arXiv:2107.10952v1 [astro-ph.HE] 22 Jul 2021

Predicting the redshift of γ-ray loud AGN using supervised machine learning

Maria Giovanna Dainotti,[1,2] Malgorzata Bogdan,[3,4] Aditya Narendra,[5] Spencer James Gibson,[6]
Blazej Miasojedow,[7] Ioannis Liodakis,[8] Agnieszka Pollo,[9,10] Trevor Nelson,[11] Kamil Wozniak,[12]
Zooey Nguyen,[13] and Johan Larrson[4]

[1]National Astronomical Observatory of Japan, Mitaka
[2]Space Science Institute, 4750 Walnut St, Suite 205, Boulder,CO,80301,USA
[3]Department of Mathematics, University of Wroclaw, Poland
[4]Department of Statistics, Lund University, Sweden
[5]Jagiellonian University, Poland
[6]Carnegie Mellon University, USA
[7]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland
[8]Finnish Center for Astronomy with ESO (FINCA), University of Turku, Finland
[9]Astronomical Observatory of Jagiellonian University, Krakow
[10]National Centre for Nuclear Research, Warsaw
[11]University of Massachusetts at Amherst, Massachusetts, USA
[12]AGH University of Science and Technology, Krakow
[13]Faculty of Astronomy, University of California, Los Angeles, California, USA

Submitted to APJ

ABSTRACT

AGNs are very powerful galaxies characterized by extremely bright emissions coming out from their central massive black holes. Knowing the redshifts of AGNs provides us with an opportunity to determine their distance to investigate important astrophysical problems such as the evolution of the early stars, their formation along with the structure of early galaxies. The redshift determination is challenging, because it requires detailed follow-up of multi-wavelength observations, often involving various astronomical facilities. Here, we employ machine learning algorithms to estimate redshifts from the observed γ-ray properties and photometric data of γ-ray loud AGN from the Fourth Fermi-LAT Catalog. The prediction is obtained with the Superlearner algorithm, using LASSO selected set of predictors. We obtain a tight correlation, with a Pearson Correlation Coefficient of 71.3% between the inferred and the observed redshifts, an average $\Delta z_{norm} = 11.6 \times 10^{-4}$. We stress that notwithstanding the small sample of γ-ray loud AGNs, we obtain a reliable predictive model using Superlearner, which is an ensemble of several machine learning models.

1. INTRODUCTION

Active Galactic Nuclei (AGN) with jets are the dominant class of objects when it comes to high-latitude ($|b| > 10$) extragalactic γ-ray sources (Abdollahi et al. 2020). The *Fermi* γ-ray space telescope has detected more than 2863 such γ-ray AGNs, the majority of which ($> 98\%$) are blazars: AGN with their jets pointed towards our line of sight. Blazars are denoted by the equivalent width of resonant emission lines in their optical spectra. Sources with broad emission lines are classified as Flat Spectrum Radio Quasars (FSRQs), whereas sources with weak or no emission lines are classified as BL Lacertae objects (BLLs). Measuring the redshift (z) of blazars has been a cumbersome and

Corresponding author: Maria Giovanna Dainotti
maria.dainotti@nao.ac.jp

observationally expensive endeavor. The situation is further complicated by the absence of emission lines in the most numerous class of γ-ray loud blazars, i.e., BL Lacs. As a result, out of the 2863 sources of the Fourth AGN *Fermi*-LAT catalog (4LAC, Ajello et al. (2020)), only 1591 have redshift estimates, ranging from $z = [0, 3]$, but mostly concentrate below $z = 2$. γ-Ray loud blazars with redshift estimates are relevant for our comprehension of the origin of the Extragalactic Background Light (EBL), which in turn let us probe the cosmic evolution of blazars (e.g., Singal et al. (2012), Singal et al. (2014), Singal (2015), Singal et al. (2013a), Chiang et al. (1995), Ackermann et al. (2015), Singal et al. (2013b) Marcotulli et al. 2020), the intergalactic magnetic field (e.g., Venters & Pavlidou 2013), star formation rate history of our universe (e.g., Fermi-LAT Collaboration et al. 2018), as well as constrain cosmological parameters (e.g., Domínguez et al. 2019). The difficulty in spectroscopically measuring redshift in a significant fraction of BL Lacs and the importance of identifying high-z blazars has led to the development of photometric estimation techniques (**photo-z**, e.g., Kaur et al. 2017, 2018; Rajagopal et al. 2020; Carrasco et al. 2015; Krakowski et al. 2016; Nakoneczny et al. 2019). However, works using such methods typically produce redshift estimates for only $\sim 6 - 13\%$ of their sample, making alternative methods necessary. Machine learning (ML) methods for obtaining photo-z estimates for AGN are becoming increasingly important in the era of big data Astronomy (e.g., D'Isanto & Polsterer 2018; Brescia et al. 2013; Brescia et al. 2019; Ilbert et al. 2008; Hildebrandt et al. 2010). Here we focus on the γ-ray emitting AGN population in the 4LAC.

In the current literature, multiple works exist which focus on extracting reliable photometric redshift of AGNs (Cavuoti et al. 2014; Fotopoulou & Paltani 2018; Logan & Fotopoulou 2020; Yang et al. 2017; Zhang et al. 2019; Curran 2020; Nakoneczny et al. 2020; Pasquet-Itam & Pasquet 2018; Jones & Singal 2017). In the current blazar literature, a lot of effort has also been placed in classifying blazars of uncertain type (e.g., Chiaro et al. 2016; Kang et al. 2019) and unidentified *Fermi* objects (e.g., Liodakis & Blinov 2019). Although these papers convey useful information about the algorithms that work well for classifying blazars, so far no analysis has been performed regarding the prediction of the redshifts of γ-ray loud blazars. Thus, we will tackle this problem by using machine and statistical learning algorithms. We apply multiple ML algorithms, such as LASSO (Least Absolute Shrinkage and Selection Operator), XGBoost (Extreme Gradient boosting), RandomForest, and BayesGLM (Bayesian generalized linear model). We follow the approach used in Dainotti et al. (2019), where some of us used the SuperLearner package to aggregate the results from multiple algorithms and predict the redshifts of γ-ray bursts.

The results of this study increases the number of blazars with inferred redshifts considerably so that we can finally obtain a more complete sample of γ-ray loud AGNs. As a result, this work will enable the solving of some crucial questions on the luminosity function and density evolution of γ-ray loud AGNs.

In Section 2, we discuss the data and predictors used. In Section 3, we outline the ML methods used, the selection of the best predictors and algorithms, and the validation of our results. In Section 4, we present the results obtained in this analysis. In Section 5, we present our results and discuss future perspectives.

2. **THE SAMPLE**

Fermi-LAT has been continuously monitoring the sky in the 50 MeV to 1 TeV range since 2008. The γ-ray properties used in this work are obtained from the 4LAC catalog (Ajello et al. 2020). **It contains 2863 sources, 658 of which are FSRQs, 1067 are BL Lacs, 1074 are blazars of uncertain type, and the remaining 64 sources are classified as radio galaxies, Narrow line Seyferts (NLSY1), and other non-blazar AGNs. Out of the 2863 sources, 1591 have a measured redshift, whose distribution is shown in Fig. 1.** For completeness of the treatment we have included also non BL LAC and non FSRQs sources in the initial scatter matrix plot in Fig. 3 to show how the variables in the sample is distributed. But, in the generalization set, we are predicting the redshift for only the BLLs.
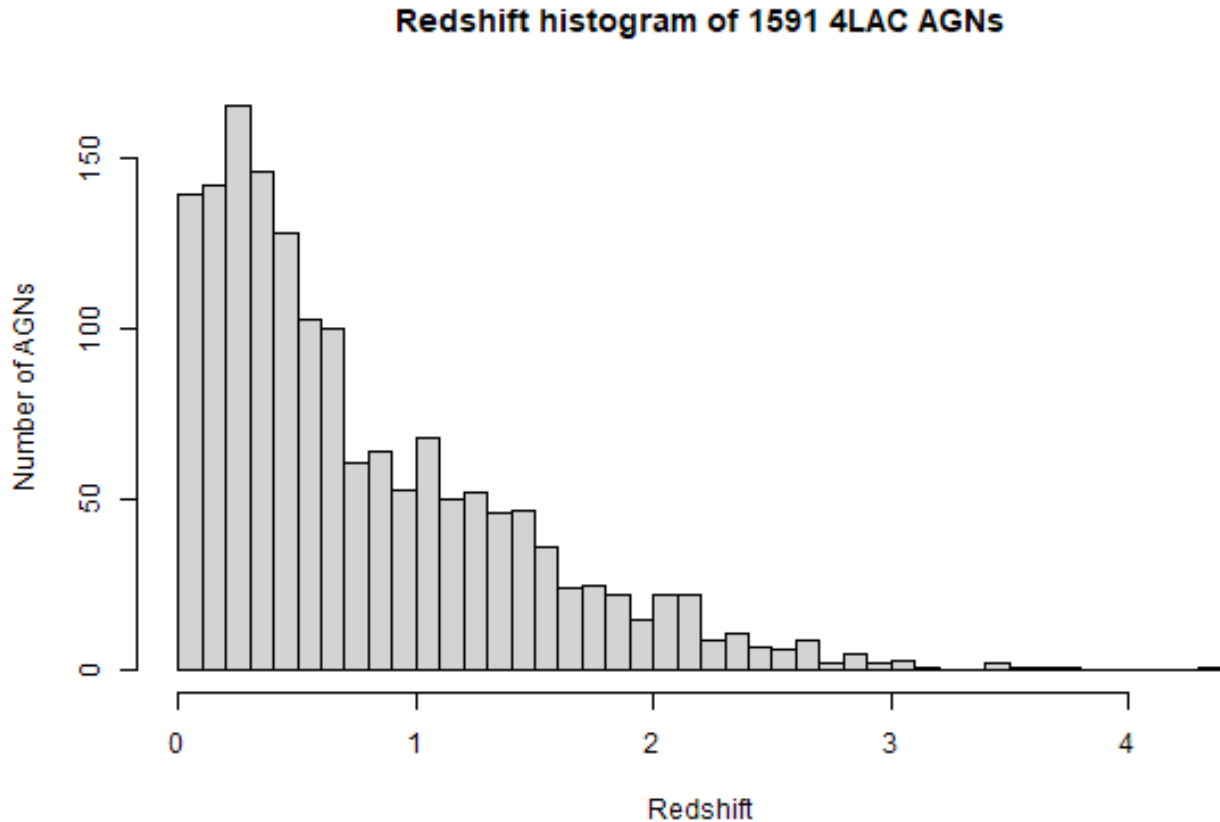
Figure 1. The redshift distribution of the entire 4LAC catalog before selection cuts and outliers removal.

Unfortunately, all of the 1591 γ-ray AGNs cannot be used for our model's training. A significant number of these γ-ray AGNs have incomplete observational data, meaning we face the problem of missing values in several parameters.

Thus, we perform cuts in the data set to remove incomplete data points leaving us with 1169 γ-ray AGNs out of 2863. These consist of 661 BLLs, 309 FSRQs, 177 unclassified AGNs, and 22 AGNs belonging to other categories. This set is split into training and generalization sets, the former consisting of the γ-ray AGNs that have observed spectroscopic redshift, while the latter consists of the γ-ray AGNs for which the redshift is not measured. Our training set consists of 793 γ-ray AGNs, made up of 422 BLLs, 308 FSRQs, 41 unclassified, and 22 other category AGNs. The 22 other category γ-ray AGNs in our training set consisted of 2 NLSY1 sources, 3 Compact-Steep spectrum Radio Source (CSS) sources, 13 Radio Galaxies (RDG) sources, and 2 sources classified as non-blazar AGNs. They are shown in Fig. 3. After we perform the cuts related to the missing data we are left with 730 γ-ray AGNs. Similarly, our generalization set consists of 376 γ-ray AGNs, of which are 239 BLLs, 1 FSRQ, and are 136 unclassified AGNs. After we perform the cuts in the generalization set we are left with 239 BLLs. Due to their dominating presence, we perform our predictions only for BLLs, and remove the 136 uncategorized AGNs. But, in the scatter matrix plot of Fig. 6, we show in black the only FSRQ from the generalization set.

