# Predicting the redshift of gamma ray loud AGNs using machine learning

Prof. Maria Giovanna Dainotti, Prof. Malgorzata Bogdan, Aditya Narendra, Spencer James Gibson, Prof. Blazej Miasojedow, Prof. Ioannis Liodakis, Prof. Agnieszka Pollo, Trevor Nelson, Kamil Wozniak

## INTRODUCTION

In this work, we are building a machine learning model which can predict the redshift of gamma-ray loud AGNs, based on their observed photometric properties.

Measuring the redshift of AGNs is a difficult and time-consuming task, as it requires detailed spectroscopic observations. Hence, there is a requirement for a method that can estimate the redshift using the photometric observations of an AGN. Over the years many such methods have been proposed, with machine learning (ML) being one of the most recent developments.
In this work we are using the Fermi 4 LAC catalog to train our ensemble machine learning model and predict the redshift.

## THE DATA

The Fermi Gamma Ray telescope has been continuously monitoring the sky in 50MeV to 1TeV range since 2008. The gamma-ray properties used in this work are obtained from the 4LAC catalog, which contains a total of 2863 sources. Out of these only 1591 AGNs have measured redshifts. Ideally, we should use all the 1591 AGNs to train our ML models, but due to incomplete observations for some sources, we need to discard them. This leads us to a total of 730 AGNs which we use for training.

These are made up of 422 BL Lacertae Objects (BLLs) and 308 Flat Spectrum Radio Quasars (FSRQs). Furthermore, we have the generalization set, which is made up of those AGN sources which do not have measured redshifts. Our trained model should ultimately predict the redshift of these AGNs.

Some of the variables are used in their logarithmic form since they span over several orders of magnitude and we predict the redshift in the scale of 1/z+1 (see Fig. 4).


Fig 1: The scatter matrix plot of the predictors used. These observed properties are used to predict the redshift. The scatter matrix plot shows their distribution with respect to eachother. The blue points denote the distribution of the generalization set.

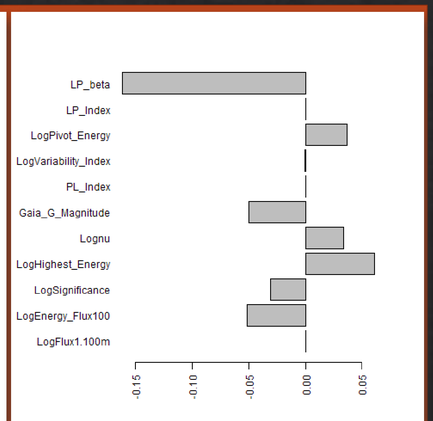


Fig 2: LASSO feature selection. The bars show the coefficient LASSO assigns to each predictor. We choose those predictors which have a coefficient more than 5%.

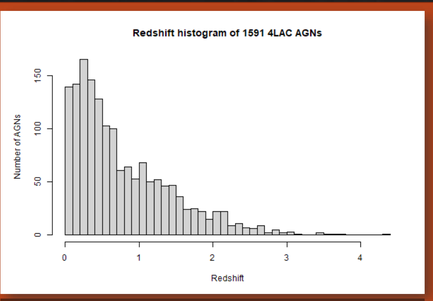


Fig 3: This figure depicts the distribution of the redshift in our data. Since there is an imbalance in the distribution of the redshifts (more low-z AGN than high-z), we apply a transformation of 1/(z+1), such that the distribution is more uniform.

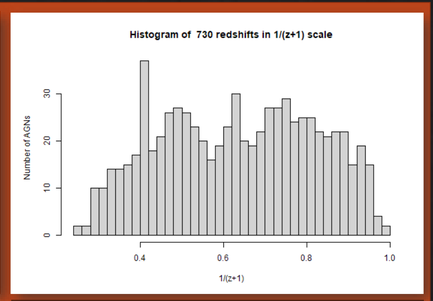


Fig 4: This is the transformed redshift distribution which has a more even distribution than pure redshift. We train the ML models to predict on this distribution, rather than directly the redshift. Such a transformation improves our final ability to better predict the redshift.

**This work has been accepted for publication in the Astrophysical Journal. The arXiv link is: https://arxiv.org/abs/2107.10952**

## FEATURE SELECTION

One of the techniques that is used in our project is called feature selection. In this technique we pick out those observed properties from the catalog which are better at predicting the redshift. For this purpose we use the LASSO method. LASSO algorithm uses a shrinkage method for linear regression by requiring the $l^1$ norm (sum of the magnitude of all vectors in the given space) of the solution vector to be less than or equal to a positive number known as the tuning parameter (λ). By varying this tuning parameter we train the model to fit the data and reduce the prediction errors.
The λ tuning parameter is tuned such that the prediction error is within 1 sigma deviation. Based on this LASSO assigns coefficients to the predictors, depicting how effective they are in predicting the redshift. We pick those predictors which have a coefficient of 5% or more. In this way, we picked 7 predictors out of a total of 14 present in the catalog. These are:

LogEnergy Flux - Logarithm in base of 10 of the energy flux, the units are in erg $cm^{-2}s^{-1}$, in the 100 MeV 100 GeV range obtained by the spectral fitting in this range.

