Indian Institute of Space Science and Technology

Department of Earth and Space Sciences

Chandra X-ray Source classification using Machine Learning

Final year project - Phase 1

Project by -
Shivam Kumaran
SC17B122
IIST

Guide
Dr. Samir Mandal , IIST
Dr Sudip Bhattacharya , TIFR
Dr Anjali Rao , IIT Indore

October 24, 2021

Contents

Abstract

This work aims at classifying globular cluster x-ray sources identified by Chandra observatory and listed in Chandra source catalogue. In globular clusters most abundant x-ray sources are CVs , LMXRBs and milli second pulsars. In the current phase we have given a scheme for classification of Low mass X-ray binaries between (Neutron star LMXRB and Black hole LMXRB). Using literature survey and other catalogues we have identified 33 BH and 83 NS already confirmed lmxrb. Features available in CSC observation table for source classification were used. Since most of the identification done prior are during outburst , but our aim is to classify source observations during outburst , so we used a flux threshold value of $10^{-12} erg/s/cm^2$ to filter out observations during outburst. In observation table more than 50% values are missing. Missing values were filled using Random Forest Imputation and the updated observation table was used for classification. Random Forest was identified as most robust classifier with test accuracy - 92% was achieved , with ROC-AUC score 0.99. With feature importance study we observed flux in low energy bands are important for identification of NS lmxrb and Black body model fit parameters are important for BH lmxrb identification.

0.1 Introduction

With the advent of very high resolution x-ray observatory like *CHANDRA* , with a resolution of 1 arcsec, the number of sources detected are increasing such that manual classification of each of them individually is not possible. With Automated machine learning algorithm we can leverage the information of already classified sources to assign a probabilistic classification to unidentified sources. In this work we are interested in classification of x-ray sources belonging to globular cluster. Due to high density In globular cluster x-ray sources are mostly - Cataclysmic variables , Low mass xray binaries milli-second pulsars[1]. For Low mass x -ray binary neutron star or a black hole accretes matter form a companion star via Roche-lobe of accretion disk. Distinction between NS and BH is confirmed either dynamically or using spectroscopy when they go into outburst.

Problem Statement *We need to identity the class of x-ray object in globular cluster detected by Chandra observatory when they were in quiescent state using the properties available in observation table of Chandra source catalogue 2.0*

0.1.1 Workflow

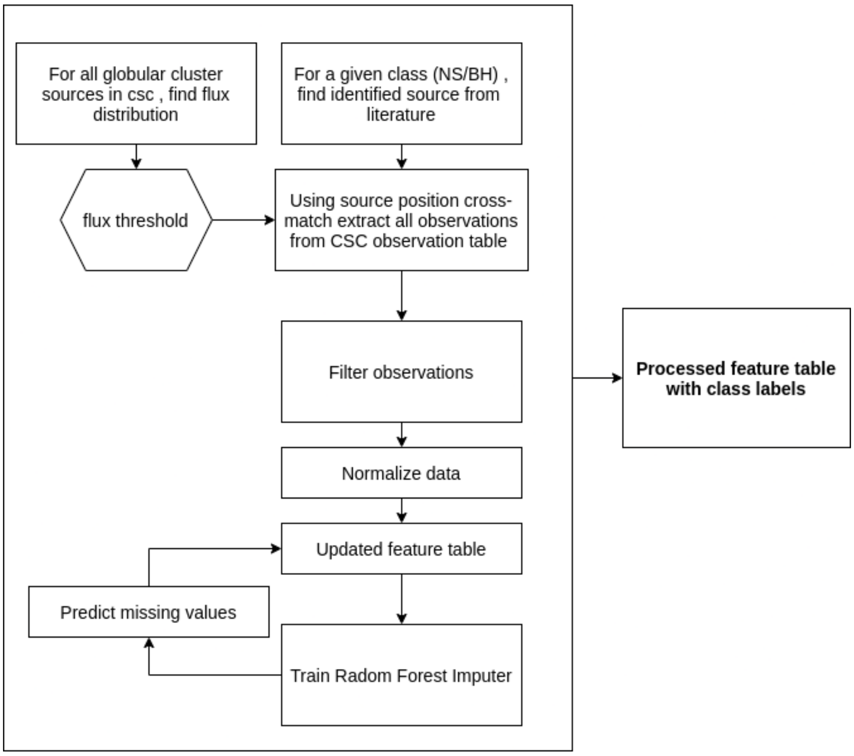
For all globular cluster sources in csc , find flux distribution
For a given class (NS/BH) , find identified source from literature
flux threshold
Using source position cross-match extract all observations from CSC observation table
Filter observations
Processed feature table with class labels
Normalize data
Updated feature table
Predict missing values
Train Radom Forest Imputer

Figure 1: Working schematic for training data preparation for two classes NS and BH lmxrb