BL Lacs and FSRQs can be very easily separated as we did when we have introduced in the Superlearner categorical variables. We here stress that this is an important point, because it means that the quality of the predictions will most probably differ, especially if the fractions of BL Lacs in the training sample and in the full population are very different. This is expected as we have already mentioned in the introduction that this could be the case because of the difficulty of obtaining their spectroscopic redshift. We also would like to stress that due to the paucity of the other classes the categorical variables have been limited to BLL and FSRQs.
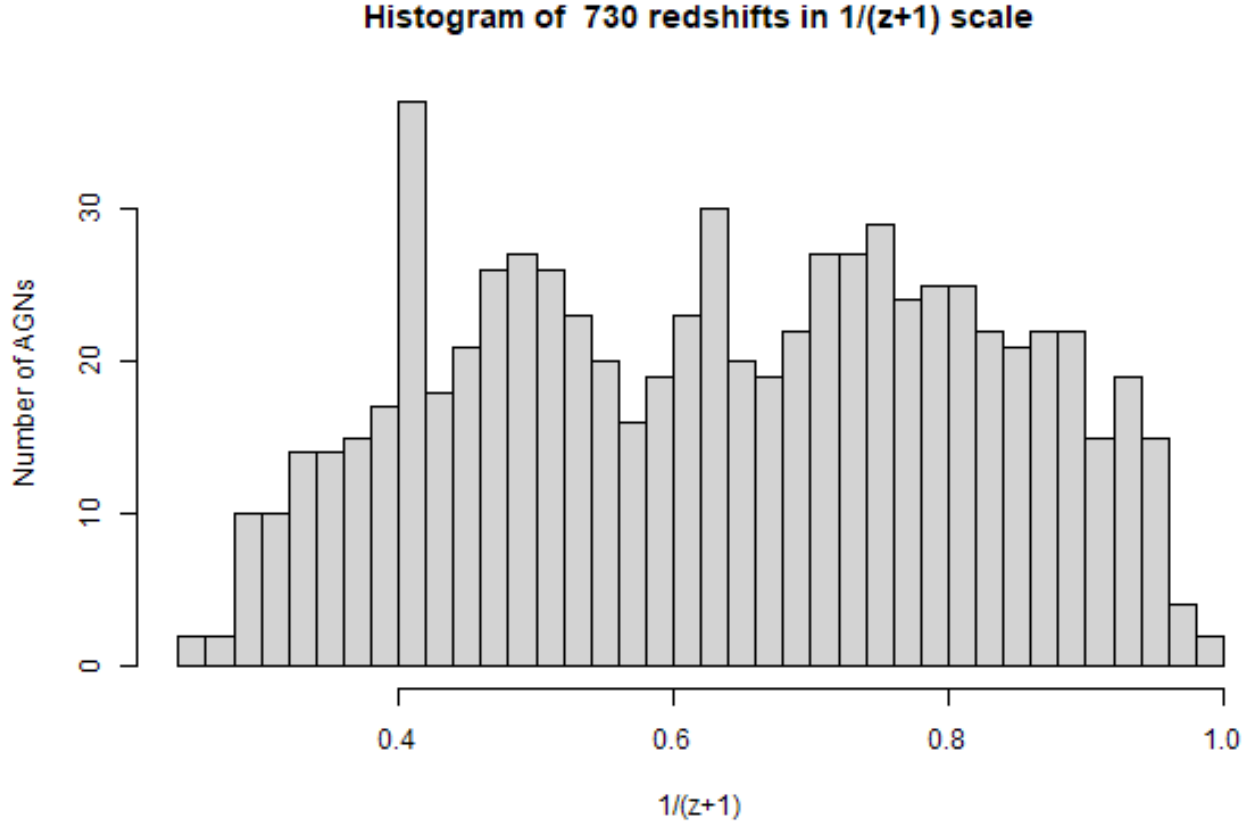
Figure 2. The histogram distribution of the redshift of our training set in 1/(z+1) scale

Regarding the predictors, 4LAC contains 13 photometric variables along with the spectroscopic redshift and names of the AGNs. It also includes the g-band magnitudes for individual sources from Gaia (Jordi et al. 2010). 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 $\frac{1}{z+1}$ (see Fig. 2). Out of these 13 variables, we take into consideration 11. We exclude fractional variability due to the incompleteness of the AGN sample and Logνfν as it is a second-order variable depending on Logν. The definition and explanation for the 11 variables are given below.

- *LogFlux* - Logarithm in the base of 10 of the integral photon flux, in photons/cm2/s, from 1 to 100 GeV.

- *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.

- *LogVariability_Index* - The sum of the log(likelihood) difference between the flux fitted in each time interval and the average flux over the 50 MeV to 1 TeV range.

- *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%.

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

- *PL_Index* - It is the photon index when fitting the spectrum with a power law, in the energy range from 50 MeV to 1 TeV.

- *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_Index* - Photon index at pivot energy (α) when fitting the spectrum (100 MeV to 1 TeV) with Log Parabola.

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

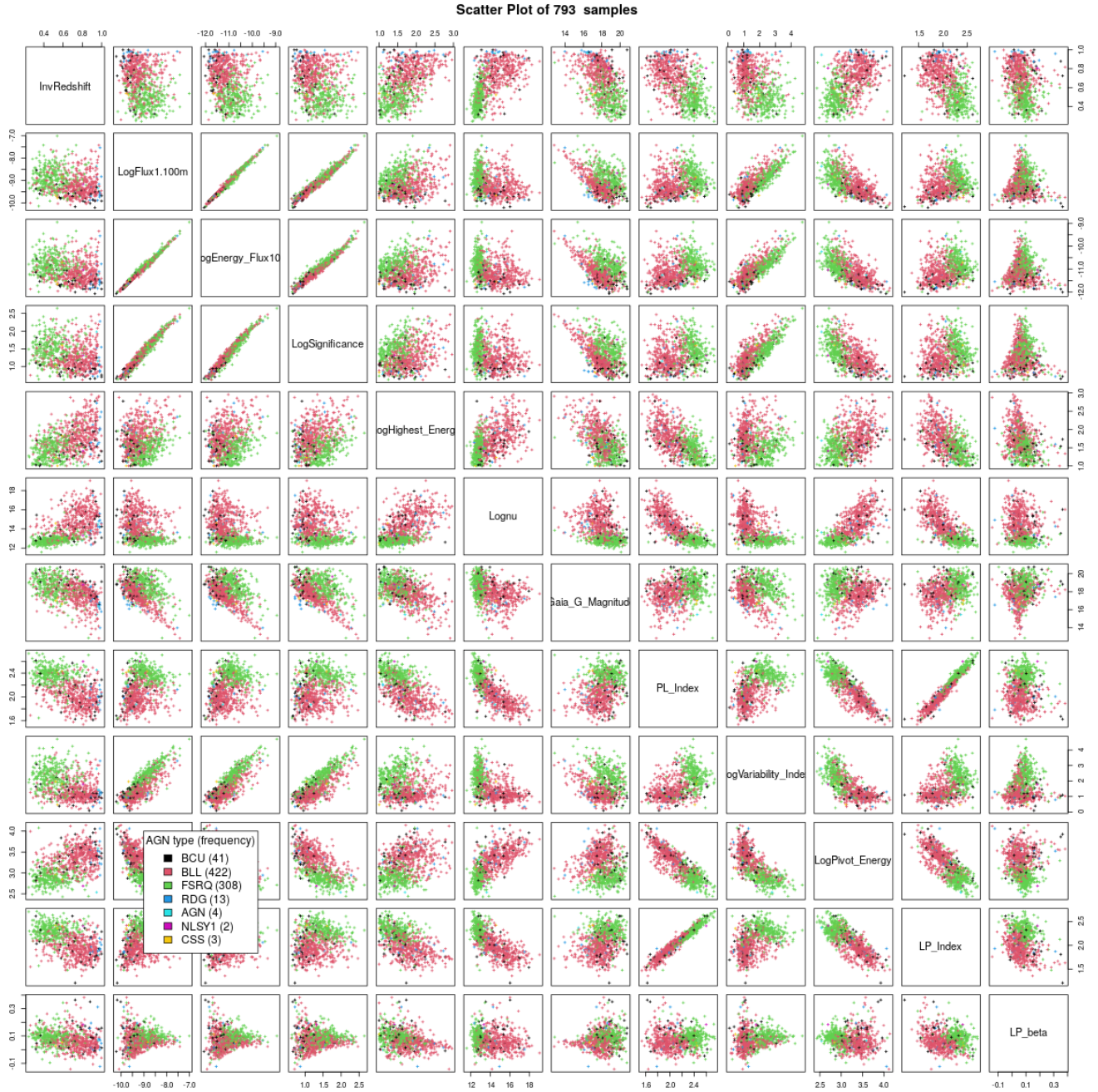- *Gaia_G_Magnitude* - Gaia Magnitude at the g-band provided by the 4LAC, taken from the Gaia Survey.

Figure 3. The full scatter matrix plot of all the variables defined above, before feature selection. Here the *InvRedshift* denotes $\frac{1}{z+1}$ scaled data.

3. METHODOLOGY

In this section, we describe, in detail, the methodology adopted for this study, from the description of the choice of the transformations adopted, the variable selection, the methods considered singularly, such as Big LASSO (a more reliable version of LASSO), XGBoost, Random Forest and Bayes GLM, to the Superlearner algorithm used to create the ensemble leading to the final prediction (see Sec. 3.4).

The statistical parameters used in order to compare our results with those of others in the field are: Bias, σ_{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 Δz_{norm} and Δz , where $\Delta z_{norm} = \frac{z_{spec} - z_{pred}}{(1 + z_{spec})}$ and $\Delta z = z_{spec} - z_{pred}$. As shown in the scatter matrix of Fig. 3, we can see the presence of multiple correlated variables such as PL_Index and LP_Index, LogEnergyFlux and LogFlux, and LogFlux and LogSignificance. Hence, we deploy a feature selection method such as LASSO which as a result naturally reduces the number of correlated variables, although it does not completely eliminate all of them.

