Predicting the Redshift of $\gamma$-Ray-loud AGNs 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$^{13,14}$, and Johan Larsson$^4$

$^1$ National Astronomical Observatory of, Mitaka Japan; maria.dainotti@nao.ac.jp
$^2$ Space Science Institute, 4750 Walnut Street, Suite 205, Boulder, CO 80301, USA
$^3$ Department of Mathematics, University of Wroclaw, Poland
$^4$ Institute of Astronomy, University of Cambridge, Cambridge, UK
$^5$ Aalto University, Helsinki, Finland
$^6$ Department of Statistics, Lund University, Sweden
$^7$ Jagiellonian University, Poland
$^8$ Carnegie Mellon University, USA
$^9$ Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland
$^{10}$ Finnish Center for Astronomy with ESO (FINCA), University of Turku, Finland
$^{11}$ University of California at San Diego, La Jolla, CA, USA
$^{12}$ AGH University of Science and Technology, Krakow, Poland
$^{13}$ Faculty of Astronomy, University of California, Los Angeles, CA, USA

Received 2021 January 6; revised 2021 July 15; accepted 2021 July 15; published 2021 October 21

Abstract

Active galactic nuclei (AGNs) are very powerful galaxies characterized by extremely bright emissions coming 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 and their formation, along with the structure of early galaxies. The redshift determination is challenging because it requires detailed follow-up of multiwavelength observations, often involving various astronomical facilities. Here we employ machine-learning algorithms to estimate redshifts from the observed $\gamma$-ray properties and photometric data of $\gamma$-ray-loud AGNs from the Fourth Fermi-LAT Catalog. The prediction is obtained with the Superlearner algorithm using a LASSO-selected set of predictors. We obtain a tight correlation, with a Pearson correlation coefficient of 71.3% between the inferred and observed redshifts and an average $\Delta z_{\text{norm}} = 11.6 \times 10^{-4}$. We stress that, notwithstanding the small sample of $\gamma$-ray-loud AGNs, we obtain a reliable predictive model using Superlearner, which is an ensemble of several machine-learning models.

Unified Astronomy Thesaurus concepts: Active galactic nuclei (16)

1. Introduction

Active galactic nuclei (AGNs) with jets are the dominant class of objects when it comes to high-latitude ($|b| > 10$) extragalactic $\gamma$-ray sources (Abdollahi et al. 2020). The Fermi Gamma-ray Space Telescope has detected more than 2863 such $\gamma$-ray AGNs, the majority of which (>98%) are blazars: AGNs with their jets pointed toward 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 observationally expensive endeavor. The situation is further complicated by the absence of emission lines in the most numerous class of $\gamma$-ray-loud blazars, i.e., BLLs. 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 in the range $z = [0, 3]$, with most concentrated below $z = 2$. The $\gamma$-ray-loud blazars with redshift estimates are relevant for our comprehension of the origin of the extragalactic background light (EBL), which in turn lets us probe the cosmic evolution of blazars (e.g., Chiang et al. 1995; Singal et al. 2012, 2014; Singal et al. 2013a; Singal et al. 2013b; Ackermann et al. 2015; Singal 2015; Marcotulli et al. 2020), the intergalactic magnetic field (e.g., Venters & Pavlidou 2013), and the 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 BLLs 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 AGNs are becoming increasingly important in the era of big data astronomy (e.g., Ilbert et al. 2008; Hildebrandt et al. 2010; Brescia et al. 2013, 2019; D’Isanto & Polsterer 2018). Here we focus on the $\gamma$-ray-emitting AGN population in the 4LAC.

In the current literature, multiple works exist that focus on extracting a reliable photometric redshift for AGNs (Cavuoti et al. 2014; Jones & Singal 2017; Yang et al. 2017; Fotopoulou & Paltani 2018; Zhang et al. 2019; Curran 2020; Pasquet-Itam & Pasquet 2018; Logan & Fotopoulou 2020; Nakoneczny et al. 2019). In the current blazar literature, a lot of effort has also gone into classifying blazars of uncertain type (e.g., Chiari 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 $\gamma$-ray-loud blazars. Thus, we will tackle this problem by using ML and statistical-
learning algorithms. We apply multiple ML algorithms, such as the Least Absolute Shrinkage and Selection Operator (LASSO), Extreme Gradient boosting (XGBoost), Random Forest, and the Bayesian generalized linear model (Bayes GLM). 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 increase 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 (Ajello et al. 2020). It contains 2863 sources, of which 658 are FSRQs, 1067 are BLLs, 1074 are blazars of uncertain type, and 64 are classified as radio galaxies, narrow-line Seyferts (NLSY1), and other nonblazar AGNs. Out of the 2863 sources, 1591 have a measured redshift, whose distribution is shown in Figure 1. For completeness of the treatment, we have also included non-BL Lac and non-FSRQ sources in the initial scatter matrix plot in Figure 3 to show how the variables in the sample are distributed. But, in the generalization set, we are predicting the redshift for only the BLLs.