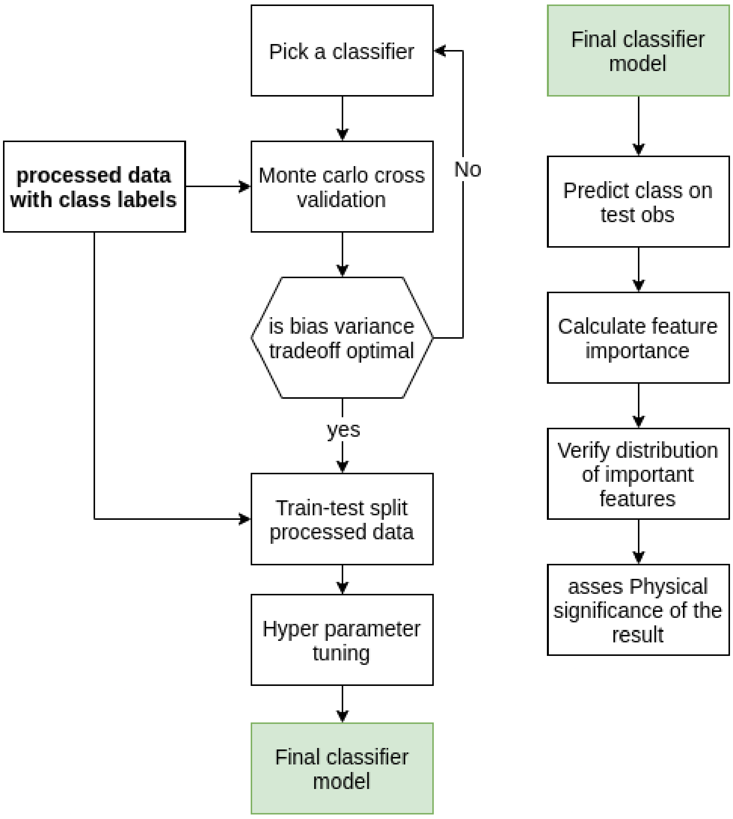
Pick a classifier
Final classifier model
No
processed data with class labels
Monte carlo cross validation
Predict class on test obs
is bias variance tradeoff optimal
Calculate feature importance
yes
Verify distribution of important features
Train-test split processed data
asses Physical significance of the result
Hyper parameter tuning
Final classifier model

Figure 2: Workflow for identification of best classifier and using this for final classification

0.2 Chandra Source Catalogue (csc-2.0)

Chandra Source Catalogue 2.0 (csc-2.0) is a catalogue of X-ray sources detected by Chandra X-ray observatory. This catalogue consist of 317,167 unique identified sources. For observation of sources for catalogue preparations , two instruments on-board the observatory ACIS (obs in 5 different bands) and HRC (one wide band) are used. The catalogue is split into three tables

- **Master Sources Table** The Master Sources Table contains 'best estimate' sources properties for each unique X-ray source in the catalog. These properties are typically determined by analyzing simultaneously the detections from the set of individual observations.
- **Stacked Observation Detections Table** contains detection properties based on observational data extracted independently from each stack of Chandra pointings in which the source is detected;
- **Per-Observation Detections Table** contains detection properties based on observational data extracted independently from each individual observation (Chandra pointing) in which the corresponding stacked observation detection is located; because a stack may include multiple observations

Since we have to filter and consider observations made during quiescent state , we work with per observation table. The source properties (Refereed to as features on-wards in this report) is given under following contextual categories.

0.2.1 Features

Relevent features used for classification

1. **Source Information**
    - Position RA-DEC
    - Galactic coordinates position
    - Exposure timings
    - Instrument used
    - Flux significance - ratio of the single-observation detection photon flux to the estimated error in the photon flux, for each band
2. **Source Fluxes** photon flux and energy flux values are calculated using aperture photometry for different bands. Also flux values calculated by model-fitting are given.
3. **Spectral Properties** Models used :
    - Black Body
    - Powerlaw
    - Bremsstrahlung

Hardness Ratio Hardness ratio between two bands (x,y : where x,y may be h,m,s,u bands) are given as

$$hard_x y = \frac{F(x) - F(y)}{F(x) + F(y)} \tag{1}$$

where F(x) and F(y) are aperture source photon flux in x and y band , where x,y may be h, m, s, u.

4. **Source Vaiability** Based on source region counts within observation source variability is calculated by the following methods.

   - Kolmogorov-Smirnov (K-S) test
   - the Kuiper's test
   - computation of the Gregory-Loredo variability probability

   Both inter-observation and intra-observation source variability properties are given in the table. We are using only intra-observation properties.

5. **Source flags**

   - pileup flag
   - saturated source flag
   - streak source flag

   These source flag were used to filter out observations.