The procedure consists of mainly two parts, as presented in the flowchart in Fig. 4. The first steps are to clean our data source by eliminating data points with missing variables and then pruning our feature set with the use of the LASSO algorithm. After this, the variables obtained as the selected ones will be used to train our model. We split our data into train and test sets composed of 657 γ-ray AGNs, and the validation set composed of 73 γ-ray AGNs. We divide the sample taking as the validation set the latest 10% of the γ-ray AGN observed. This choice is the same as taking the validation set randomly since there is no preferential order in redshift when we choose the validation set. This is just for one test, but as we show in the Sec. 3.4 we also apply the 10-fold cross-validation (hereafter called 10fCV) 100 times to avoid choosing a validation sample that may not be representative of the whole sample. We will use Superlearner which includes the optimized XGBoost, Random forest, Bayes GLM, and Big LASSO. Details of such an optimization are mentioned in Sec. 3.3. After training this ensemble on our data, we obtain our trained model, which leads us to the prediction on the redshifts.
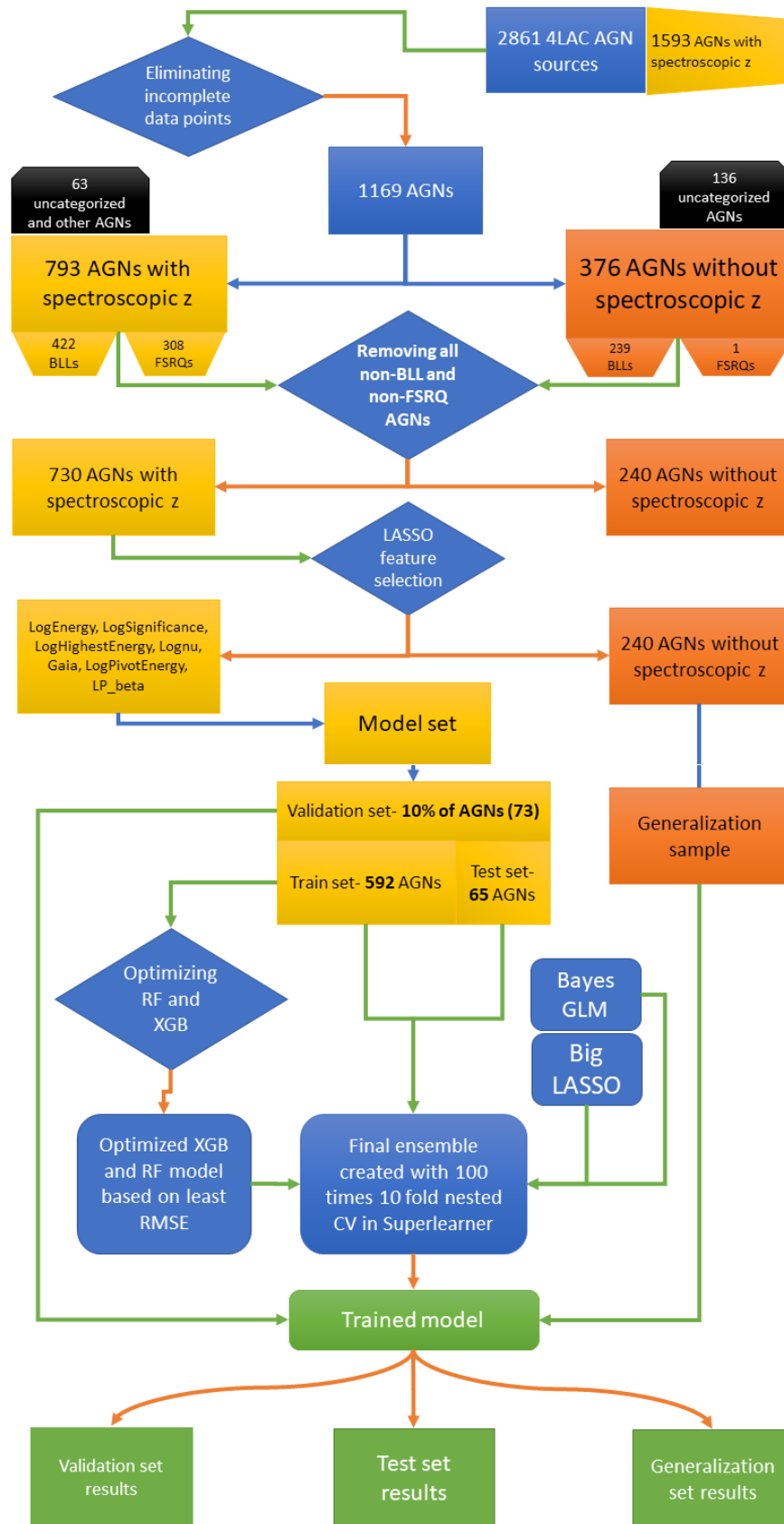
Figure 4. Methodology flowchart: the rectangular boxes represent data sets, the parallelograms the γ-ray AGN categories, the rhombus indicates functions performed, rounded rectangles indicate the ML algorithms used, the green lines show the direction of the input, orange lines the output, and blue lines indicate the splits and changes in the data set. The color-coding indicates the following: yellow indicates the data with spectroscopic z, orange the ones without spectroscopic z, green the results and, blue indicates the intermediate steps or datasets.

3.1. *Feature selection*

We apply the LASSO method to prune our features and obtain a more effective subset for redshift prediction. The 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). This penalization allows the model to select a subset of features and discards the rest by setting their coefficients to 0 (Tibshirani 1996). The tuning parameter is responsible for deciding the shrinkage coefficient applied to the estimated vector. As a consequence, the model is easier to interpret with a smaller number of features and usually has a smaller prediction error than the full model. The prediction error is the RMSE between the predicted and the observed redshifts, which is minimized during the one hundred times 10fCV training. As a measure of the prediction errors we quote the RMSE value, as well as the σ_{NMAD}. For our analysis, we use the GLMNET function with the LASSO selection feature (Hastie et al. 2017; Tibshirani et al. 2012). We pick the $\lambda.1se$ value, which is the maximum λ value for which the error is within 1 standard deviation (Friedman et al. 2010a) and its corresponding coefficients for the features. The coefficients assigned by LASSO to each of them are displayed in Fig. 5 and we choose only the non-zero coefficient features. To better visualize the parameter space of these features we plot them in the scatter matrix plot shown in Fig. 6 , along with the generalization set. LASSO feature selection shows that some of the variables that were strongly correlated are naturally eliminated, but we are still left with two correlated variables: LogEnergyFlux and LogSignificance. This means that for LASSO both features are relevant. Since LogSignificance is providing the information on the detectability of the γ-ray AGN and this is relevant to the final prediction of the redshift; thus, we decided to retain it. On the other hand, from a statistical point of view, it is not necessary to remove correlated variables, since the aim here is to reach a greater accuracy on the prediction of the redshift. Nevertheless, we have shown in the Appendix (See Fig. 17) that the results do not change at the level of 1% for σ_{NMAD} (Normalized Median Absolute Deviation), RMSE and Correlation when we consider to manually discard this variable.
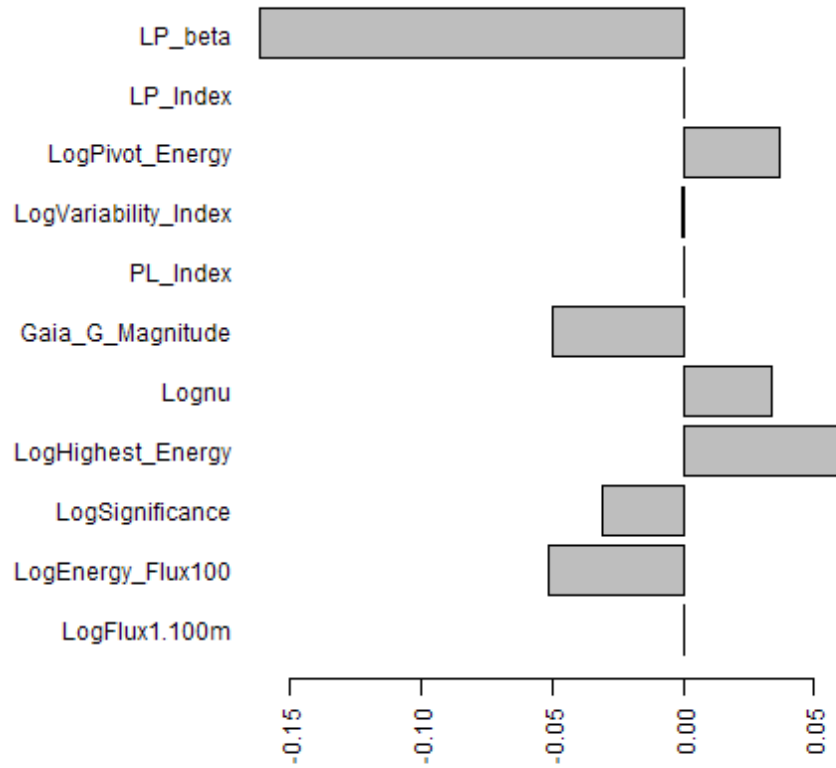
Figure 5. The coefficients assigned to the features by LASSO at the λ.1se value. We only keep the coefficient features > 0.
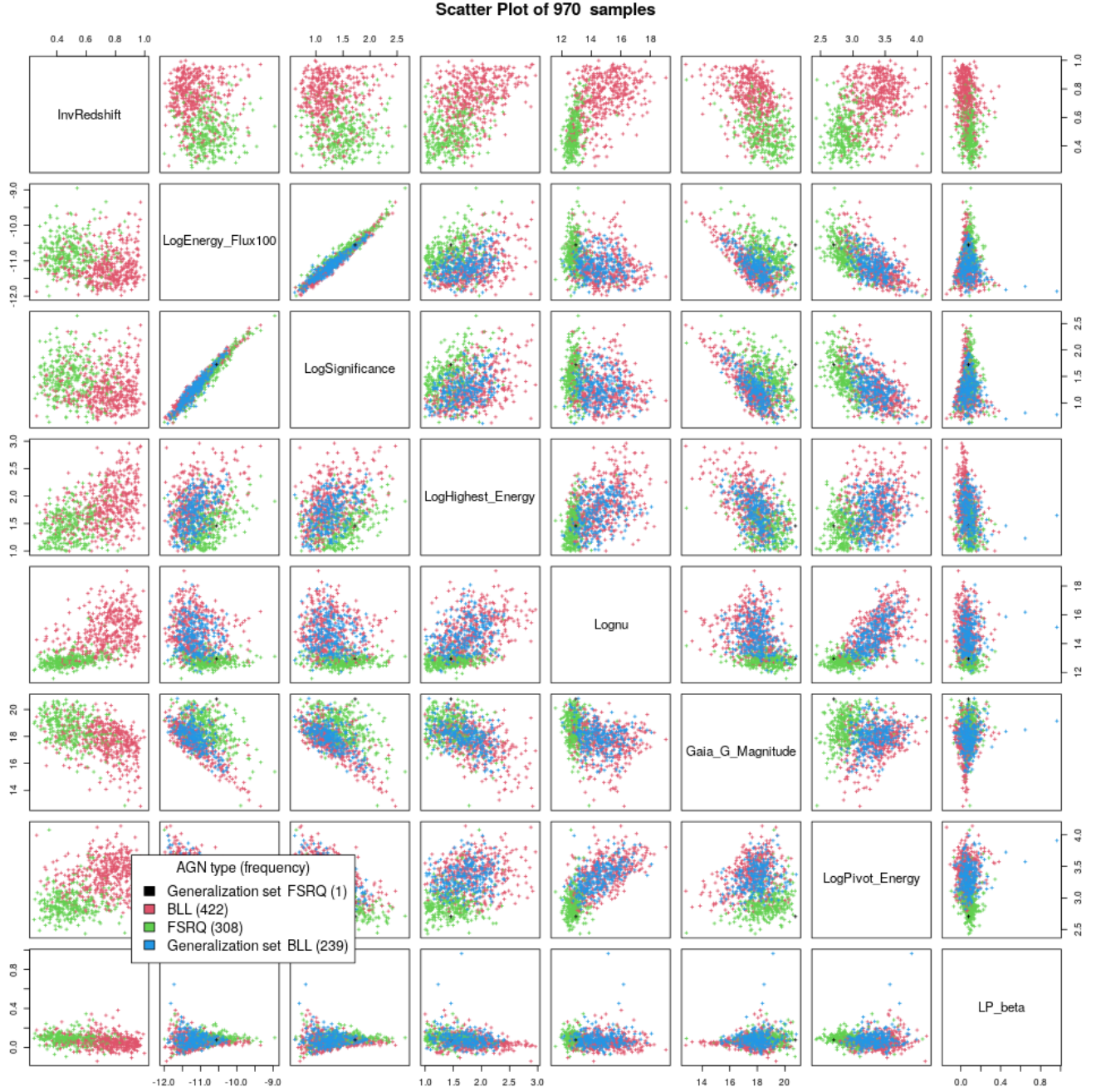
Figure 6. The full symmetric scatter matrix plot shows the response (in our case the *InvRedshift*) and predictor variables. The different γ-ray AGN categories are color-coded according to the legend displayed on the plot. The values in the parenthesis indicate the number of γ-ray AGNs present in the data set.

In addition, we clarify that we performed the analysis with both $\log_{10}(1 + z)$ and $\frac{1}{z+1}$, the distribution of the latter shown in Fig. 2. The choice of transformation arises from the fact that the results related to the choice of $\frac{1}{z+1}$ present the smallest σ_{NMAD} and smaller Δz_{norm} (normalized variation in redshift), thus leading us to use this transformation.