LogSignificance - The source detection significance in Gaussian sigma units, on the range from 50 MeV to 1 TeV.

Log Highest Energy - Measured in GeV, it is the energy of the highest energy photon detected for each source, selected from the lowest instrumental background noise data, with an associated probability of more than 95%.

Logv - Logarithm in base of 10 of the synchrotron peak frequency in the observer frame, measured in Hz.

LogPivot Energy - The energy, in MeV, at which the error in the differential photon flux is minimal, derived from the likelihood analysis in the range from 100 MeV - 1 TeV.

LP β - the spectral parameter (β) when fitting with Log Parabola spectrum from 50 MeV to 1 TeV.

Gaia G Magnitude - Gaia Magnitude at the g-band provided by the 4LAC, taken from the Gaia Survey

## SUPERLEARNER

Superlearner (SL) is an ML algorithm that allows us to combine multiple algorithms into an ensemble. An ensemble is a collection of ML algorithms that work together to give a prediction that is more accurate than any of the constituent models. This is a great way to overcome the short comings of individual models and obtain better results.
Indeed, SL works in this way. It take the methods we specify, and then trains each of them individually on the data.
It trains the models using the 10 fold cross validation (10fCV) method, where by the data is split into 10 equal portions, the models are trained on 9 of them and then they predict the redshift for the $10^{th}$ portion. This is done iteratively till all the 10 portions have been predicted.
Then, SL compares the errors on the predictions of each model and combines them by assigning a coefficient, such that the final prediction has a lower error than any of the constituent models.

## MACHINE LEARNING MODELS

Machine learning (ML) is a field of computer science where we mimic the way a human brain learns. It focuses on analyzing the data and finding hidden patterns using sophisticated algorithms. In our project we train the ML models on the AGNs that have observed redshifts. We basically ask the models to understand the underlying relations between the observed photometric properties and redshift, such that we can use these relations to predict the redshift in the future.
There exist many machine learning algorithms which can be used to figure out these underlying relations. In our project we are using four methods, namely, Random Forest, XGBoost (Extreme Gradient Boosting), Bayesian GLM (Generalized Linear Models), and Big LASSO (Least Absolute Square Selection Operator).

Random Forest and XGBoost belong to the class of ML algorithms called regression trees. These work by partitioning the data based on the values of the independent variables and averaging the value of the dependent variables.

Baysian GLM is a baysian inference of the linear model. It determines the most likely estimate of the response variable (in our case the redshift) given the particular set of predictors and the prior distribution on the set of regression parameters. It works on the Fisher principle: “what value of the unknown parameter is most likely to generate the observed data”.

Finally, Big LASSO is a modification of the usual LASSO method, designed to work with large data sets.

## RESULTS

Having trained our ML models, we need to obtain a realistic estimate of how well the model performs in a real world scenario. For this, we use the 10fCV method, performed 100 times where the splits in the data are randomized. The final results are an average of these 100 iterations. We perform the procedure 100 times so as to derandomize and stabilize the results, and to ensure that we are not under or over fitting. Once we obtain the predictions of the redshifts via this process, we compare them to the observed redshifts we already have and calculate multiple statistical parameters. The statistical parameters used in order to compare our results with those of others in the field are: Bias,$\sigma_{NMAD}$(normalized median absolute deviation), Pearson correlation r, RMSE (root mean square error), and standard deviation (σ). We quote the measured values of these parameters for $\Delta z_{norm}$ and $\Delta z$, where $\Delta z_{norm} = z_{spec} - z_{pred} / (1+z_{spec})$ and $\Delta z = z_{spec} - z_{pred}$. We also quote the catastrophic outlier percentage, which is the number of predictions that lie outside the 2σ error bars.

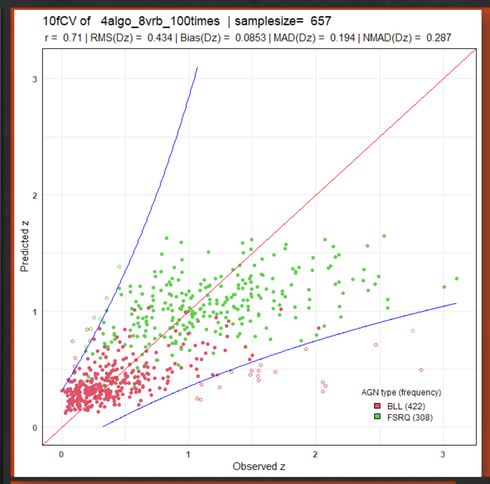


Fig 4: This shows the correlation plot between the observed and predicted redshifts.