0.3 Source observation data preparation

0.3.1 Data collection

In chandra source catalogue there is no classification for source is given. But for supervised learning for any machine learning classifier , we need already labelled examples. In literature and in previously available catalogue we can pick out identified sources and check whether CSC has observed this source or not.

Methodology

- For a given class of object , we look for identified x-ray sources in literature and/or in other catalogues.
- Find out RA-DEC of those sources
- Use HEASARC table cross-correlation to find matching sources in CSC. Cross correlation radius is taken as minimum of 10" or
- Download the closest match source observation data if available.
- If more than one source is cross-matched , keep only the best cross-matched source.

For the current phase of project we are concerned with a robust classification of Neutron Star and Black Hole Low mass x-ray binary. We collected source observations corresponding to these. Information about the sources and their detection type with reference is given in the following table

0.3.2 Obs Filtering

- Flux filtering
- Pileup-filter
- streak source filter
- saturation filter
- significance filter

Flux filter

In maximum cases the identified sources are classified only when they goes into outburst (for the case of Xray binaries). But our aim is to classify sources in CSC which are in quiescent state , which have quiet different properties even for the same source. So we need to make sure that the data our classifier is trained on comes from the sources in quiescent state so we need to take observations corresponding to the source when they were in quiescent state.

During outburst variation of luminosity of LMXRB in quiescent state is $L_X \approx 10^{32} erg/s$. During outburst luminosity goes as high as $L_X \approx 10^{36} - 1p^{38} erg/s$ Now if we are considering only galactic sources, using distance limits (distance to the center of the galaxy - $\approx 8kpc$, diameter of the galaxy $\approx 15Kpc$) we can come up with a boundary value of minimum flux for outburst sources.

Dist | L_x	10^{36}	10^{38}
1kPc	8.4×10^{-9}	8.4×10^{-7}
8kPc	1.3×10^{-10}	8.4×10^{-8}
15kPc	3.7×10^{-11}	3.7×10^{-9}

Hence from the table above, we can conclude that if we take flux less than $10^(-12) erg/s/cm^2$, we can make sure that the obs is not during outburst. We also look at the flux distribution of all the x-ray sources belonging to chandra source catalogue.

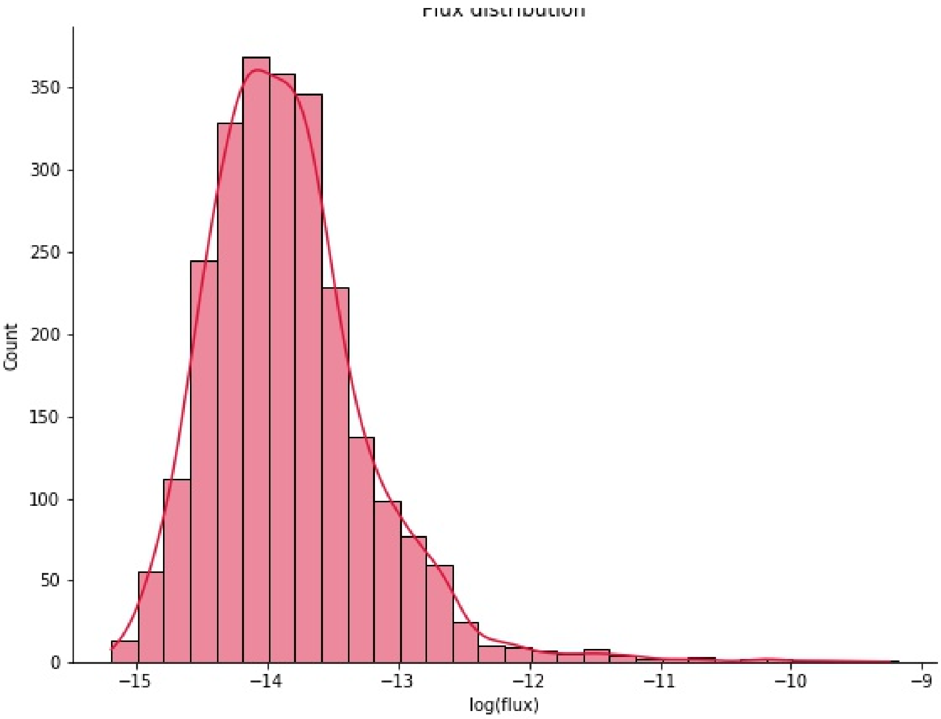

Figure 3: Globular cluster x-ray sources flux distribution. Flux values considered are in ACIS Broad-band filter

The figure 3 shows the broad band filter energy flux distribution for GC X-ray sources. It is clear that most of the sources are well below flux level $10^{-12} erg/s/cm^2$.

