Classifying ROSAT X-ray Sources

By: Lauren Janicke, Janice Lee, Peicheng Qiu, Jenny Shan
Advisor: Peter Freeman

Carnegie Mellon University
Statistics & Data Science

Introduction

In 1990, the ROSAT X-ray telescope was launched to observe "X-ray binaries", a class of binary stars that are luminous in X-rays.[1] While some X-ray sources have an extended shape—like the expanding gas cloud of a supernova remnant—most are point-like and thus hard to classify by visual inspection only. With follow-up observations of the sources with three other telescopes, including the brightness measurements from the optical regime by Gaia and SDSS and the infrared measurements by WISE, it is possible to differentiate the source types.

The goal of our study is to learn a statistical model that takes in X-ray and brightness measurements of astronomical objects and produces an accurate classification of quasars and galaxies.

Data

Our dataset has 4198 astronomical bodies and 26 predictor variables. There are five classes for the response variable: quasars, broad-line active galactic nuclei (BLAGN), narrow-line active galactic nuclei (NLAGN), galaxies, and stars. Quasars and broad-line active galactic nuclei were combined to form one class, and galaxies and narrow-line active galactic nuclei were combined to form another. Stars were removed from the data considered in the statistical models. Some predictor variables were log-transformed for better visualization and analysis.

Predictor Variable Name	Description
RXS_(ExiML, CRate, Ext, LOGGALNH, SRC_FLUX)	ROSAT observations: detection likelihood, source X-ray count rate, source extent in ROSAT CCD pixels, log-base-10 of hydrogen column density (cm^(-2)), source flux in 0.1-2.4 keV band (erg/cm^2/sec)
ALLW_(W1,W2,W3,W4,J,H,K)mag	source magnitudes as measured by WISE, in 7 infrared (IR) bands
SDSS_MODELMAG_(u,g,r,i,z)	first version of source magnitudes as measured by SDSS, in 5 optical and near-IR bands
SDSS_FIBER2MAG_(u,g,r,i,z)	second version of source magnitudes as measured by SDSS, in 5 optical and near-IR bands
Z_BEST	best estimate of source redshift
GAIA_DR2_phot_(g,bp,rp)_mean_mag	source magnitudes as measured by Gaia, in 3 optical bands

Table 1: Predictor Variables and Descriptions

The pairwise plot of ROSAT observations shows a strong correlation between the log of RXS_CRate, the log of RXS_ExiML, and the RXS_SRC_FLUX variables. The scatter plots and density plots, however, show a lot of overlap between the classes.

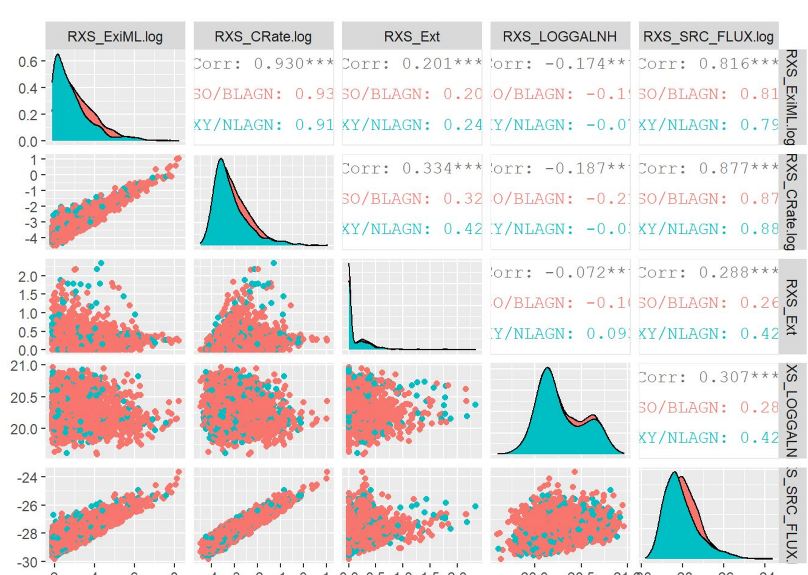

Figure 1: Pairwise Plot of ROSAT data

Methods

- We utilized various classification techniques to build binary classifiers
  - Main methods: logistic regression, forward/backward subset selection, LASSO, Ridge Regression, classification tree, random forest, XGBoost, KNN, Naive Bayes, and SVM with linear, polynomial, and radial kernels
  - Vif reduction was used to address multicollinearity within the data for the regression models.
- The highest AUCs are from lasso regression on the full predictor space, boosting, and random forest, which yields 0.934, 0.933, 0.933, respectively.
  - LASSO regression is a shrinkage method performing both variable selection and regularization
  - Boosting is a family of algorithms that consists of iteratively learning weak classifiers and add them to a final stronger learner
  - Random Forest is an ensemble learning model that involves constructing and aggregating multiple decision trees
- We decided to use our LASSO regression model to generate final predictions because it has the greatest AUC.

Analysis

Model	AUC
Lasso Regression (non-vif)	0.934
Boosting	0.933
Random Forest	0.933
Backward Selection (non-vif)	0.930
Forward Selection (non-vif)	0.930
Logistic Regression (non-vif)	0.929
Backward Selection (vif-reduced)	0.928
Forward Selection (vif-reduced)	0.928
Lasso Regression (vif-reduced)	0.928
Logistic Regression (vif-reduced)	0.928
Ridge Regression (non-vif)	0.926
Ridge Regression (vif-reduced)	0.923
SVM - Linear	0.917
SVM - Polynomial	0.916
Naive Bayes	0.879
SVM - Radial	0.869
KNN	0.738
Decision Tree	0.655

Table 2: AUCs for the Models Considered

- The AUC for models on full predictor space are unanimously greater than that for models with vif-reduced inputs. While models on full predictor space yield to more accurate predictions, the issue with multicollinearity may undermine their inferential ability.
- The AUC for our non-vif LASSO Regression model is 0.934. This is also the highest AUC among those AUCs from all the models. The AUC for LASSO is visualized below.