Unfortunately, not all of the 1591 γ-ray AGNs can be used for our model’s training. A significant number of these γ-ray AGNs have incomplete observational data, meaning that 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 redshifts, 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 AGNs, and 22 other-category AGNs. The 22 other-category γ-ray AGNs in our training set consist of two NLSY1 sources, three compact steep-spectrum radio sources, 13 radio galaxy sources, and two sources classified as nonblazar AGNs. They are shown in Figure 3. After we perform the cuts related to the missing data, we are left with 730 γ-ray AGNs. Thus, we perform our predictions only for BLLs and remove the 136 uncategorized 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 Figure 6, we show in black the only FSRQ from the generalization set.

The BLLs and FSRQs can be very easily separated, as we did when we introduced categorical variables into the Superlearner method. We stress here that this is an important point, because it means that the quality of the predictions will most probably differ, especially if the fractions of BLLs in the training sample and 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 BLLs and FSRQs.

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 several orders of magnitude, and we predict the redshift in the scale of \( \frac{1}{z+1} \) (see Figure 2). Out of these 13 variables, we take 11 into consideration. We exclude fractional variability due to the incompleteness of the AGN sample and log(\( \Delta \nu \)), as it is a second-order variable depending on log(\( \nu \)). The definitions and explanations of the 11 variables are given below.

1. Log\( \text{Flux} \)—Logarithm in the base of 10 of the integral photon flux, in photons cm\(^{-2}\) s\(^{-1}\), from 1 to 100 GeV.
2. Log\( \text{Energy Flux} \)—Logarithm in the 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.
3. Log\( \text{Significance} \)—The source detection significance in Gaussian \( \sigma \) units in the range from 50 MeV to 1 TeV.
4. Log\( \text{Variability Index} \)—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.
5. Log\( \text{Highest Energy} \)—Measured in GeV, 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%.
6. Log\( \nu \)—Logarithm in the base of 10 of the synchrotron peak frequency in the observer frame, measured in Hz.
7. PL\_Index—The photon index when fitting the spectrum with a power law, in the energy range from 50 MeV to 1 TeV.
8. Log\( \nu \text{Flux} \)—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 to 1 TeV.
9. LP\_Index—Photon index at pivot energy (\( \alpha \)) when fitting the spectrum (100 MeV to 1 TeV) with Log Parabola.
10. LP\( \beta \)—The spectral parameter (\( \beta \)) when fitting with the Log Parabola spectrum from 50 MeV to 1 TeV.
11. Gaia\_G\_Magnitude—Gaia magnitude at the \( g \) band provided by the 4LAC, taken from the Gaia Survey.

### 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, and 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 Section 3.4).

The statistical parameters used in order to compare our results with those of others in the field are bias, \( \sigma_{\text{NMAD}} \) (normalized median absolute deviation), Pearson correlation \( r \), root mean square error (RMSE), and standard deviation (\( \sigma \)). We quote the measured values of these parameters for \( \Delta z_{\text{norm}} \) and \( \Delta z \), where \( \Delta z_{\text{norm}} = \frac{z_{\text{pred}} - z_{\text{spec}}}{\sigma_{\text{NMAD}}} \) and \( \Delta z = z_{\text{spec}} - z_{\text{pred}} \).

As shown in the scatter matrix of Figure 3, we can see the presence of multiple correlated variables, such as PL\_Index and LP\_Index, Log\( \text{Energy Flux} \) and Log\( \text{Flux} \), and Log\( \text{Flux} \) and Log\( \text{Significance} \). 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 mainly consists of two parts, as presented in the flowchart in Figure 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 training and test sets composed of 657 \( \gamma \)-ray AGNs and a validation set composed of 73 \( \gamma \)-ray AGNs. We divide the sample, taking as the validation set the latest 10% of the \( \gamma \)-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 Section 3.4, we also apply the tenfold cross-validation (hereafter called 10CV) 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 given in Section 3.3. After training this ensemble on our data, we obtain our trained model, which leads us to the prediction on the redshifts.

