1 (FN:FP cost)**	**1:10 (FN:FP cost)**	**1:100 (FN:FP cost)**	**Area Under ROC (AUC, %)**
Gaussian-QDA	90.782	91.523	92.276	92.424	91.53
KM-Logistic	98.383	92.845	87.283	86.181	92.85

The resulting "chimeric" models behave as expected, with the Gaussian-QDA model exceling in pulsar classification contexts where the FP:FN costs are known. Meanwhile, the KM-Logistic performs better in contexts where simple accuracy is desired. Nevertheless, both perform strongly in the others' environment, and optimize the sensitivity-specificity trade-off similarly.

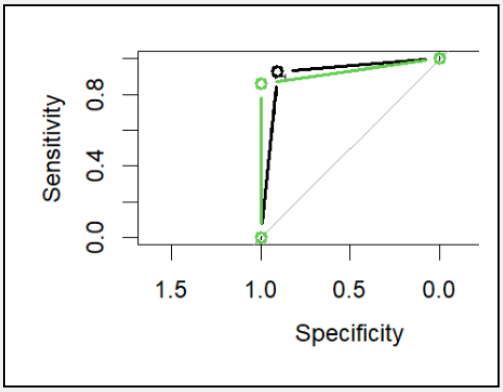


Fig. 3.2

Still Searching

There is no celestial object yet discovered that spin with greater ferocity than a pulsar. They are whirring generators, launching narrow emission beams throughout the cosmos. Differentiating their signals from the noise, cosmological or man-made, poses countless challenges throughout the classification pipeline. Powerful radio telescopes are built, radio signal metrics are formalized, and creative combinations of technology and human are needed to catalog them (Lyon, 2016). This analysis was one of many attempts to utilize supervised and unsupervised machine learning techniques to aid in the prediction of pulsars. It did so with a cost-sensitive approach, comparing various metrics used to evaluate the predicted classifications, such as simple and balanced accuracies, and the area under the ROC curve. Unsupervised clustering, while too holistic for effective pulsar prediction on its own, proves to significantly improve accuracies when combined with classification models. The Gaussian Mix ~ QDA and K means ~ Logistic Regression paired models are robust across cost structures, with the former exceling in environments where the costs of false positives are high to the scientific organization, and the latter excelling where such costs are negligible, or assumed exogenous to the model.

Why go to such lengths, building powerful radio telescopes, formalizing radio pulse metrics, fitting and evaluating the full spectrum of learning models, all in the hopes of improving the accuracy of pulsar classifications? As an unnamed educator at the Pisgah Astronomical Research Institute (PARI) eloquently stated:

> *"Pulsars are very bizarre, almost unfathomable, celestial objects. It is almost impossible to imagine an object propelling radiation into space at close to the speed of light or an object more massive than our Sun that would easily fit inside a city. We study pulsars because they are so oddly fascinating and so different from anything experience here on Earth. They are sites of extreme physics: density, gravity, magnetic and electric field."*
>
> (Sensing the Radio Sky, PARI)

In short, the pulsar classification problem resides on one of the many paths toward a better understanding of the universe, and ourselves. In the extreme electromagnetic and gravitational conditions created by Pulsars, forces of nature and particles that we still deem the most fundamental disintegrate and implode. In their absence, we are left with a deafening

unknowing. Yet, distantly, these mysterious pulsars sweep their emission beams across the vacuum, beckoning us to understand (Kramer, 2016).

Fig. 5.1: (Crass, Wilkins, Lucky)

Works Cited

as9736. “Classifying Dead Stars.” *ALL YOUR BASE ARE BELONG TO US*, 12 Mar. 2019, allofyourbases.com/2019/03/09/classifying-dead-stars/.

“Chapter 6 Pulsars.” *6 Pulsars‣ Essential Radio Astronomy*, www.cv.nrao.edu/~sransom/web/Ch6.html.

Chauta, Matt. “Astronomers Discover Fastest-Spinning Pulsar.” *National Radio Astronomy Observatory*, 5 Apr. 2017, public.nrao.edu/news/astronomers-discover-fastest-spinning-pulsar/.

Dunbar, Brian. “NASA's Fermi Finds Youngest Millisecond Pulsar, 100 Pulsars To-Date.” *NASA*, NASA, 18 Mar. 2019, www.nasa.gov/mission_pages/GLAST/news/young-pulsar.html.

“HTRU.” *HTRU | Max Planck Institute for Radio Astronomy*, www.mpifr-bonn.mpg.de/research/fundamental/htru.

“HTRU2.” *UCI Machine Learning Repository: HTRU2 Data Set*, archive.ics.uci.edu/ml/datasets/HTRU2.

Institute of Astronomy - Design by D.R. Wilkins and S.J. Crass. “Institute of Astronomy.” *Lucky Imaging | Institute of Astronomy*, www.ast.cam.ac.uk/research/lucky/.

Kramer, Michael. “Pulsars as Probes of Gravity and Fundamental Physics.” *ArXiv.org*, 13 June 2016, arxiv.org/abs/1606.03843.

Lyon, Robert James. “Why Are Pulsars Hard To Find?. *Research Explorer | The University of Manchester*, 2016. www.research.manchester.ac.uk/portal/en/theses/why-are-pulsars-hard-to-find(f15226ec-355d-4794-b2b8-e0a8e793948e).html.

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

“Pulsar.” *Wikipedia*, Wikimedia Foundation, 4 July 2022, en.wikipedia.org/wiki/Pulsar.

PARI. *Sensing the Radio Sky*, campus.pari.edu/radiosky/lessons/pulsars/10.shtml.

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

Trivedi, S., Pardos, Z.A. & Heffernan, N.T., 2015. The Utility of Clustering in Prediction Tasks. *arXiv.org*. Available at: https://doi.org/10.48550/arXiv.1509.0