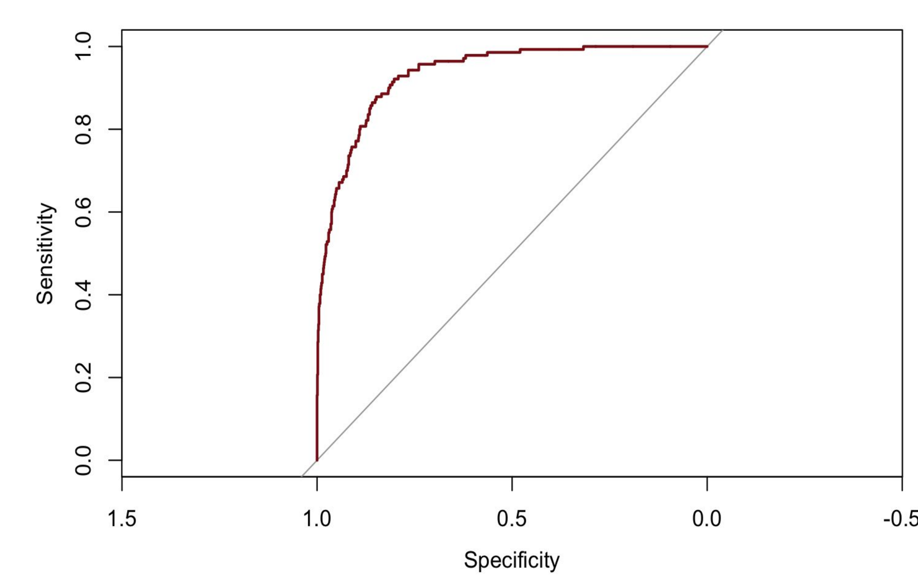

Figure 2: ROC Curve of LASSO Regression

		LASSO Predictions	
		Galaxy/ NLAGN	Quasar/ BLAGN
Response Variable	Galaxy/ NLAGN	113	27
	Quasar/ BLAGN	118	929

Table 3: Confusion Matrix of LASSO Regression

- The MCR for our non-vif LASSO Regression model is 0.12.
- The sensitivity and specificity are both about 0.86.
- The plot below shows the coefficients of the predictor variables given different log of lambda values. The log of our best lambda was -9.094. This means 24 predictor variables contribute to the model.

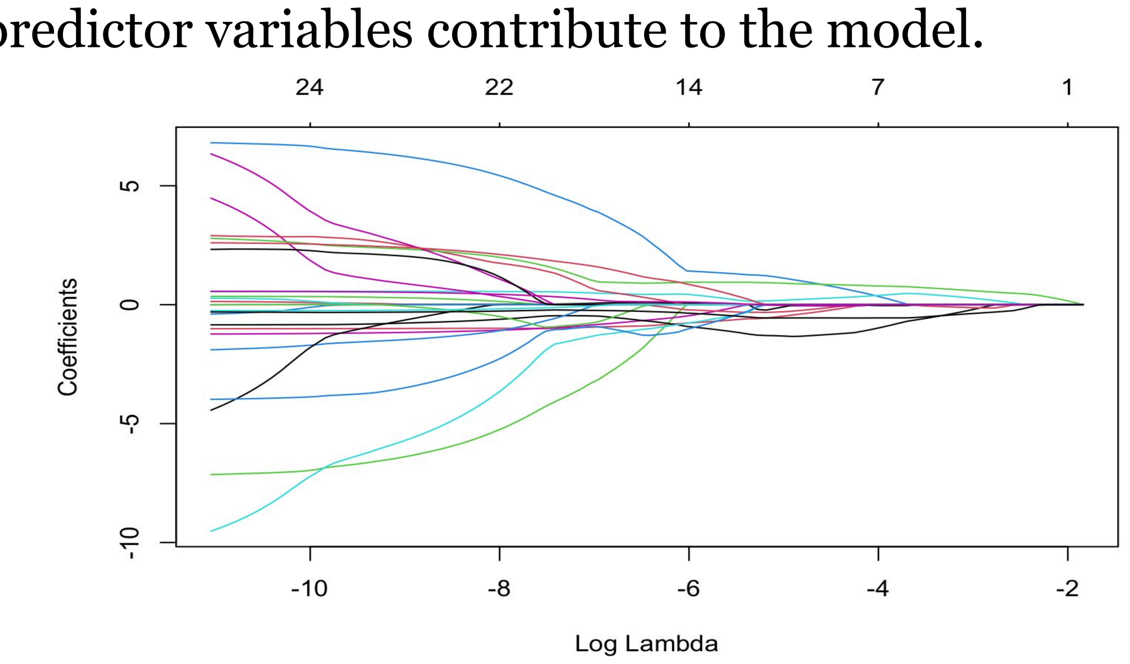

Figure 3: Lambda of LASSO Regression

Conclusions

We conclude that we can indeed classify quasars and broad-line active galactic nuclei versus galaxies and narrow-line active galactic nuclei with relative accuracy. The metric used for determining the optimal model was having the highest AUC. Using this metric, we found the statistical model that optimally performed this classification was a lasso regression on the full predictor space, with a misclassification rate of 12%.

References

1. Comparat, J. et al. 2020, Astronomy & Astrophysics, in press (arXiv:1912.03068)