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 $\ell^1$ 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 ($\lambda$). This penalization allows the model to select a subset of features and discards the rest by setting their coefficients to zero (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 observed values.

Figure 3. Full scatter matrix plot of all variables defined above, before feature selection. Here InvRedshift denotes $\frac{1}{1+z}$ scaled data.
Figure 4. Methodology flowchart. The rectangular boxes represent data sets, the parallelograms represent the γ-ray AGN categories, the rhombus indicates functions performed, the rounded rectangles indicate the ML algorithms used, the green lines show the direction of the input, the orange lines show the output, and the 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 data without spectroscopic $z$, green the results, and blue the intermediate steps or data sets.
redshifts, which is minimized during the 100 times 10fCV training. As a measure of the prediction errors, we quote the RMSE value, as well as the $\sigma_{\text{NMAD}}$. 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 $\lambda$ value for which the error is within 1 standard deviation (Friedman et al. 2010) and its corresponding coefficients for the features. The coefficients assigned by LASSO to each of them are displayed in Figure 5, and we choose only the nonzero coefficient features. To better visualize the parameter space of these features, we plot them in the scatter matrix plot shown in Figure 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. LogSignificance is providing the information on the detectability of the $\gamma$-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 Figure 17) that the results do not change at the level of 1% for $\sigma_{\text{NMAD}}$, RMSE, and correlation when we consider manually discarding this variable.

In addition, we clarify that we performed the analysis with both $\log_{10}(1+z)$ and $\frac{z}{2+z}$, with the distribution of the latter shown in Figure 2. The choice of transformation arises from the fact that the results related to the choice of $\frac{z}{2+z}$ present the smallest $\sigma_{\text{NMAD}}$ and smaller $\Delta z_{\text{norm}}$ (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 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.

1. 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.

2. The Random Forest algorithm generates multiple independent regression trees and averages them to obtain a more accurate prediction (Breiman 2001; Miller et al. 2015; Green et al. 2019; Valencia et al. 2019). An extremely difficult task is how to choose the optimal depth of such a tree, namely, to decide what the number of partition levels is. 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 rather built on each other using residuals from the previous step until the culmination of trees forms a stronger regression model.

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

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

5. 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, MAP, estimator). It works on the Fisher principle: “what value of the unknown parameter is most likely to generate the observed data.” The Bayes GLM method is more numerically and computationally stable 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. 2010).

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 overfitting). We tune these to best fit our data without over- or underfitting. In Figure 8, the top left and right panels show the variation of
the RMSE and correlation, respectively, related to the number of trees in the model. The RMSE and correlation minimize and maximize, respectively, at a depth of 5 and at a number of 500 trees. However, since depths 4 and 5 give very similar results, to avoid the risk of overfitting usually associated with a higher depth, we choose a max depth of 4 and proceed to test the model performance while varying the learning rate; see bottom panels of Figure 8. The optimal learning rate in our case is 0.01. In the bottom left panel of Figure 8, we plot the RMSE variation, and in the right panel, we plot 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, the 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 two. We vary the number of trees and the maximum number of nodes. The RMSE and correlation variations are shown in the top left and right plots of Figure 9, respectively. We observe that a value of 200 for maximum nodes gives the least RMSE and a 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 Figure 9.

---

**Figure 6.** The full symmetric scatter matrix plot shows the response (in our case, 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 parentheses indicate the number of γ-ray AGNs present in the data set.
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 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 also contemporaneously have the plateau of the correlation coefficient (see bottom left panel of Figure 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 Bayes GLM, there are no tunable hyperparameters, as is the case for XGBoost and Random Forest. Instead, we specify a formula based on which the redshift is predicted. The formula used is a linear combination of all of the features we consider:

\[
\frac{1}{z_i + 1} = f(\sum K_i).
\]

Figure 7. 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.

Here \( K \) belongs to a set of features described in Section 3.1 and presented in Figure 5, and \( i \) denotes each $\gamma$-ray AGN in the training set that is used in the model fitting.