**Correlation R = 0.71**

**RMSE = 0.434**

**Bias (Δz) = 8.5 x $10^{-2}$**
**Bias ($\Delta z_{norm}$) = 11.6 x $10^{-4}$**

**$\sigma_{nmad}(\Delta z)$ = 0.287**
**$\sigma_{nmad}(\Delta z_{norm})$ = 0.192**

**Catastrphic outliers = 5%**

| Experiment | Bias ($\Delta z_{norm}$) | Sigma ($\Delta z_{norm}$) | NMAD ($\Delta z_{norm}$) |
|---|---|---|---|
| Superlearner | 0.001 | 0.19 | 0.19 |
| Brescia et al. 2013 (best case) | 0.004 | 0.069 | 0.029 |
| Laurino et al. | 0.095 | 0.16 | ... |
| Ball et al. | 0.095 | 0.18 | ... |
| Richards et al. | 0.115 | 0.28 | ... |

Fig 5: Comparision of our results with those of others in the literature.

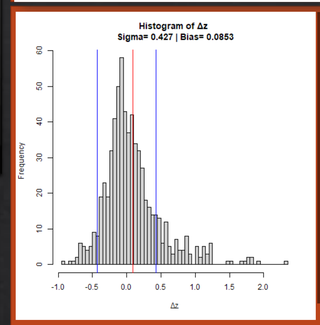


Fig 6: This shows the distribution of the Δz. Ideally we should see a histogram concentrated at 0. Due to the imbalance in our training data, we see that the difference in redshifts for high-z AGNs is large. This can only be rectified by using a more uniform distribution to train out ML models.

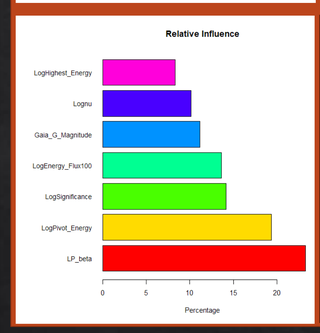


Fig 7: The relative influence of our chosen predictors are shown here. We see that the LP_beta property has the highest influence in predicting the redshift, followed by LogPivot_Energy. It is noteworthy that in case of gamma-ray loud AGNs, luminosity is not the most effective predictor of redshift, unlike in AGN classes

## PREDICTION ON THE GENERALIZATION SET

Having trained out models, we want to use it to predict the redshift of the AGNs in the generalization set. For that purpose, we first removed all the non-BLL AGNs. This was done because the generalization se is dominated by BLL and we want to investigate the redshift predictions on this category. Additionally, we also trim the generalization set so that we are not extrapolating, meaning we are not trying to predict the redshift of those AGNs which lie beyond the parameter space the model has been trained on. The histograms shown in Fig. 8 depict the predicted redshift over the observed ones.

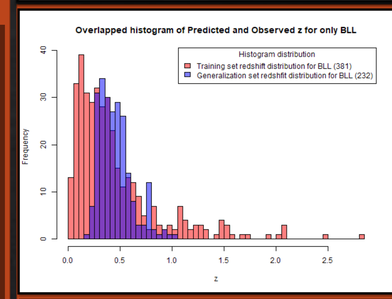


Fig 8: The histogram distribution of the generalization set over the training set, exclusively for BLLs.

## CONCLUSIONS AND FUTURE WORK

- In this project we developed a methodology to predict the redshift of gamma-ray loud AGNs using their observed properties.
- This increases the size of the 4LAC catalog with measured redshifts by 61%.
- Currently, to the best of our knowledge, no work in the blazar literature attempts to estimate the redshift using their observed γ-ray characteristics
- We aim to further improve our prediction by increasing the sample size, using more robust ML models and eliminating missing data.
- We are currently also implementing this methodology to predict the redshift of GRBs using their plateau emission parameters.

## REFERNCES

1. Ball, N. M., Brunner, R. J., Myers, A. D., et al. 2008, The 449 Astrophysical Journal, 683, 12–21, doi: 10.1086/589646
2. Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., & Mercurio, A. 2013, The Astrophysical Journal, 772, 140
3. Dainotti, M., Petrosian, V., Bogdan, M., et al. 2019, arXiv preprint arXiv:1907.05074
4. Fermi-LAT Collaboration, Abdollahi, S., Ackermann, M., et al. 2018, Science, 362, 1031, doi: 10.1126/science.aat8123
5. Laurino, O., D'Abrusco, R., Longo, G., & Riccio, G. 2011, Monthly Notices of the Royal Astronomical Society, 418, 2165–2195, doi: 10.1111/j.1365-2966.2011.19416.x
6. Liodakis, I., & Blinov, D. 2019, MNRAS, 486, 3415, doi: 10.1093/mnras/stz1008
7. Richards, G. T., Myers, A. D., Gray, A. G., et al. 2008, The Astrophysical Journal Supplement Series, 180, 67–83, doi: 10.1088/0067-0049/180/1/67