3.2. *The ML algorithms used in our analysis*

By adopting an ML approach, we leverage the built-in algorithms that learn from the training set and we test out predictions on the test set. We employ the trained models to predict the redshift of sources for which the redshift
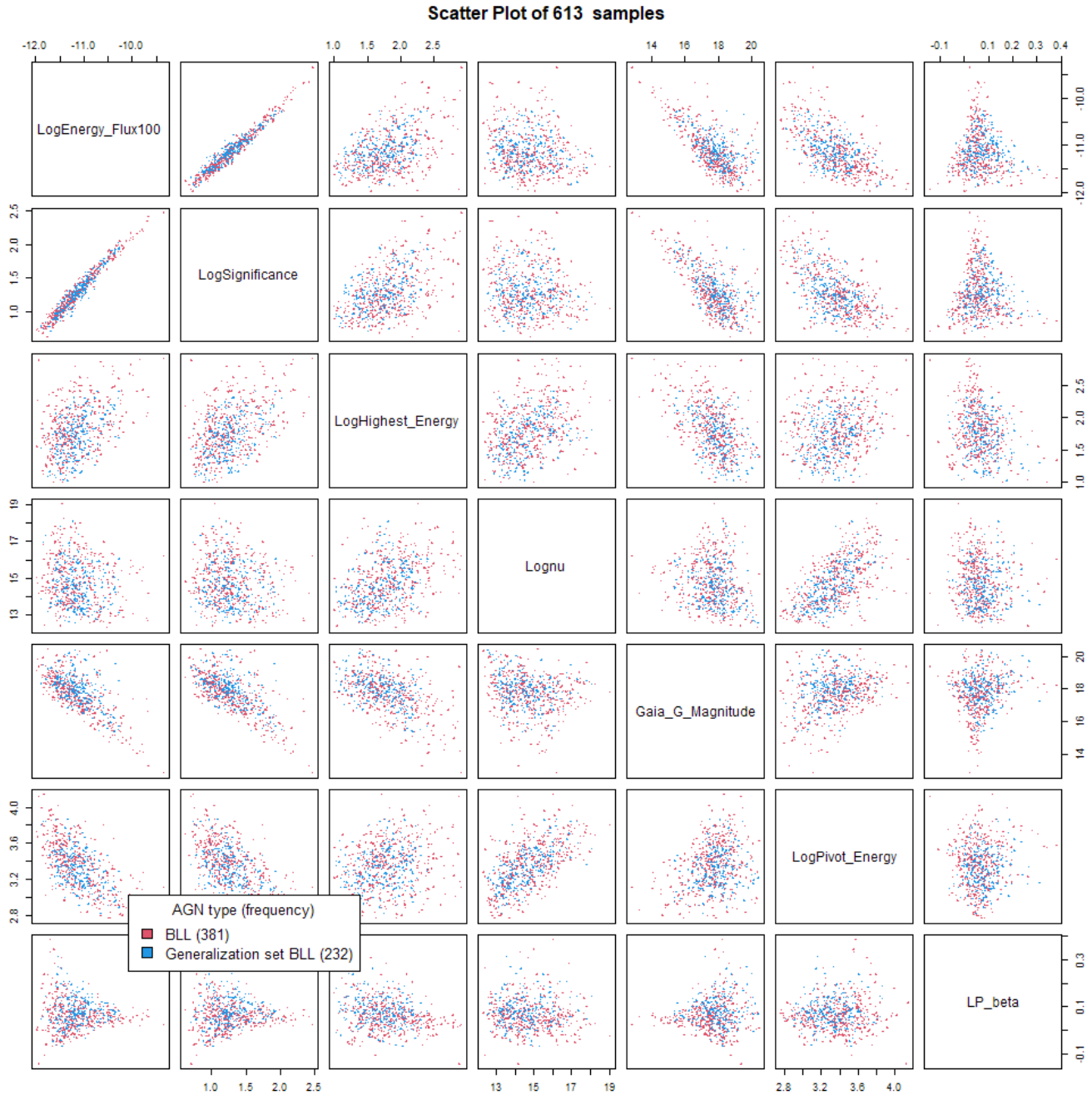
Figure 7. The scatter matrix plot for BLLs in the generalization and training set. The generalization set BLLs are shown in blue, while the training set BLLs are shown in red.

has not been measured. These optimized methods are combined into an ensemble using the Superlearner package, providing us with a better prediction than any single algorithm. The ML algorithms used here are summarized in the following itemized points:

- Regression trees build the predictor by partitioning the data based on the values of the independent variables and averaging the value of the dependent variables. Examples of regression trees are XGBoost and Random Forest. Indeed, both the XGBoost and Random Forest algorithms utilize multiple regression trees to increase their predictive power.

- The Random Forest algorithm generates multiple independent regression trees and averages them to obtain a more accurate prediction (Breiman 2001; Valencia et al. 2019; Green et al. 2019; Miller et al. 2015). An extremely difficult task is how to choose the optimal depth of such a tree, namely to decide which is the number of partition levels. In gradient boosting, the final predictor is built as a weighted sum of simple tree predictors. Compared to the Random Forest method, regression trees are not generated independently but built on each other using residuals from the previous step, until the culmination of trees forms a stronger regression model.

- The XGBoost algorithm is an amelioration of the gradient boosting method (Chen & Guestrin 2016; Friedman et al. 2000; Friedman 2001, 2002) and it also leverages poor predictors. It uses a more regularized model formalization to control overfitting, and thus give better performance.

- Big LASSO is a computationally efficient implementation of the LASSO algorithm in R (Zeng & Breheny 2017). The Big LASSO is an implementation that allows us to compute and analyze big multidimensional data sets quickly and efficiently.

- Bayes GLM is a bayesian inference of the generalized 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 (Maximum A Posteriori estimator, MAP). It works on the Fisher principle: "what value of the unknown parameter is *most likely* to generate the observed data". BayesGLM method is more numerically and computationally stable as compared to normal GLM models. It employs a student-t prior distribution for the regression coefficients. Then, given the observed data, the likelihood function for these parameters is calculated. The likelihood function and priors are combined to produce the posterior distributions from which we obtain the MAP estimators of the desired parameters (Birnbaum 1962; Hastie & Tibshirani 1987, 1990; Friedman et al. 2010b).

3.3. *Optimizing Algorithms*

It should be noted that these results are obtained after performing 10fCV on our data set. For the XGBoost algorithm, we have the option to vary the number of regression trees, the depth, and the learning rate (the so-called shrinkage coefficient, which shrinks the predictions of a tree to prevent over-fitting). We tune these to best fit our data without over or under-fitting. In Fig. 8 top left and right panels show the variation of the root mean square error (RMSE) and correlation, respectively, related to the number of trees in the model. The RMSE and correlation minimize and maximize, respectively, at a depths of 5 and at a number of 500 trees. However, since depth 4 and 5 gives very similar results, to avoid the risk of over-fitting usually associated with a higher depth we choose a max depth of 4 and proceeded to test the model performance while varying the learning rate, see bottom panels of Fig. 8. The optimal learning rate in our case is 0.01. In the left bottom panel of Fig. 8, we plot the RMSE variation, and on the right the Pearson Correlation coefficient (r). In summary, our final XGB optimized model consists of 500 trees, with a depth of 4 and a shrinkage coefficient of 0.01.

A similar analysis is performed for Random Forest as well. We tune the number of trees, depth, and the maximum number of nodes based on which model has the lowest RMSE and maximum correlation value. We started with a default value for the number of variables that will be randomly sampled (from here on denoted as mtry), which is 2. We vary the number of trees and the maximum number of nodes. The RMSE and Correlation variation are shown in the top left and right plots of Fig. 9, respectively. We observe that a value of 200 for maximum nodes gives the least RMSE and maximum correlation at 400 trees. Next, we keep the maxnode parameter constant and vary the mtry value from 2 to 4. The RMSE and Correlation plots are shown in the bottom panel of Fig. 9. Among the different values of mtry tested, we see that mtry=2 gives us the best results in terms of the highest correlation coefficient and the smallest RMSE. Furthermore, the number of trees is selected to be 600, as this gives the second smallest RMSE, but since in this region we have contemporaneously also the plateau of the Correlation coefficient (see left bottom panel of Fig. 9) 600 is the most favored value. In addition, when the RMSE is similar as in the 600 and 900 trees we prefer the smaller number of trees to prevent overfitting. In the case of BayesGLM, there are no tuneable hyperparameters, as instead it is for XGBoost and RF. Instead, we specify a formula based on which the redshift is predicted. The formula used is a linear combination of all the features we consider:

$$\frac{1}{z_i + 1} = f(\sum K_i) \tag{1}$$

Here K belongs to a set of features described in Sec. 3.1 and presented in Fig. 5, and i denotes each γ-ray AGN in the training set which is used in the model fitting.

The Big LASSO algorithm is an extension of LASSO. Hence its optimization is done identically, i.e its λ hyperparameter is tuned based on its internal CV such as to obtain the model with the least RMSE. As a result, there is no need for us to explicitly handle its optimization.
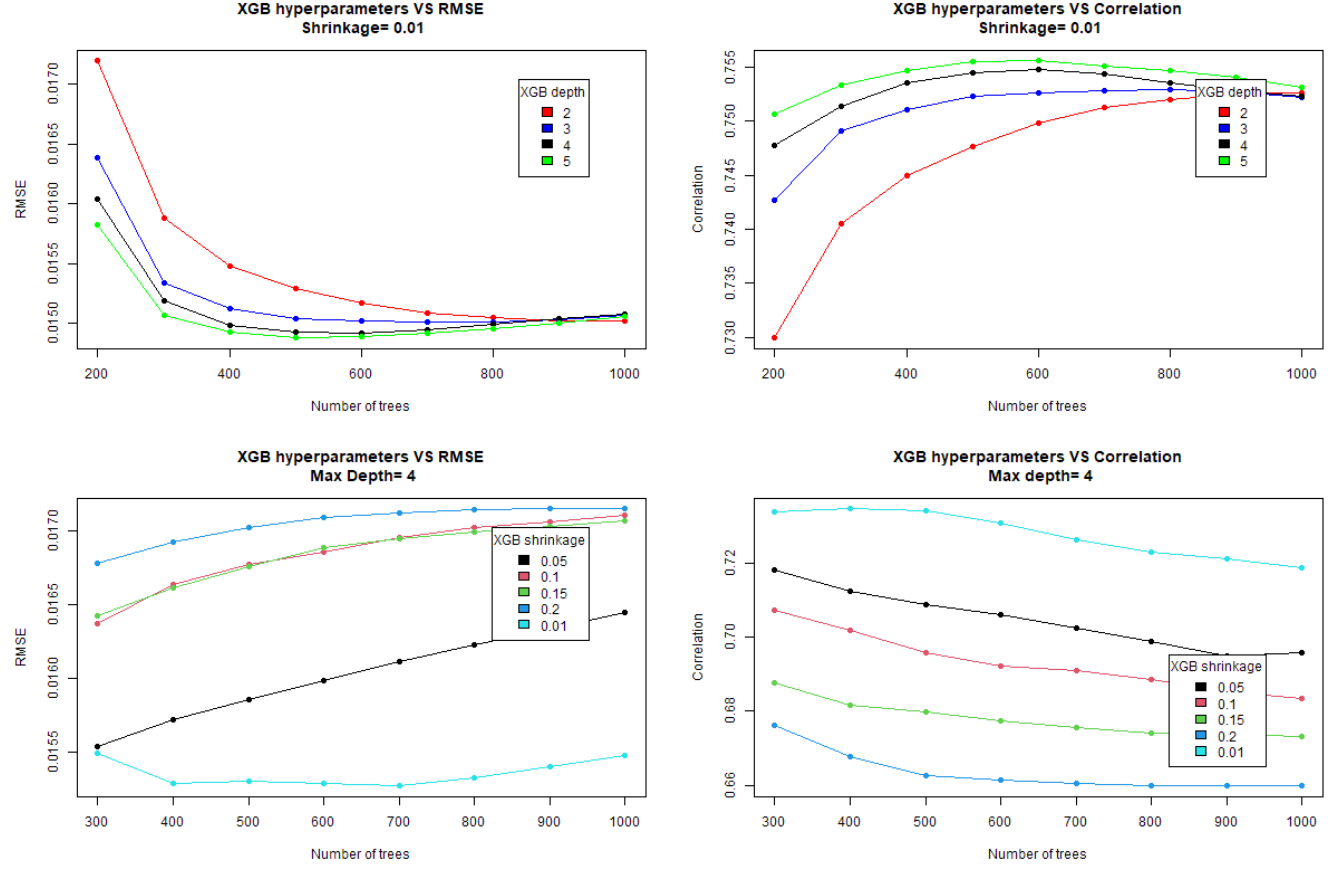

Figure 8. Variation of the RMSE and Correlation coefficient versus the number of trees, for different depths (upper panels) and shrinkage coefficients (lower panels).

Figure 9. The panels show random forest optimization plots. The upper left and right panels present the RMSE and Correlation vs. the number of trees, respectively. This is performed with a fixed value of mtry=2 and different values of RF maxnodes=(50, 100, 150, 200) color-coded with red, blue, black, and green, respectively. The bottom left and right panels present the same plots as the upper panel, but with the fixed value of Maxnodes=200 and with mtry=2,3,4 indicated with red, blue and black, respectively.