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

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, it is called the 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, 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 zero. 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 data set is randomly partitioned into 10 complementary subsets. SuperLearner is trained on nine 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 data set. Thus, our prediction results are the average of 100 independent SuperLearner predictions. This allows for stabilization and derandomization of our results. Given the paucity of our data set, 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 Figure 4).

In Figure 10, the left panel shows the correlation plot between the observed and predicted redshift in $\frac{z + 1}{\sigma}$ (left panel) and linear scale (right panel). The blue lines indicate the 2σ
Figure 9. Random Forest optimization plots. The top 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 Random Forest 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 top panels but with a fixed value of maxnodes = 200 and with mtry = 2, 3, and 4 indicated with red, blue, and black, respectively.

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.
cones for each of the plots, where \( \sigma \) is calculated in the
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 \( 2\sigma \) line is not
straight on the linear scale and is shown with the following
formula:

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

We obtain a Pearson correlation \( r = 0.71 \) in the linear scale,
with \( \sigma_{\text{NMAD}}(\Delta z_{\text{norm}}) = 0.192 \) and \( \sigma_{\text{NMAD}}(\Delta z) = 0.287 \). We
obtain a low bias for \( \Delta z_{\text{norm}} \) at \( 11.6 \times 10^{-4} \) and \( \Delta z \) at
\( 8.5 \times 10^{-2} \). 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 \( \gamma \)-ray AGNs for which
\( |\Delta z| > 2\sigma \) and thus lie outside
the cone presented in Figure 10. In the top panels of Figure 11,
we present the distribution of our linear scale RMSE and the
relative influence of the features in our data over the 100 10f
nested CV runs in the left and right panels, respectively. In the
bottom panels of Figure 11, the NMAD and the differential

Figure 11. In all panels, the results are obtained with the 100 10fCV. The top left and right panels show the histogram of the RMSE and the relative influence of our
chosen predictors, respectively. The bottom left and right panels show the NMAD distribution and linear correlation distribution, respectively.

The Astrophysical Journal, 920:118 (20pp), 2021 October 20
Dainotti et al.
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 the distance–brightness relation, which is due to selection biases (see Singal et al. 2013b, 2012, 2014, 2013a; Singal 2015, as we have discussed in the Introduction). Indeed, a very recent study (Qu et al. 2019; Zeng et al. 2021) has been performed on the 4LAC to evaluate the dependence of the BLL luminosity on the redshift. For completeness, we also present the results from a sample that is not used in the CV step at all, alongside the prediction of the model on an internal test set in Figure 13. With this validation set, we have a catastrophic outlier percentage of 7%, thus comparable with the previous values. In the top left panel of Figure 12, we show the histogram of Δz, with the red line indicating the bias and the blue line ±1σ, while in the top right panel of Figure 12, we present the histogram of Δz_norm, with the red line indicating the normalized bias and the blue line ±1σ normalized.

We present the residual plot in Figure 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 value for Δz = 0.39. Additionally, we compare our results with other work in the field, such as Richards et al. 2008 (type 1 broad-line quasars from the Sloan Digital Sky Survey, 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 the Galaxy Evolution Explorer, GALEX), and Brescia et al. (2013; quasars from

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 indicates the bias. The bottom plots show the box plot representation of the above frequency histogram.
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 as small compared to the sample investigated in the mentioned paper. Hence, these results highlight that further enlargement and enhancements of the 4LAC data will produce more precise results in the near future.

### 4.1. Bias Correction

As it can be seen from Figure 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 is shown by the cyan and purple dashed lines in Figure 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.

The bias corrections for BLLs and FSRQs follow this equation:

$$U_{\text{prediction}} = a \times U_{\text{observed}} + b,$$

where $U_{\text{prediction}} = \frac{1}{z_{predicted}+1}$, $U_{\text{observed}} = \frac{1}{z_{observed}+1}$, and $a$ and $b$ are the slope and intercept of the 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 $\gamma$-ray AGNs that have estimates of the redshift. Based on the results shown in the previous section, we have so far reached a trained model that enables predictions for 4LAC $\gamma$-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 overlaps 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 Figure 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 FSRQs, 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 $\gamma$-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 Figure 7, showing the distribution of the very same seven predictors

![Figure 13. Correlation of the validation set predicted $\frac{1}{z+1}$ vs. observed $\frac{1}{z+1}$ (left panel) and predicted $z$ vs. observed $z$ (right panel).](image-url)