Table 1: Total collected sources and observations , before obs filtering

	Num of sources	Num of Obs
NS lmxrb	84	493
BH lmxrb	33	227

Table 2: Number of sources and corresponding observations after all the filters applied

	Num of sources	Num of Obs
NS lmxrb	48	302
BH lmxrb	27	158

[1]

0.4 Pre-training / Data prepossessing

Before we feed raw data to any machine learning algorithm , we must pre-process data to make it suitable for certain classifier algorithm. Data pre-processing have a far more

[1] More information about individual sources and how their class were identified is to be included in appendix

impact on classification quality

0.4.1 Data Scaling

In out data-set , the magnitude scale have a very high difference , for example all 'flux' features are $\sim 10^{-12}$ whereas variability features are of the order of ~ 10. Difference in order of features results in assigning higher importance to features with higher magnitudes which may not be very importance for source identification when network based algorithm are used , hence we need to normalise data.

- Normalisation for each column of features :

  $$xi = (xi - max)/(max - min) \tag{2}$$

  Such that for each observations , feature all the feature values are within 0-1

- Standardisation

  $$xi = (xi - mean)/var \tag{3}$$

  such that the distribution of each feature becomes standard , with '0' mean and unit variance

0.4.2 Data Imputation

Not all the observation in CSC are taken in all the bands. For observations which are taken only in two bands , flux values in other bands are missing. Hardness values and model fitting parameters are missing. Missing data is a major challenge for any classification task and can introduce large variance in the classifier. Data imputation techniques used :

- Zero Imputation : fill in all the missing value with zero
- Mean imputation : Missing values filled with feature-mean
- Median imputation : Missing values filled by feature-median
- Correlation imputation : We take the advantage of available values and their correlation among features to fill in missing data.
  1. correlation coefficient for each pair of feature is computed (feature-feature correlation matrix)
  2. For each observation (say x_i) , for missing feature , say f_i , highest correlated feature , say f_j is selected.
  3. for x_i , using value of feature f_j , feature f_j id filled with linear regression output from feature (f_i , f_j).
- Imputation Using Random Forest : If two sources are similar to each other , then we can use one to fill in the missing value of the other. With Random forest we can find the proximity of each pair of observation called *proximity matrix* , where the element of this matrix correspond to obs $(x_i - x_j)$ and the value if the fraction of number of decision trees of random forest in which both the observation follows the same path and end up in same leaf node. This method is based on the fact that two similar obs are likely to follow same path along decision tree.

Iterative procedure for data filling

1. Start by filling missing values with corresponding feature median
2. Calculate proximity matrix using Random forest
3. For each obs x_i , for each missing feature f_i , fill it with weighted average of available f_i value across all x_ks with weighing factor proximity- $(x_i - x_k)$
4. repeat step 3 - 4
5. After 6-7 iterations filled in values converge.

0.5 Classifiers

We are using supervised machine learning methods for classification. For now the problem is binary classification (NS LMXRB - BH LMXRB). Starting with basic classifier like Logistic regression , we have tried following classifiers

- Logistic regression
- K-Nearest Neighbour
- Fully connected network
- Convolution Neural network
- Random Forest classifier

0.6 Best Classifier Scheme

We have a collection of different data-scaling method , data-imputation method and classifiers , and thus different permutation gives different classification schematic. Best classifier should give accuracy and the variance in accuracy should be small (i.e accuracy should not be affected by training sample variations). To compare performance of each of the combinations , we use **Monte-Carlo Cross validation** to get optimal BIAS-VARIANCE trade-off :

1. select data-scaling and data-imputation method and pass data through corresponding processing pipeline
2. Pick the classification algorithm
3. From dataset , 20 % is randomly sampled for test , rest 80% is used for training , (split is done with stratification , making sure class balance in test data)
4. Train the classifier on training set.
5. Measure network accuracy on test data , called TEST-ACCURACY
6. Step 3-5 is repeated 32 times , for each classifier.

0.6.1 MC validation Result

From the figure 4 , shows that except Random forest , data set without normalisation have a very high variance in test accuracy and also a lower mean accuracy. In most of the cases , standardisation works better. For RF case , data scaling does not have much impact.

From the plot 5 , we see that for KNN , FC and CNN Correlation imputation works best with accuracy distribution skewed towards higher side. For random forest , correlation imputation has resulted in lower accuracy, whilst median imputation works best for RF.