Since every ML method has its advantages in a given parameter space and in our case in different redshift ranges, we leverage each of the methods by using Superlearner, described in the next subsection.

3.4. *SuperLearner*

In our approach, we have three different types of sets: the training, the test, and the generalization sets. The training set is used to train the model based on the observed variables for which we already know the response variable, while the test set is used to validate the accuracy of the model, the generalization set is the one for which the redshift is unknown and the ML algorithm is applied for inferring this information. First, we use LASSO and select important features based on the data from the training set. Then, we construct the prediction model using the Superlearner ensemble algorithm which includes the optimized XGBoost, Random Forest, Bayes GLM, and Big LASSO. In our case, since the test set has never been used in the training set, then it is called validation data set.

SuperLearner (Van der Laan et al. 2007) is an algorithm that utilizes k-fold CV to estimate the performance of ML algorithms. It creates an optimal weighted average of the input models, i.e., an ensemble. Namely, the SuperLearner provides coefficients that reflect the relative importance of each learner against the others in the ensemble. Besides this feature, Superlearner can test the predictive power of multiple ML models or the same model, but with different settings. The weights of the algorithms always sum up to 1 and are always equal to or greater than 0. Using these coefficients, we can group the highest weighted algorithms into an ensemble and improve the prediction more than any single algorithm (Polley & Van der Laan 2010).

We use the functions implemented in the statistical software R, particularly the SuperLearner package.

In 10fCV the dataset is randomly partitioned into 10 complementary subsets. The SuperLearner is trained on 9 of these subsets and the resulting model is employed to infer the values in the remaining subset, which plays the role of the test set. The process is iterated 10 times, with each subset playing the role of the test set. The SuperLearner parameters are automatically set to optimize the prediction for all test sets (i.e., all data points). Following statistical practice, we repeat this whole procedure 100 times to make the prediction less dependent on the selection of the specific random partition of the dataset. Thus, our predictions result as the average of 100 independent SuperLearner predictions. This allows for stabilization and de-randomization of our results. Given the paucity of our dataset, this is a crucial step in analyzing the performance of our model.

4. **RESULTS**

Our final training set consists of 657 γ-ray AGNs with observed redshifts. We separate 73 γ-ray AGNs as a validation set that is not used for any training (see Fig. 4).
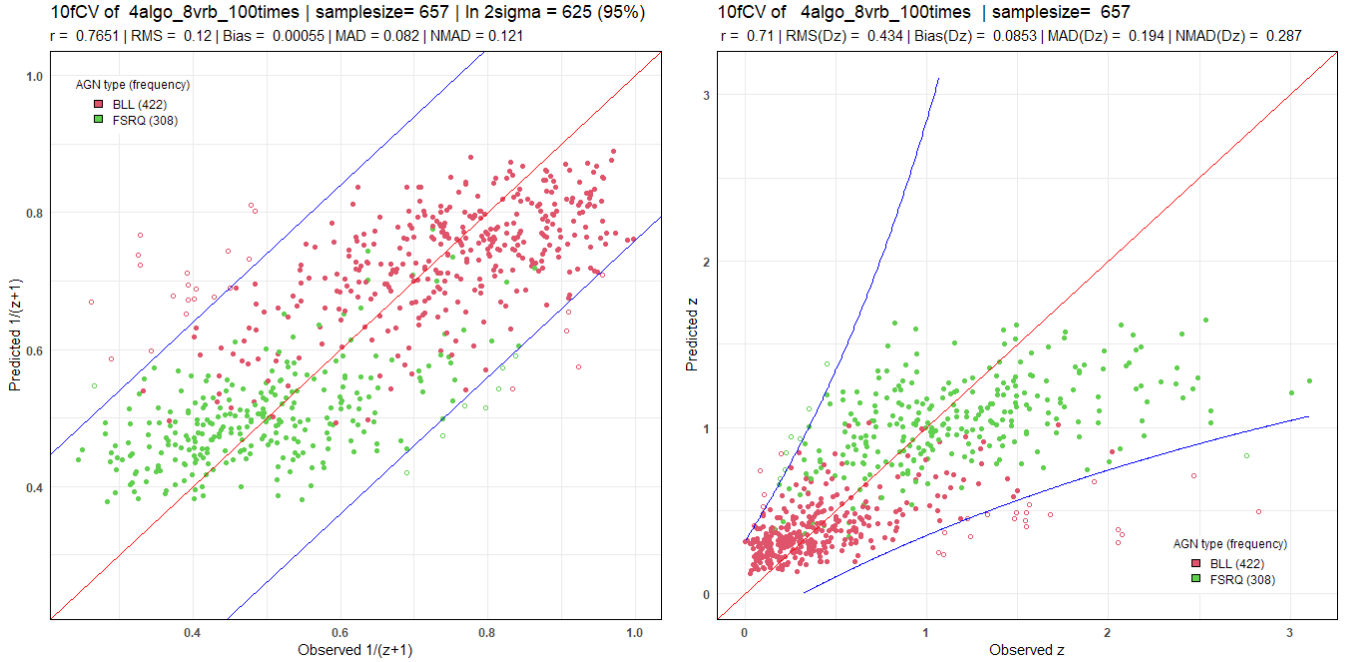

Figure 10. The left panel shows the observed vs. predicted redshift in the $\frac{1}{z+1}$ scale, while the right panel shows the observed vs. predicted redshifts in the linear scale.

In Fig. 10 the top panel shows the correlation plot between the observed and predicted redshift in $\frac{1}{z+1}$ (left panel) and linear scale (right panel). The blue lines indicate the 2σ cones for each of the plots where the σ is calculated in the $\frac{1}{z+1}$ scale as follows:

$$\frac{1}{z_p+1} = \frac{1}{z_s+1} \pm 2\sigma,$$

where z_s is the spectroscopic redshift and z_p is the photometric redshift. Due to the choice of our scaling, the 2σ line is not straight on the linear scale and is shown in the following formula-

$$z_p = z_s \left[\frac{1 \pm 2\sigma(z_p+1)}{1 \mp 2\sigma} \right] \pm \frac{2\sigma}{1 \mp 2\sigma}.$$

We obtain a Pearson Correlation $r = 0.71$ in the linear scale, with the $\sigma_{NMAD}(\Delta z_{norm}) = 0.192$ and $\sigma_{NMAD}(\Delta z) = 0.287$. We obtain a low bias for Δz_{norm} at 11.6×10^{-4} and for Δz at 8.5×10^{-2}. We also have a low percentage of catastrophic outliers at 5% of our total sample. The so-called catastrophic outliers are the outliers in ML nomenclature (Jones & Singal 2020)). More specifically, these catastrophic outliers are the γ-ray AGNs for which $|\Delta z| > 2\sigma$, and
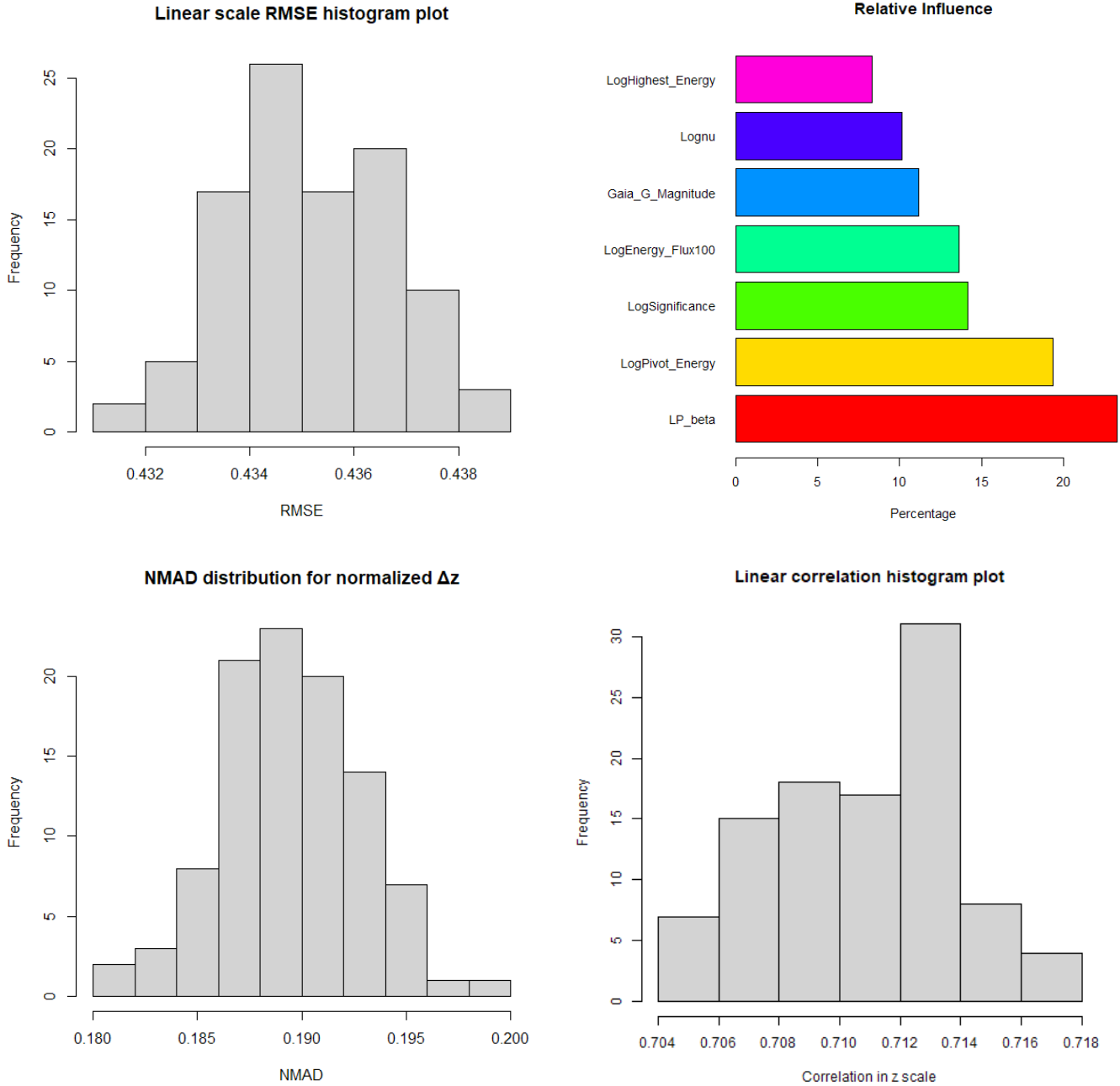
Figure 11. In all panels the results are obtained with the one hundred 10fCV. Top left and right panels show the histogram of the RMSE, and the relative influence of our chosen predictors, respectively. Bottom left and right panels show the NMAD distribution, and linear Correlation distribution, respectively.

thus lie outside the cone presented in Fig. 10. In the upper panel of Fig. 11, we present the distribution of our linear scale RMSE, and the relative influence of the features in our data over the one hundred 10f nested CV runs, in the upper left and right panels, respectively. In the bottom panel of Fig. 11, the NMAD and the differential distribution of the correlation coefficient are shown in the left and right panels, respectively. We note here that in our analysis the redshift of γ-ray AGNs is not just an effect of distance-brightness relation, which is due to selection biases (see Singal et al. (2013b), Singal et al. (2012), Singal et al. (2014), Singal (2015), Singal et al. (2013a), as we have discussed in the introduction). Indeed, a very recent study (Qu et al. (2019) and Zeng et al. (2021)) has been performed on the 4LAC catalog to evaluate the dependence of the BLLs luminosity on the redshift. For completeness, we also present the

Experiment	Bias (Δz_{norm})	Sigma (Δz_{norm})	NMAD (Δz_{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	...

Table 1. Comparison of our results with those of other ML-based photometric redshift estimation techniques. The empty spaces indicate a lack of available data for those cases.

results from a sample that is not used in the CV step at all, alongside with the prediction of the model on an internal test set in Fig. 13. With this validation set, we have a catastrophic outlier percentage of 7%, thus comparable with the previous values. In the left upper panel of Fig. 12, we show the histogram of Δz indicating with the red line indicating the bias and with the blue line the $\pm 1\,\sigma$; while in the right upper panel of Fig. 12 we present the histogram of Δz_{norm} with the red line the normalized bias, and with the blue line the $\pm 1\sigma$ normalized.