**Table 1**

| Experiment          | Bias ($\Delta_{\text{norm}}$) | $\sigma$ ($\Delta_{\text{norm}}$) | NMAD ($\Delta_{\text{norm}}$) |
|---------------------|-------------------------------|-----------------------------------|---------------------------------|
| 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                              | ...                             |

**Note.** The empty spaces indicate a lack of available data for those cases.
chosen by the LASSO features in Figure 6. The blue points belong to the new trimmed generalization set, and as we can see, all of 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 $\gamma$-ray AGNs, which is 97% of the total number. These 232 $\gamma$-ray AGNs are all BLLs. We would like to clarify here that the objects in the generalization sample that are classified as Blazar candidates of unknown type (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 Figure 16. As shown in our previous results (see Figure 10), 95% of our predictions fall within the $2\sigma$ 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 $2\sigma$ errors. We performed the Kolmogorov–Smirnov (K-S) test to evaluate if the extracted redshift distribution comes from the observed redshift distribution in the training set. As a result, we find that the null hypothesis that the two distributions come from the same parent population is rejected at a level of less than $10^{-16}\%$. Since we are not taking the error bars into account, the K-S test gives us that the two distributions are different. Thus, we decided to investigate this issue by performing the K-S test again on the singular distribution of the variables, and we also confirmed 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 \( \gamma \)-ray AGNs at low redshift and the brightest \( \gamma \)-ray AGNs at higher redshift.

Our model without accounting for the bias correction predicts a 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 zero and 3, which better resembles the distribution of true redshifts. When the originally predicted redshift (Superlearner prediction) is close to 0.5, we are at the borders of the generalization limits, namely, close to the intercept values \( b \), and cannot predict the true redshift well.

To be more specific, our sample contains FSRQs and BLLs in similar numbers (655 FSRQs and 686 BLLs). However, it is easier to measure redshift in FSRQs given their prominent broad emission lines. Given the observational difficulties in measuring redshifts for BLLs, the sources in our study might not be a representative sample of the BL Lac population. There is a nonzero probability chance for sources to be misclassified, or even for the \( \gamma \)-ray source to be misassociated with a counterpart. Moreover, our sample contains only 60 nonblazar \( \gamma \)-ray AGNs whose \( \gamma \)-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 nonblazar \( \gamma \)-ray AGNs (a factor of 2 improvement) between the 3LAC and 4LAC (as well as earlier catalogs), future Fermi catalogs will allow us to further address the shortcomings of our current sample.

5. Conclusion

In this work, we have crafted a methodology to predict the redshift of \( \gamma \)-ray-loud AGNs from the 4LAC using their observed \( \gamma \)-ray properties. We used categorical variables to distinguish among \( \gamma \)-ray AGN types and the LASSO algorithm to select the most predictive variables. We selected the ML models based on the coefficient of the predictive power obtained with Superlearner after we 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 \( \gamma \)-ray AGN redshifts. By computing the relative influence of these observed properties, we also determined which of them are the best predictors. The application of these methods to the 4LAC \( \gamma \)-ray AGN catalog for the BLL sources for which the redshift is unknown increases the size of the data set of \( \gamma \)-ray AGNs with known redshift by 61\%, thus allowing one to reach a larger sample. This new data set will have the great advantage of being complete for a given flux limit with a higher percentage. This enlarged sample of \( \gamma \)-ray AGNs, in turn, will allow us to determine the luminosity function, its evolution, and the density evolution of \( \gamma \)-ray AGNs with improved accuracy. With a sample of 657 \( \gamma \)-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 100 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 \( 5.4 \times 10^{-4} \); if we instead consider the results in the \( z \) scale \( r = 0.71 \), the RMSE (\( \Delta z_{\text{norm}} \)) = 0.43, the bias (\( \Delta z_{\text{norm}} \)) is \( 1.2 \times 10^{-3} \), and \( \sigma_{\text{NMAD}} = 0.192 \).

We then predict the redshift of 232 BLLs that do not have the observed redshift and plot them against the observed redshift. Most \( \gamma \)-ray AGNs without an estimation of redshift lie in the range \( 0.18 \leq z \leq 1.02 \).