From these two plots 5 4 , it is evident that Random Forest has least variance and higher test accuracy. *Random Forest is our classifier of choice* For comparison Random Forest imputation is tested with RF classifier. From plot 6 , we see that RF imputation has the least variance and highest mean test accuracy. Also mean train accuracy is lower for RF , however this shows least over fitting case. *RF imputation is our choice of data imputation* and since we have selected RF classifier , data scaling does not have much effect (this behaviour is expected since RF works based on decision trees), so to avoid floating point calculation accuracy we choose to use *Normalisation*

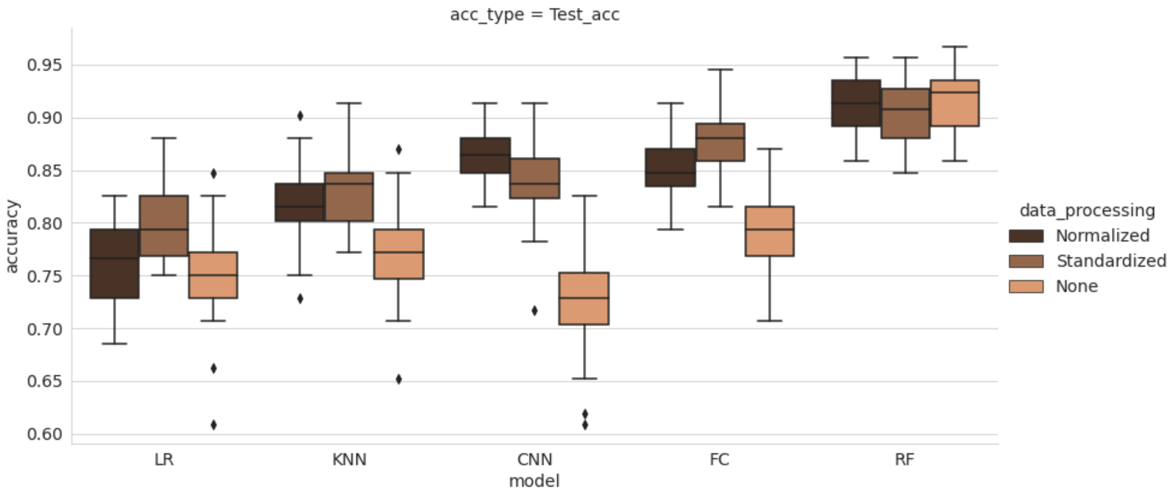

Figure 4: Caption

[H]