We present the residual plot in Fig. 14 bottom right panel. The lack of any increasing or decreasing trend of the redshift between the residuals and the fitted values is evidence of the goodness of our fit. Furthermore, the R^2 value for our result is 0.508, and the (Interquartile Range, IQR) value for $\Delta z = 0.39$. Additionally, we compare our results with other works done in the field, such as Richards et al. (2008) (Type-1 broad line quasars from SDSS), Laurino et al. (2011) (Optical galaxies and quasars from SDSS), Ball et al. (2008) (Main sample galaxies, luminous red galaxies and quasars from SDSS and GALEX), and Brescia et al. (2013) (Quasars from SDSS+GALEX+WISE+UKDISS). The comparisons are shown in Table 1.

We stress that even though our results do not always achieve a more precise prediction than some of the cases shown in Table 1, they are still comparable to them, and we need to take into account that our training set is at least twice smaller compared to the sample investigated in the mentioned paper. Hence, these results highlight that further enlargement and enhancements to the 4LAC dat will produce more precise results in the near future.
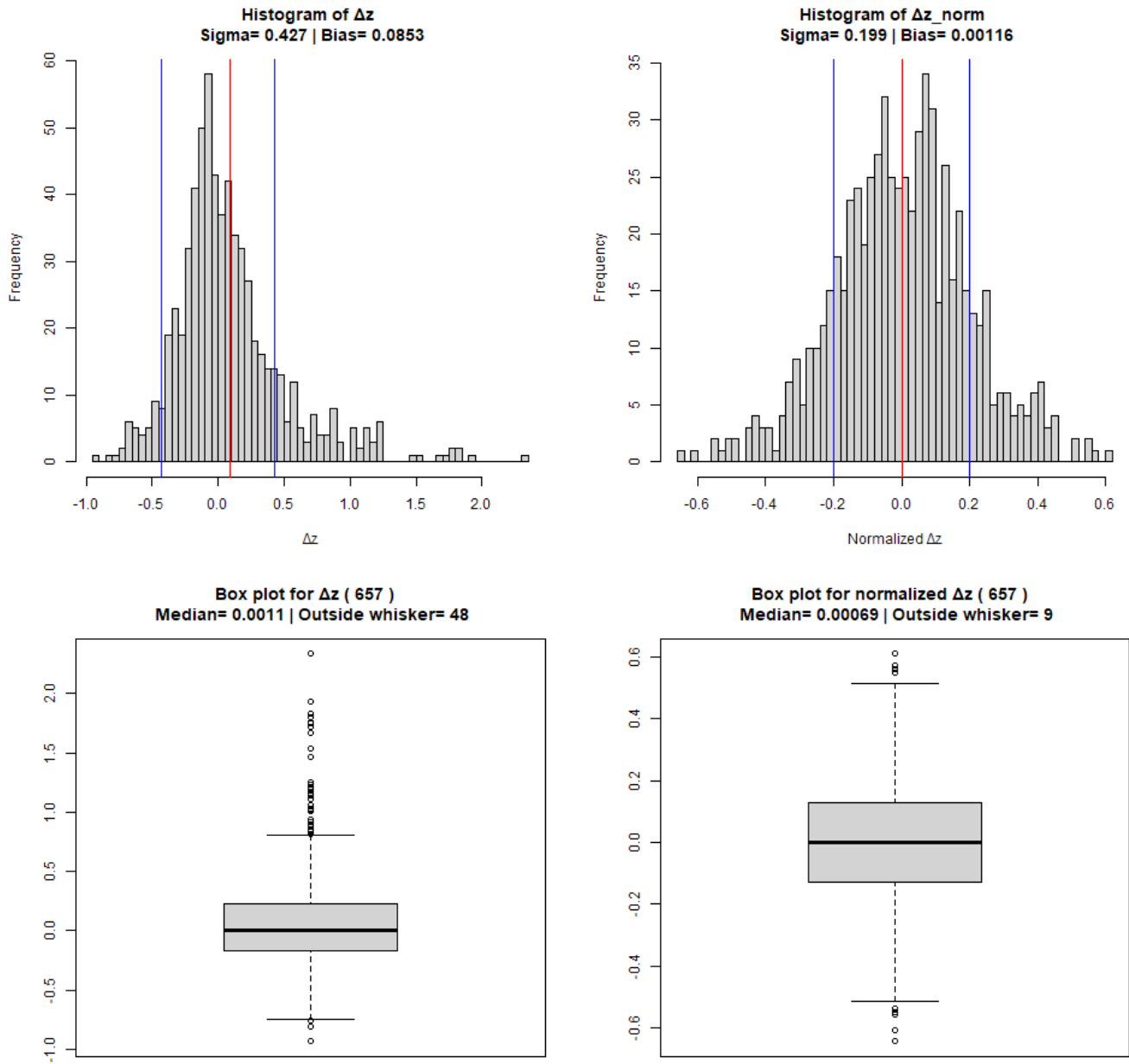
Figure 12. The differential distribution of the frequencies Δz and Δz_{norm} are shown in the left and right panels, respectively. The blue lines indicate the σ value and the red line the bias. The bottom plots show the box plot representation of the above frequency histogram, respectively.
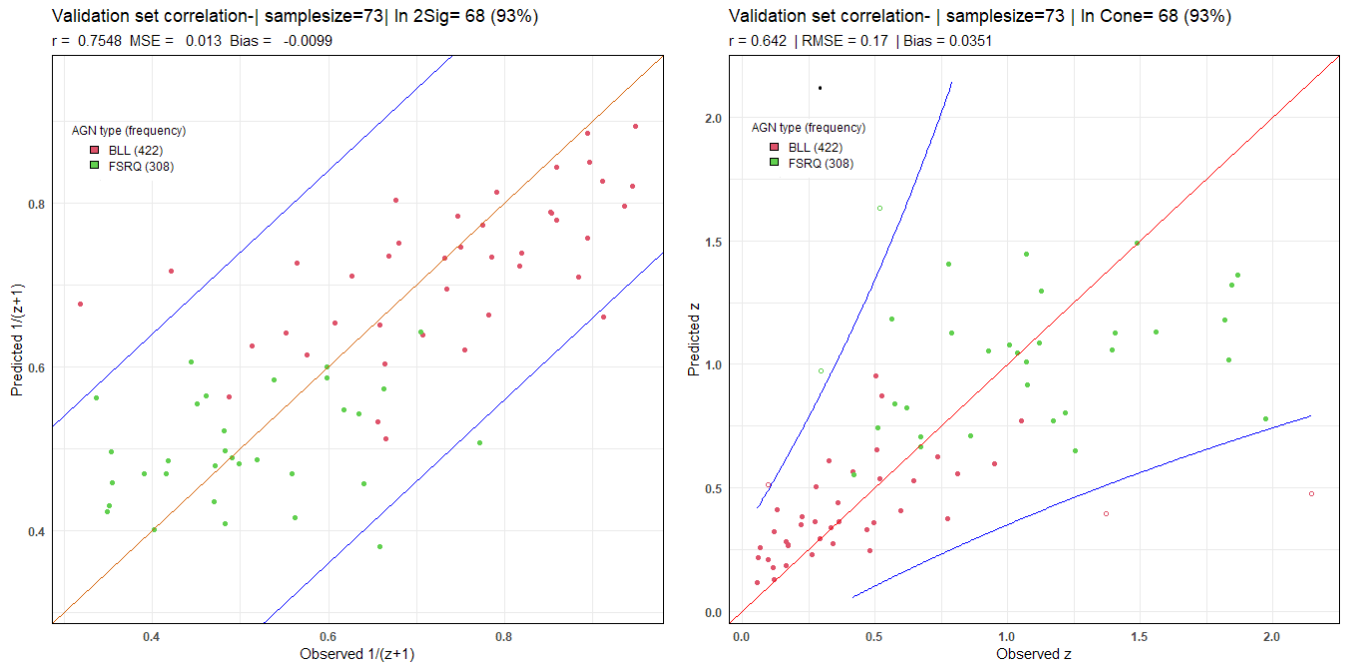
Figure 13. The Correlation of validation set predicted $\frac{1}{z+1}$ vs. the observed $\frac{1}{z+1}$ (upper left panel) and the predicted z vs. the observed one (upper right panel).
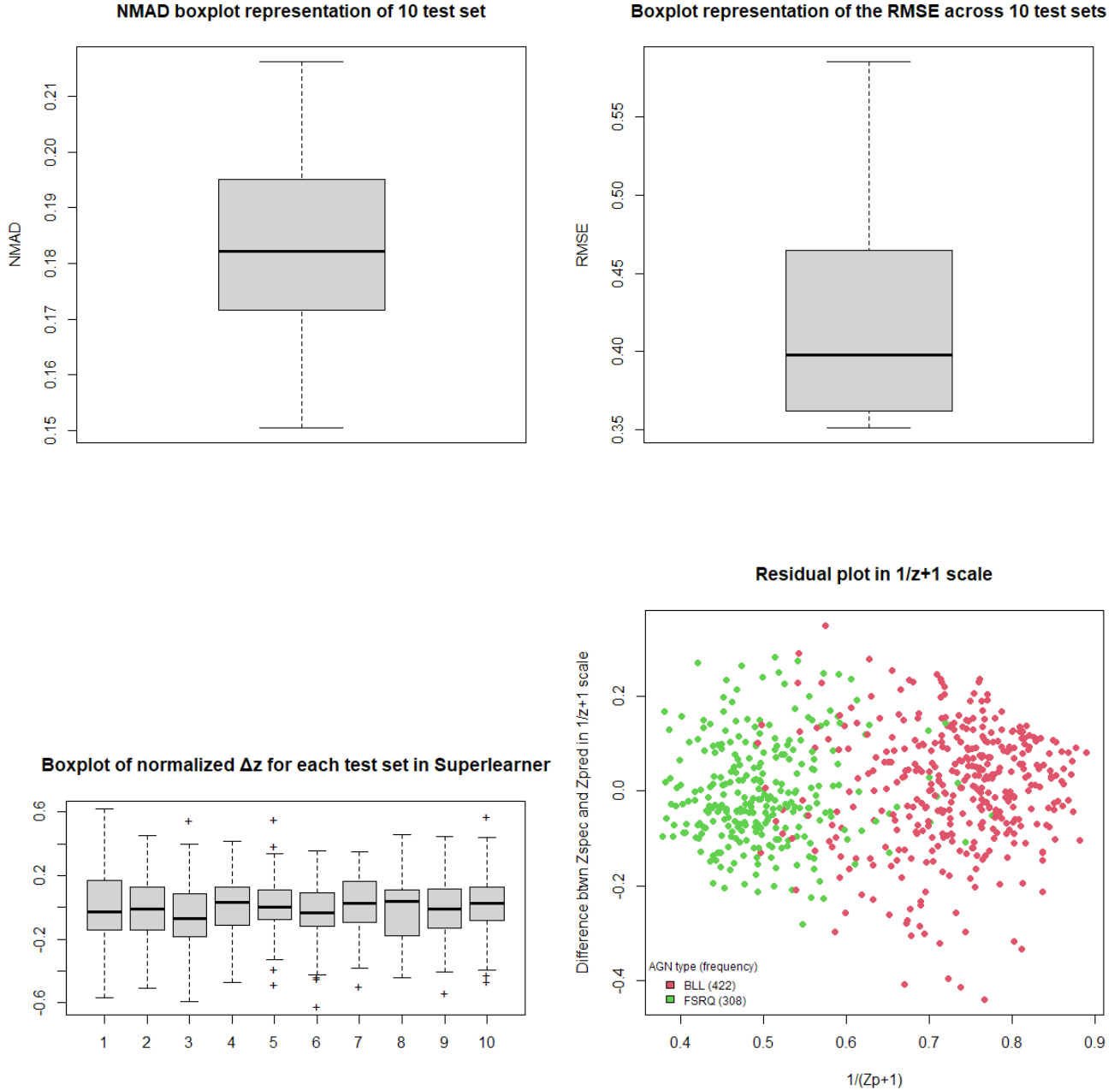
NMAD boxplot representation of 10 test set

Boxplot representation of the RMSE across 10 test sets

Boxplot of normalized Δz for each test set in Superlearner

Residual plot in 1/z+1 scale

AGN type (frequency)
- BLL (422)
- FSRQ (308)

Figure 14. The top left and right panels show the boxplot of the NMAD values for the internal Superlearner test set, and the boxplot of the RMSE values for ten internal Superlearner test sets, respectively. The bottom left panel shows the Δz distribution for the ten Superlearner test sets. Bottom right panel: The residuals VS the Superlearner predictions for each of the test sets.