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

This work presents results from the European Space Agency (ESA) space mission Gaia. Gaia data are being processed by the Gaia Data Processing and Analysis Consortium (DPAC). Funding for the DPAC is provided by national institutions, in particular the institutions participating in the Gaia MultiLateral Agreement (MLA). The Gaia mission website is https://www.cosmos.esa.int/gaia. The Gaia archive website is https://archives.esac.esa.int/gaia. M.G.D. thanks Trevor Hastie for the interesting discussion on overfitting problems. We also thank Raymond Wayne for the initial computation and discussions about balanced sampling techniques that will be implemented in subsequent papers. This research was supported by Polish National Science Centre grant UMO-2018/30/M/ST9/00757 and Polish Ministry of Science and Higher Education grant DIR/WK/

Figure 16. Differential distribution of the predicted redshift of 232 BLLs from the generalization set (blue histogram) vs. the training set (orange histogram). The top panel shows the distribution in linear scale, while the bottom panel shows the distribution in $\frac{1}{1+z}$ scale.
Finally, we would like to thank Sarthak Das and Subham Kedia for their help in performing the 10fCV analysis of the Superlearner algorithm.

Appendix

In this Appendix, we discuss how it is crucially important to show how the models used together with an ensemble perform 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 Section 3.3.

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.

We show the 100 10fCV results related to the RMSE, the NMAD distribution for the normalized $\Delta z$, and the linear correlation. For completeness of the discussion, we show the results when we exclude LogSignificance from our analysis; see Figure 17.

| Algorithm  | RMSE | Linear Correlation | Bias ($\Delta z_{\text{norm}}$) ($\times 10^{-4}$) | NMAD $\Delta z_{\text{norm}}$ |
|------------|------|---------------------|-----------------------------------------------|-------------------------------|
| 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                          |
| Big Lasso  | 0.02  | 0.69               | 2.2                                          | 0.19                          |
| Bayes GLM  | 0.02  | 0.69               | 8.6                                          | 0.19                          |

| Scaling   | Mean Square Error | Linear Correlation | Bias ($\Delta z_{\text{norm}}$) ($\times 10^{-4}$) | NMAD $\Delta z_{\text{norm}}$ |
|-----------|-------------------|---------------------|-----------------------------------------------|-------------------------------|
| log(z + 1)| 0.427             | 0.70               | 223                                          | 0.2                           |
| $\frac{1}{z+1}$ | 0.435            | 0.71               | 11.6                                         | 0.19                          |
Figure 17. Linear scale correlation plot when LogSignificance is not included.

Figure 18. Correlation plot in linear scale. The values for statistical parameters are shown on the plot itself.
Next, we present the results when we use only a single variable, LogEnergyFlux, for the prediction using our ensemble in Figure 18.

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 Figure 19.

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.

**ORCID iDs**

Maria Giovanna Dainotti @ https://orcid.org/0000-0003-4442-8546

Zooey Nguyen @ https://orcid.org/0000-0001-9219-3505

**References**

Abdollahi, S., Acero, F., Ackermann, M., et al. 2020, ApJS, 247, 33

Ackermann, M., Ajello, M., Albert, A., et al. 2015, ApJL, 813, L41

Ajello, M., Angioni, R., Axelsson, M., et al. 2020, ApJ, 892, 105

Ball, N. M., Brunner, R. J., Myers, A. D., et al. 2008, ApJ, 683, 12

Birnbaum, A. 1962, J. Amer. Stat. Assoc., 57, 269

Breiman, L. 2001, Mach. Learn., 45, 5

Brescia, M., Cavuoti, S., D’Abrusco, R., Longo, G., & Mercurio, A. 2013, ApJ, 772, 140

Brescia, M., Salvato, M., Cavuoti, S., et al. 2019, MNRAS, 489, 663

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

Cavuoti, S., Brescia, M., D’Abrusco, R., Longo, G., & Paolillo, M. 2014, MNRAS, 437, 968

Chen, T., & Guestrin, C. 2016, in KDD ’16: Proc. of the 22nd ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining (San Francisco, CA: ACM), 785

Chiang, J., Fichtel, C., Von Montigny, C., Nolan, P., & Peterson, V. 1995, ApJ, 452, 156

Chiari, G., Salvetti, D., La Mura, G., et al. 2016, MNRAS, 462, 3180

Curran, S. 2020, MNRAS, 493, L70

Dainotti, M., Petrov, V., Bogdán, M., et al. 2019, arXiv:1907.05074

D’Isanto, A., & Polsterer, K. L. 2018, A&A, 609, A111

Domínguez, A., Wojtak, R., Finke, J., et al. 2019, ApJ, 885, 137

Fermi-LAT Collaboration, Abdollahi, S., Ackermann, M., et al. 2018, Sci, 362, 1031