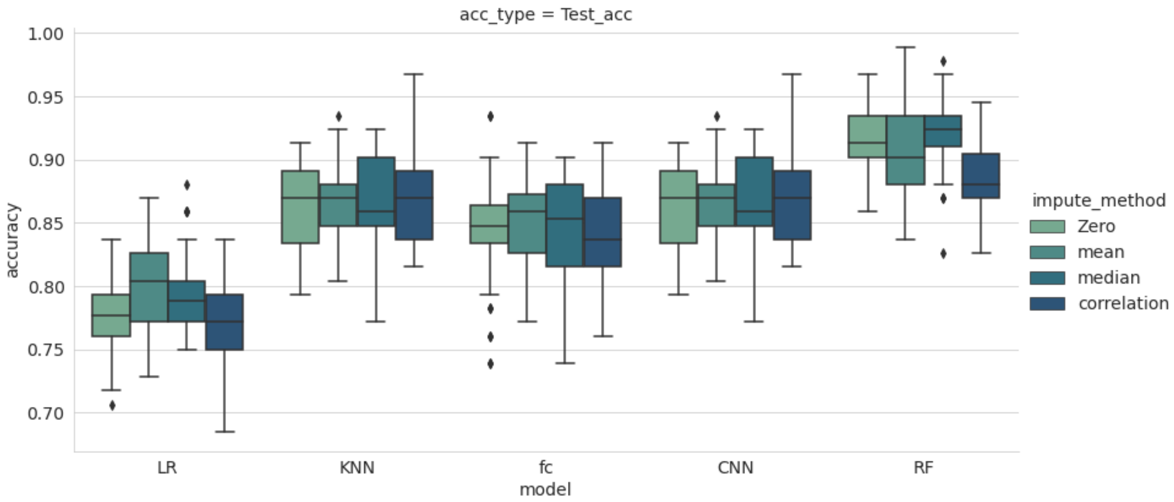

Figure 5: Caption

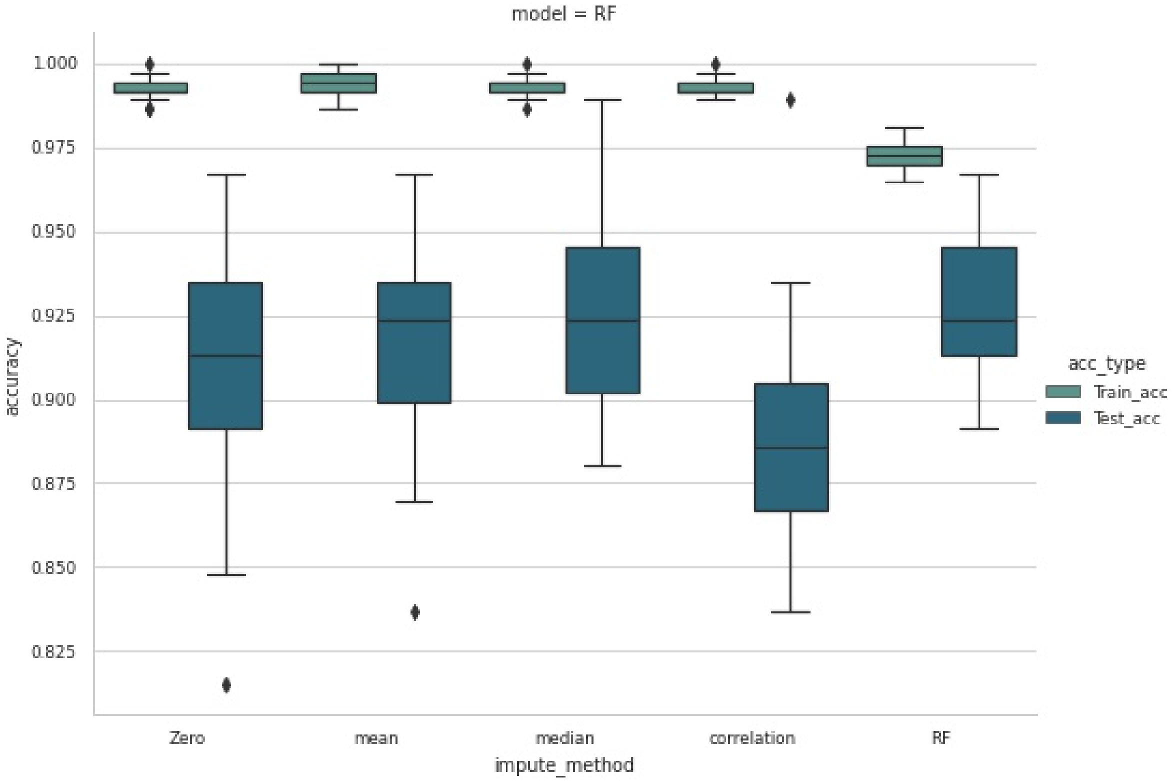

Figure 6: Caption

Schematic selected

- Data scaling - Normalisation
- Data Imputation - Random Forest Imputation
- Classifier - Random Forest

0.7 Training

Data-set is normalised, and imputed with **Random-forest-imputer**. We then used this as our updated feature table. Observations rows are reshuffled, and split as :

Table 3: Dataset Train-test split

	fraction	Total obs	BH nos.	NS Nos.
Training sample	0.8	368	126	242
Test sample	0.2	92	32	60

At first RF is trained with default values. Accuracy and prediction probability is increased with hyper-parameter tuning .For random forest following hyper-parameters are needed to be tuned

- umber of decision trees
- Maximum depth of decision trees

Tuning is done in two steps

- Random search withing a range of parameter values
- Grid Search over finer range of values about previously best selected values via random search.

0.7.1 Training Result

Predicted Probability Quality

A better classifier is one with predicted probabilities close to 1.

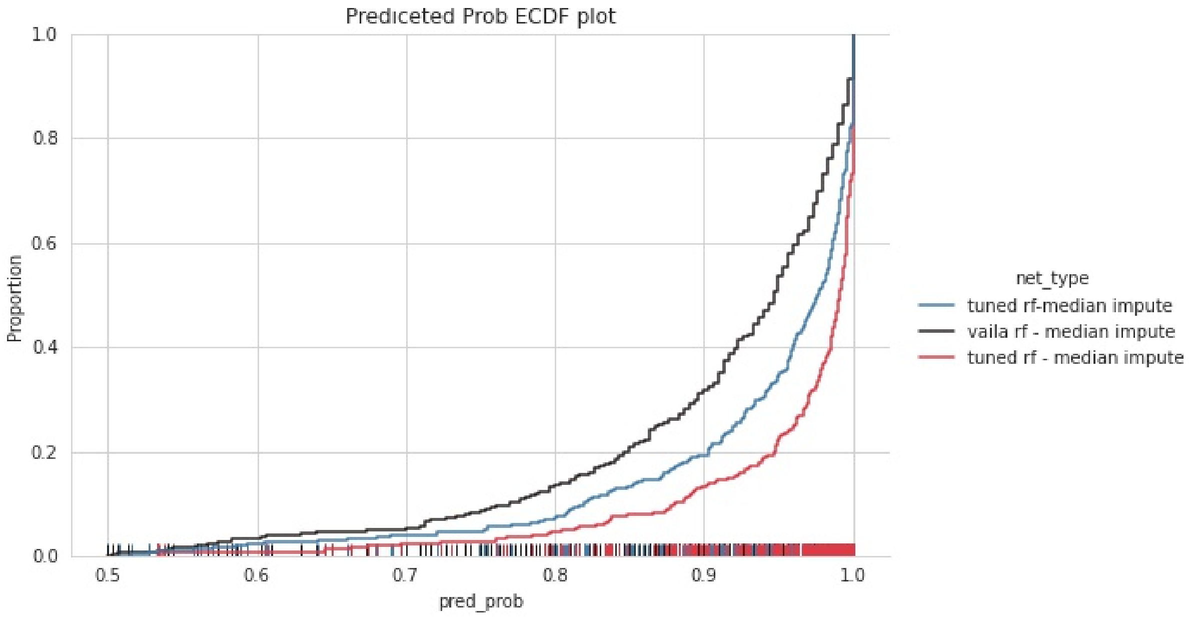

Figure 7: Empirical Cumulative distribution curve for predicted probabilities for - Default RF with median imputation (), tuned RF with median imputation , tuned RF with RF imputation