4.1. *Bias correction*

As it can be seen from Fig. 10 left panel, the higher redshift AGNs are being predicted at a lower value. This is a clear signature of our predictions being biased. To correct for this, we fit a linear model between the observed and predicted redshifts in the $\frac{1}{z+1}$ scale. We fit linear models for both BLLs and FSRQs separately, which are shown by the cyan and purple dashed lines in Fig. 15, left panel. The black dotted line represents the linear fit for both BLLs and FSRQs together. We can see clearly that the fitted lines deviate from the 1:1 line.
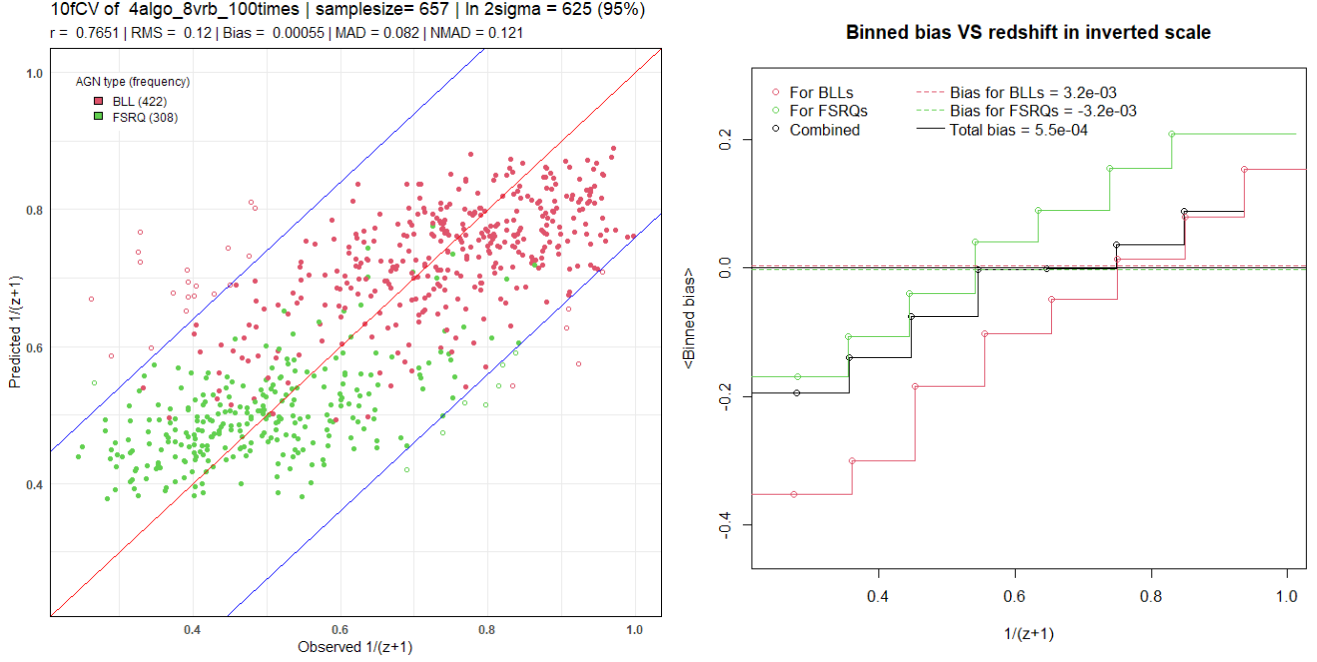
Figure 15. Left Panel: Linear regression fitting between the predicted and observed redshifts. The cyan and purple lines show the linear fit for BLL and FSRQs, respectively. Right Panel: Plot of binned $\frac{1}{z+1}$ vs mean bias. The average bias for BLLs and FSRQs is 3.2×10^{-3} and -3.2×10^{-3}, respectively.

The bias corrections for BLLs and FSRQs follow this equation:

$$U_{prediction} = a * U_{observed} + b, \tag{2}$$

where $U_{prediction} = \frac{1}{Z_{predictions}+1}$, $U_{observed} = \frac{1}{Z_{observed}+1}$, a and b are the slope and the intercept of linear fit, respectively. We obtain a different value of a and b for BLLs and FSRQs. These quantify the bias present in our analysis. For BLLs : $a = 0.29$ and $b = 0.51$. For FSRQs : $a = 0.29$ and $b = 0.35$.

4.2. Prediction on the generalization set

Our initial aim, as already indicated in the introduction, is to increase the number of 4LAC γ-ray AGNs that have estimates of the redshift. Based on the results shown in the previous section, we have reached so far a trained model which enables predictions for 4LAC γ-ray AGNs that fall within its trained parameter space. Indeed, for the generalization set, it is of crucial importance to ensure that the generalization set parameter space should overlap with our training set as much as possible. We start with a great advantage with this data set, since based on the scatter matrix plot in Fig. 6 we can observe that there is a significant overlap in the training (red and green data points for BLLs and FSRQs, respectively) and the generalization set (blue and black points for BLLs and FSRQ, respectively). Hence, the trained model has the advantage of extrapolating less when predicting the redshift of the generalization set. For the generalization set, we decide to retain γ-ray AGNs based on the condition that the values of their predictors should fall within the maximum and minimum values of the corresponding predictor in the training set. This way, we can achieve more reliable redshift predictions with minimal extrapolation.

To better evaluate how the generalization set overlaps with the training set, we present a scatter matrix plot in Fig. 7, showing the distribution of the very same seven predictors chosen by the LASSO features in Fig. 6. The blue points belong to the new trimmed generalization set, and as we can see, all the points fall well within the training set data points, as shown by the red points.

After we perform these cuts in the parameter space, we are left with 232 γ-ray AGN which is 97% of the total number. These 232 γ-ray AGNs are all BLLs. We would like to clarify here that the objects in the generalization sample that are classified as BCU, or uncategorized, are excluded when we are performing our predictions. We also exclude the single FSRQ that we have in our generalization set, so as to focus solely on BLLs for our predictions. Thus, the trimming of the variables does not influence the total number of redshifts we predict. We present the results of our

analysis in Fig. 16. As shown in our previous results (see Fig. 10), 95% of our predictions fall within the 2σ error bars. We expect a similar scenario for the predictions on the generalization set. Here, the blue histogram bars represent the median of the predictions on the generalization set, not taking into account the 2σ errors. We performed the Kolmogorov Smirnov Test (KS) test to evaluate if the extracted redshift distribution comes from the observed redshift distribution in the training set. As a result, we obtained that the null hypothesis that the two distributions come from the same parent population is rejected at the level of less than $10^{-16}\%$. Since we are not taking into account the error bars, hence the KS test gives us that the two distributions are different. Thus, we decided to investigate this issue by performing the KS test again on the singular distribution of the variables and we confirm also that the null hypothesis of similarity is rejected. Thus, it is not surprising that the two redshift distributions are not similar. Nevertheless, we do not necessarily expect the distributions of the redshift to be similar from a statistical point of view, since selection biases are at play and it is possible, as mentioned earlier, that we observe the faintest γ-ray AGNs at low redshift and the brightest γ-ray AGNs at higher redshift.

Our model without accounting for the bias correction predicts the redshift for BLLs between 0.5 and 1. With the application of the bias correction, the predicted redshifts are extended to cover the whole interval between 0 and 3, which better resembles the distribution of true redshifts. When the originally predicted redshift (Superlearner prediction) is close to 0.5, then we are at the borders of the generalization limits, namely close to the intercept values b and can not predict the true redshift well.
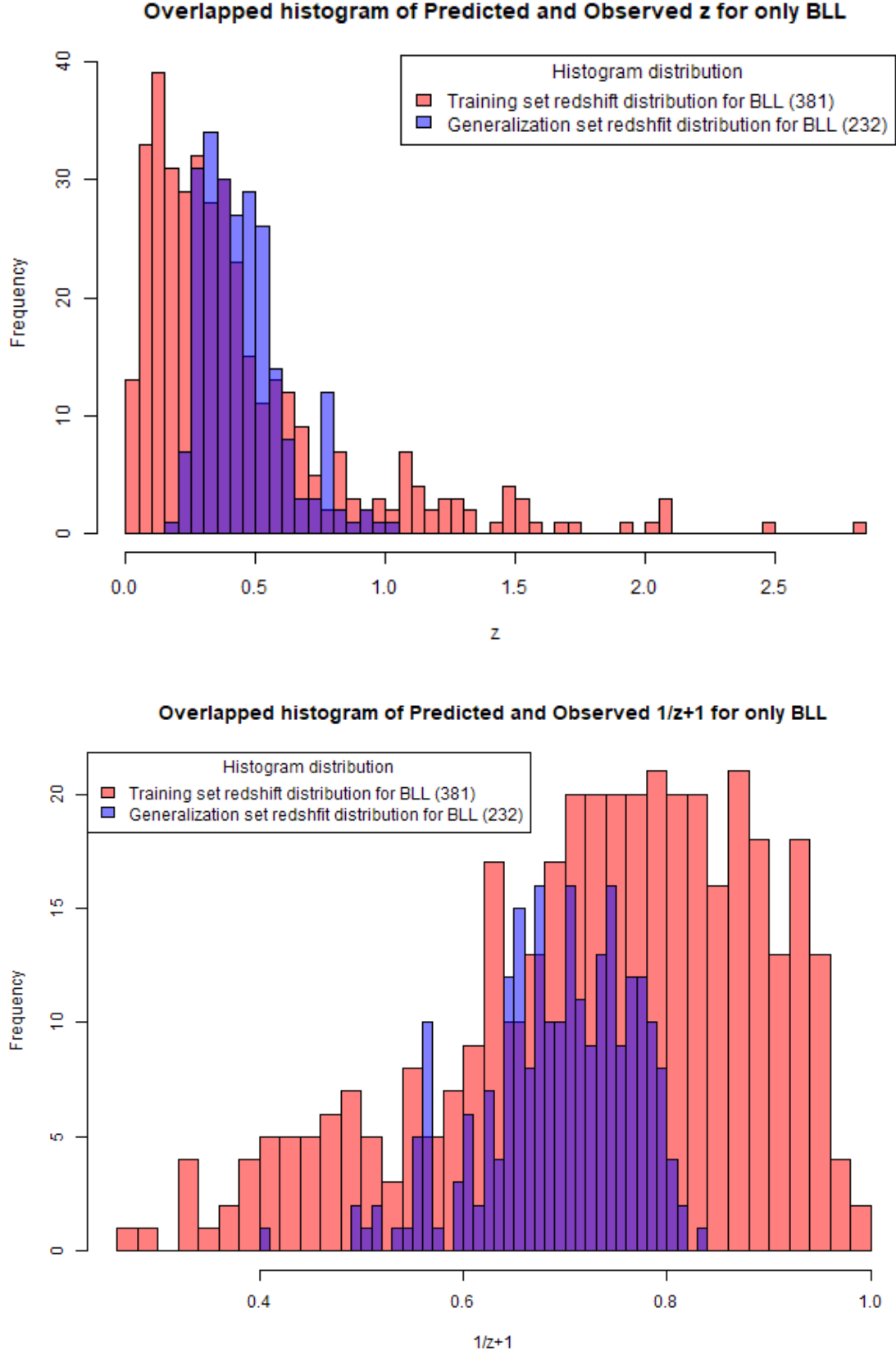
Figure 16. The differential distribution of the predicted redshift of 232 BLLs from the generalization set (blue histogram) vs. training set (orange data points). The upper panel shows the distribution in linear scale, while the bottom panels shows the distribution in the $\frac{1}{z+1}$ scale.