Fotopoulou, S., & Paltani, S. 2018, A&A, 619, A14

Friedman, J., Hastie, T., & Tibshirani, R. 2010, J. Stat. Softw., 33, 1

Friedman, J., Hastie, T., Tibshirani, R., et al. 2000, AnSta, 28, 337

Friedman, J. H. 2001, AnSta, 29, 1189

Friedman, J. H. 2002, Comput. Stat. Data. Anal., 38, 367

Green, S. B., Ntampaka, M., Nagai, D., et al. 2019, ApJ, 884, 33

Hastie, T., & Tibshirani, R. 1987, J. Amer. Stat. Assoc., 82, 371

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

Hastie, T. J., & Tibshirani, R. J. 1990, Generalized Additive Models, Vol. 43 (Boca Raton, FL: CRC press)

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

Ilbert, O., Capak, P., Salvato, M., et al. 2008, ApJ, 690, 1236

Jones, E., & Singal, J. 2017, A&A, 600, A113

Jones, E., & Singal, J. 2020, PASP, 132, 024501

Jordi, C., Gebran, M., Carrasco, J. M., et al. 2010, A&A, 523, A48

Kang, S.-J., Fan, J.-H., Mao, W., et al. 2019, ApJ, 872, 189

Kaur, A., Rau, A., Ajello, M., et al. 2018, ApJ, 859, 80

Kaur, A., Rau, A., Ajello, M., et al. 2017, ApJ, 834, 41

Kraakowski, T., Malek, K., Bilić, M., et al. 2016, A&A, 596, A39

Launino, O., D’Abrusco, R., Longo, G., & Riccio, G. 2011, MNRAS, 418, 2165

Liodakis, I., & Blinov, D. 2019, MNRAS, 486, 3415

Logan, C., & Fotopoulou, S. 2020, A&A, 633, A154

Figure 19. Linear correlation plot when using LP_beta and LogPivotEnergy in our ensemble.
Marcotulli, L., Ajello, M., & Di Mauro, M. 2020, BAAS, 52, 1
Miller, A., Bloom, J., Richards, J., et al. 2015, ApJ, 798, 122
Nakoneczny, S., Bilicki, M., Solarz, A., et al. 2019, A&A, 624, A13
Nakoneczny, S. J., Bilicki, M., Pollo, A., et al. 2021, A&A, 649, A81
Pasquet-Itam, J., & Pasquet, J. 2018, A&A, 611, A97
Polley, E. C., & Van der Laan, M. J. 2010, Super Learner In Prediction, Working Paper 266, UC Berkeley
Qu, Y., Zeng, H., & Yan, D. 2019, MNRAS, 490, 758
Rajagopal, M., Kaur, A., Ajello, M., et al. 2020, ApJ, 898, 18
Richards, G. T., Myers, A. D., Gray, A. G., et al. 2008, ApJS, 180, 67
Singal, J. 2015, MNRAS, 454, 115
Singal, J., Ko, A., & Petrosian, V. 2013a, in IAU Symp. 340, Multiwavelength AGN Surveys and Studies (Cambridge: Cambridge Univ. Press), 149
Singal, J., Ko, A., & Petrosian, V. 2014, ApJ, 786, 109
Singal, J., Petrovian, V., & Ajello, M. 2012, ApJ, 753, 45
Singal, J., Petrovian, V., & Ko, A. 2013b, American Astronomical Society HEAD Meeting, 13, 300
Tibshirani, R. 1996, J. R. Stat. Soc. Series B, 58, 267
Tibshirani, R., Bien, J., Friedman, J., et al. 2012, J. R. Stat. Soc. Series B, 74, 245
Valencia, D., Paracha, E., & Jackson, A. P. 2019, ApJ, 882, 35
Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. 2007, Statistical Applications in Genetics and Molecular Biology (Berlin: De Gruyter), 6
Venters, T. M., & Pavlidou, V. 2013, MNRAS, 432, 3485
Yang, Q., Wu, X.-B., Fan, X., et al. 2017, AJ, 154, 269
Zeng, H., Petrovian, V., & Yi, T. 2021, ApJ, 913, 120
Zeng, Y., & Breheny, P. 2017, arXiv:1701.05936
Zhang, K., Schlegel, D. J., Andrews, B. H., et al. 2019, ApJ, 883, 63