Figure 7 shows empirical cumulative probability distribution for different classification schemes. It shows the proportion of observations for which predicted probability is less than certain value. For *tuned RF* , *with RF imputation* , less than about 10 % of observations predicted probabilities are less than 0.9. This method is more confident in predicting class.

0.7.2 Feature Importance

RF classifier assign feature importance based on **Gini impurity** , which shows that amount of discrimination brought about by each feature. Feature importance for each individual class is computed as :

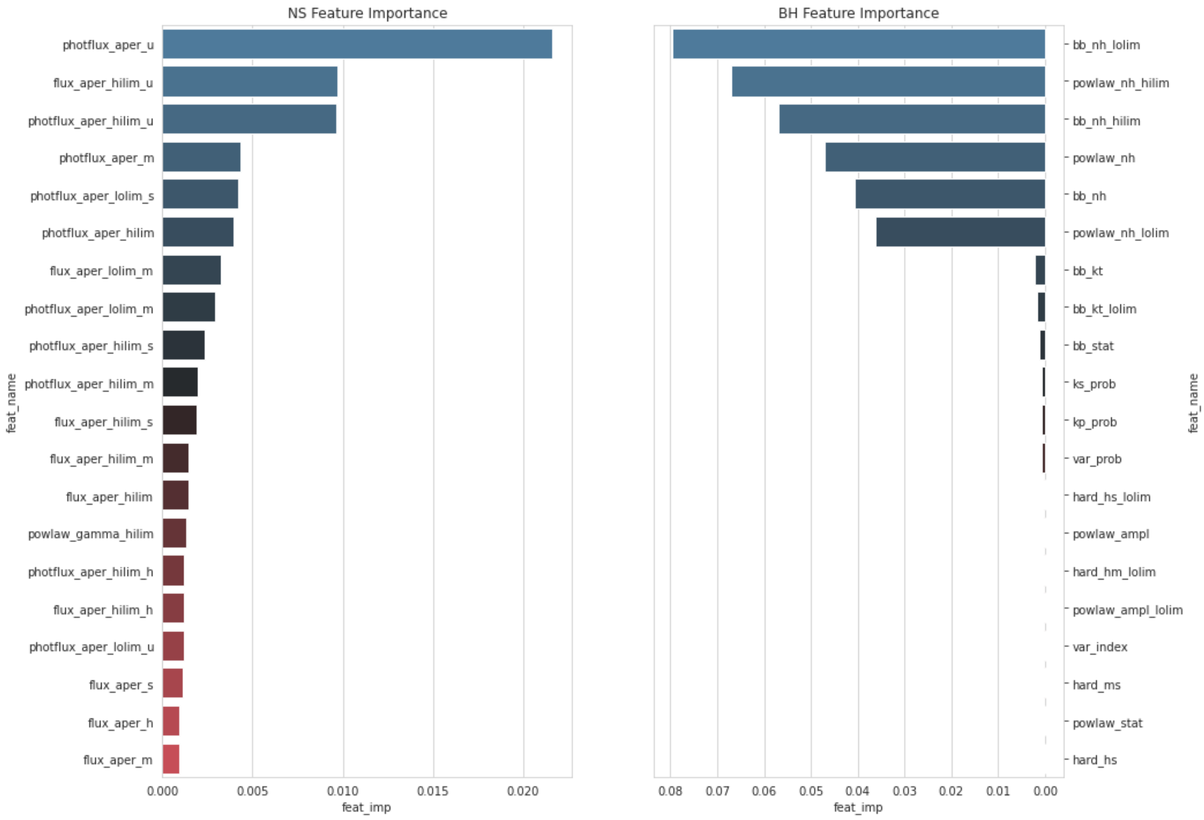

Figure 8: Gini impurity based Feature importance for NS and BH

Observation

For Neutron Star XRB more importance is assigned to photon flux and energy flux features in ultra-soft , soft and medium band. For Black hole XRB , important features are Black Body model fit parameters, especially column density. Since these features are importance there is a clear distinction in the distribution of thee feature cor two classes