To be more specific, our sample contains FSRQs and BL Lacs in similar numbers (655 FSRQs and 686 BL Lacs). However, it is easier to measure redshift in FSRQs given their prominent broad emission lines. Given the observational difficulties in measuring redshifts for BL Lacs, the sources in our study might not be a representative sample of the BL Lac population. There is a non-zero probability chance for sources to be mis-classified, or even the γ-ray source to be mis-associated with a counterpart. Moreover, our sample contains only 60 non-blazar γ-ray AGN whose γ-ray properties potentially evolve differently with redshift. All of the above may hamper the accuracy of the ML models. However, given the improvement in localization accuracy, the number of sources, and the number of non-blazar γ-ray AGN (a factor of two improvement) between the 3LAC and 4LAC (as well as earlier catalogs), future *Fermi* catalogs will allow us to address further the shortcomings of our current sample.

5. **CONCLUSION**

In this work, we have crafted a methodology to predict the redshift of γ-ray loud AGN from the 4LAC catalog, using their observed γ-ray properties. We used categorical variables to distinguish among γ-ray AGN types and the LASSO algorithm to select the most predictive variables. We select the ML models based on the coefficient of the predictive power obtained with Superlearner after we have performed the optimization of the models. We trained several ML algorithms on these properties by using Superlearner and used the trained models to predict the γ-ray AGN redshifts. By computing the relative influence of these observed properties, we also determine which of them are the best predictors. The application of these methods to the 4LAC γ-ray AGN catalog for the BLLs sources for which the redshift is unknown increases 61% the size of the data set of γ-ray AGNs with known redshift, thus allowing to reach a larger sample. This new data set will have the great advantage to be complete for a given flux limit with a higher percentage. This enlarged sample of γ-ray AGNs, in turn, will allow us to determine the luminosity function, its evolution, and the density evolution of γ-ray AGNs with improved accuracy. With a sample of 657 γ-ray AGNs with measured redshifts, we have shown that using the Superlearner method can provide predicted redshifts that correlate with the observed redshift to a high degree of accuracy. We obtain, after performing one hundred 10f nested CV, an average Pearson Correlation coefficient, $r = 0.77$ in the $\frac{1}{z+1}$ scale and RMSE= 0.12 and a bias of 5.4×10^{-4}; if we consider the results instead in the z scale $r = 0.71$, the $\text{RMSE}(\Delta z_{norm}) = 0.43$, the bias$(\Delta z_{norm})$ 1.2×10^{-3} and $\sigma_{NMAD} = 0.192$.

We then predict the redshift of 232 BLLs that do not have the observed redshift and plot them against the observed redshift. Most γ-ray AGNs without the estimation of redshift lie between $0.18 \leq z \leq 1.02$.

Previous work utilizing ML algorithms focused primarily on the classification of γ-ray AGNs. Currently, to the best of our knowledge, no work in the **blazar** literature attempts to estimate the redshift using their observed γ-ray characteristics. This is a pioneering work in γ-ray AGN redshift estimation and will hopefully usher in follow-up studies that can improve our predictive capabilities even further.

6. APPENDIX

In this Appendix, we discuss how it is crucially important to show how the models used together with an ensemble performs better than the singular methods. In Table. 2 we show the RMSE, linear correlation, Bias, and NMAD scores of the individual algorithms used in the ensemble and the final Superlearner ensemble score. Based on the RMSE and the linear correlation values, we can clearly see that the Superlearner ensemble performs better. The singular model scores presented here are 10fCV and we ran them with the same optimization parameters shown in Sec. 3.3.

Algorithm	Root Mean Square error	Linear correlation	Bias (Δz_{norm}) ($\times 10^{-4}$)	NMAD Δz_{norm}
SuperLearner	0.014	0.71	11.6	0.19
XGB	0.015	0.70	22.6	0.19
RF	0.015	0.70	15	0.20
BigLasso	0.02	0.69	2.2	0.19
BayesGLM	0.02	0.69	8.6	0.19

Table 2. The 10fCV risk estimates of individual algorithms and the Superlearner ensemble.

Our choice of using $\frac{1}{z+1}$ scaling for the redshift instead of $\log(z+1)$ is based on the result presented in Table. 3. These results are obtained after performing a 10fCV using the two different scalings.

Scaling	Mean square error	Linear Correlation	Bias (Δz_{norm}) ($\times 10^{-4}$)	NMAD Δz_{norm}
log(z+1)	0.427	0.70	223	0.2
$\frac{1}{z+1}$	0.435	0.71	11.6	0.19

Table 3. The MSE, Correlation, bias, and NMAD of two different redshift scaling.

We show the one hundred 10fCV results related to the RMSE, the NMAD distribution for the normalized Δz, and the linear correlation. For completeness of the discussion, we show the results when we exclude LogSignificance from our analysis, see Fig. 17 .

Figure 17. The linear scale correlation plot, when LogSignificance is not included.

Next, we present the results when we use only a single variable, LogEnergyFlux, for the prediction using our ensemble, in Fig.18.

Figure 18. Correlation plot in linear scale. The values for statistical parameters are shown on the plots themselves.

It is clear that when we use only one predictor even though it has a high relative influence (the flux), the prediction we achieve for the redshift is poor compared to the prediction we obtain with the full set of LASSO selected predictors.

Additionally, we show the results obtained when using our two most predictive features, i.e LP_beta and LogPivotEnergy in Fig. 19.

Figure 19. The linear correlation plot when using LP_beta and LogPivotEnergy in our ensemble.

These two have the highest relative influence in our feature set, but, using them independently does not lead to accurate results as the entire feature set does.

REFERENCES

Abdollahi, S., Acero, F., Ackermann, M., et al. 2020, ApJS, 247, 33, doi: 10.3847/1538-4365/ab6bcb

Ackermann, M., Ajello, M., Albert, A., et al. 2015, The Astrophysical Journal Letters, 813, L41

Ajello, M., Angioni, R., Axelsson, M., et al. 2020, ApJ, 892, 105, doi: 10.3847/1538-4357/ab791e

Ball, N. M., Brunner, R. J., Myers, A. D., et al. 2008, The Astrophysical Journal, 683, 12–21, doi: 10.1086/589646

Birnbaum, A. 1962, Journal of the American Statistical Association, 57, 269

Breiman, L. 2001, Machine learning, 45, 5

Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., & Mercurio, A. 2013, The Astrophysical Journal, 772, 140

Brescia, M., Salvato, M., Cavuoti, S., et al. 2019, MNRAS, 489, 663, doi: 10.1093/mnras/stz2159

Carrasco, D., Barrientos, L. F., Pichara, K., et al. 2015, Astronomy & Astrophysics, 584, A44

Cavuoti, S., Brescia, M., D'Abrusco, R., Longo, G., & Paolillo, M. 2014, Monthly Notices of the Royal Astronomical Society, 437, 968

Chen, T., & Guestrin, C. 2016, in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794

Chiang, J., Fichtel, C., Von Montigny, C., Nolan, P., & Petrosian, V. 1995, The Astrophysical Journal, 452, 156

Chiaro, G., Salvetti, D., La Mura, G., et al. 2016, MNRAS, 462, 3180, doi: 10.1093/mnras/stw1830

Curran, S. 2020, Monthly Notices of the Royal Astronomical Society: Letters, 493, L70

Dainotti, M., Petrosian, V., Bogdan, M., et al. 2019, arXiv preprint arXiv:1907.05074

D'Isanto, A., & Polsterer, K. L. 2018, A&A, 609, A111, doi: 10.1051/0004-6361/201731326

Domínguez, A., Wojtak, R., Finke, J., et al. 2019, ApJ, 885, 137, doi: 10.3847/1538-4357/ab4a0e

Fermi-LAT Collaboration, Abdollahi, S., Ackermann, M., et al. 2018, Science, 362, 1031, doi: 10.1126/science.aat8123

Fotopoulou, S., & Paltani, S. 2018, Astronomy & Astrophysics, 619, A14

Friedman, J., Hastie, T., & Tibshirani, R. 2010a, Journal of Statistical Software, 33, 1. https://www.jstatsoft.org/v33/i01/

—. 2010b, Journal of statistical software, 33, 1

Friedman, J., Hastie, T., Tibshirani, R., et al. 2000, Annals of statistics, 28, 337

Friedman, J. H. 2001, Annals of statistics, 1189

—. 2002, Computational statistics & data analysis, 38, 367

Green, S. B., Ntampaka, M., Nagai, D., et al. 2019, The Astrophysical Journal, 884, 33

Hastie, T., & Tibshirani, R. 1987, Journal of the American Statistical Association, 82, 371

Hastie, T., Tibshirani, R., & Tibshirani, R. J. 2017, arXiv preprint arXiv:1707.08692

Hastie, T. J., & Tibshirani, R. J. 1990, Generalized additive models, Vol. 43 (CRC press)

Hildebrandt, H., Arnouts, S., Capak, P., et al. 2010, Astronomy & Astrophysics, 523, A31

Ilbert, O., Capak, P., Salvato, M., et al. 2008, The Astrophysical Journal, 690, 1236

Jones, E., & Singal, J. 2017, Astronomy & Astrophysics, 600, A113

—. 2020, Publications of the Astronomical Society of the Pacific, 132, 024501

Jordi, C., Gebran, M., Carrasco, J. M., et al. 2010, A&A, 523, A48, doi: 10.1051/0004-6361/201015441

Kang, S.-J., Fan, J.-H., Mao, W., et al. 2019, The Astrophysical Journal, 872, 189

Kaur, A., Rau, A., Ajello, M., et al. 2018, ApJ, 859, 80, doi: 10.3847/1538-4357/aabdec

—. 2017, ApJ, 834, 41, doi: 10.3847/1538-4357/834/1/41

Krakowski, T., Małek, K., Bilicki, M., et al. 2016, Astronomy & Astrophysics, 596, A39

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

Liodakis, I., & Blinov, D. 2019, MNRAS, 486, 3415, doi: 10.1093/mnras/stz1008

Logan, C., & Fotopoulou, S. 2020, Astronomy & Astrophysics, 633, A154

Marcotulli, L., Ajello, M., & Di Mauro, M. 2020, in American Astronomical Society Meeting Abstracts, Vol. 235, American Astronomical Society Meeting Abstracts #235, 405.06

Miller, A., Bloom, J., Richards, J., et al. 2015, The Astrophysical Journal, 798, 122

Nakoneczny, S., Bilicki, M., Solarz, A., et al. 2019, Astronomy & Astrophysics, 624, A13

Nakoneczny, S. J., Bilicki, M., Pollo, A., et al. 2020, Photometric selection and redshifts for quasars in the Kilo-Degree Survey Data Release 4. https://arxiv.org/abs/2010.13857

Pasquet-Itam, J., & Pasquet, J. 2018, Astronomy & Astrophysics, 611, A97

Polley, E. C., & Van der Laan, M. J. 2010

Qu, Y., Zeng, H., & Yan, D. 2019, Monthly Notices of the Royal Astronomical Society, 490, 758

Rajagopal, M., Kaur, A., Ajello, M., et al. 2020, ApJ, 898, 18, doi: 10.3847/1538-4357/ab96c4

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

Singal, J. 2015, Monthly Notices of the Royal Astronomical Society, 454, 115

Singal, J., Ko, A., & Petrosian, V. 2013a, Proceedings of the International Astronomical Union, 9, 149

—. 2014, The Astrophysical Journal, 786, 109

Singal, J., Petrosian, V., & Ajello, M. 2012, The Astrophysical Journal, 753, 45

Singal, J., Petrosian, V., & Ko, A. 2013b, AAS/High Energy Astrophysics Division# 13, 300

Tibshirani, R. 1996, Journal of the Royal Statistical Society Series B, 58, 267

Tibshirani, R., Bien, J., Friedman, J., et al. 2012, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74, 245

Valencia, D., Paracha, E., & Jackson, A. P. 2019, The Astrophysical Journal, 882, 35

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. 2007, Statistical applications in genetics and molecular biology, 6

Venters, T. M., & Pavlidou, V. 2013, MNRAS, 432, 3485, doi: 10.1093/mnras/stt697

Yang, Q., Wu, X.-B., Fan, X., et al. 2017, The Astronomical Journal, 154, 269

Zeng, H., Petrosian, V., & Yi, T. 2021, The Astrophysical Journal, 913, 120

Zeng, Y., & Breheny, P. 2017, arXiv preprint arXiv:1701.05936

Zhang, K., Schlegel, D. J., Andrews, B. H., et al. 2019, The Astrophysical Journal, 883, 63