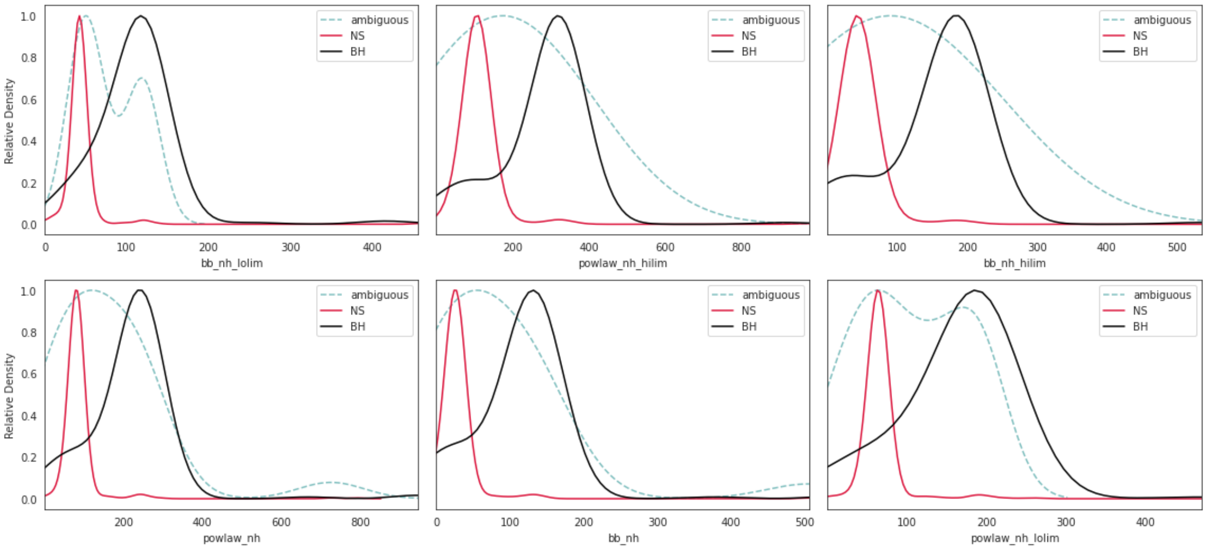

Figure 9: Feature distribution plot for features that were identified as important for BH lmxrb. Each plot shows the distribution of features for both classes. Distribution of features for objects that were assigned ambiguous class is also shown

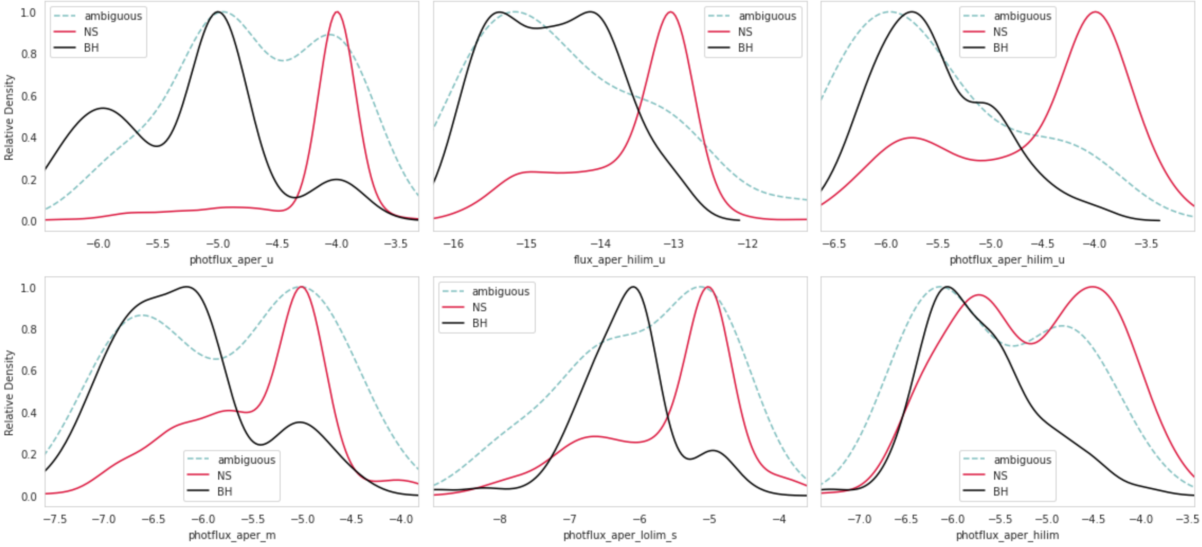

Figure 10: Feature distribution plot for features that were identified as important for Neutron Star lmxrb. Each plot shows the distribution of features for both classes. Distribution of features for objects that were assigned ambiguous class is also shown

0.8 Conclusion and future workplan

Bibliography

[1] D. Pooley, Pooley, and D. “Globular cluster X-ray sources”. In: *MmSAI* 87 (2016), p. 547. ISSN: 0037-8720. URL: https://ui.adsabs.harvard.edu/abs/2016MmSAI..87..547P/abstract (page 1).