Draft version May 4, 2021
Typeset using LaTeX **preprint** style in AASTeX631

Measuring Photometric Redshifts of DESI Bright Galaxy Sample using various Machine Learning Algorithms*

Yasha Kaushal,[1] Ali Beheshti,[1] and Jeffrey Newman[1]

[1]*Department of Physics and Astronomy*
3941 O'Hara St, Pittsburgh, PA 15213
University of Pittsburgh, USA

ABSTRACT

The next generation cosmology experiments like Euclid, Large Synoptic Survey Telescope and Square Kilometre Array will rely on photometric redshifts rather than spectroscopic redshifts as obtaining spectroscopic data for billions of objects is both time and resource expensive. Hence, high accuracy and robust photometric redshift measurement is critical. In this study we test the performance of some machine learning algorithms in predicting the photometric redshifts of DESI Bright Galaxy Sample using different training sample sizes, different feature sets and different hyper-parameters. We found least outlier fractions for Keras Neural Networks and Random Forest Regression.

Keywords: Machine Learning, Regression, Photometric Redshifts, DESI Survey, Random forest Regression, K nearest neighbor, Gradient Boosting, XGBoost, CatBoost, Neural Networks, Multi-Layer Perceptron, Keras, Artificial Neural Networks, Gaussian Process Regression

1. INTRODUCTION

Photometric redshift is an estimate for the recession velocity of an astronomical object such as a galaxy or quasar using its easier to measure properties like color, magnitude, size, light profile, without measuring its spectrum. Most of the information in determining this is coming from photometry which is the brightness of an object viewed through various filters giving relatively broad pass band of colours such as red light, green light, or blue light. Once we have the redshift, using Hubble's Law we obtain the distance to the observed object. Historically, two main classes of methods have been used to obtain this, one is template fitting and other is machine learning. Template fitting methods use templates of galaxy spectral energy distributions (SEDs) for different galaxy types that can be redshifted to fit the photometry. It is highly dependent on whether the templates are representative of the observed galaxy sample and are corrected for how emission lines affect the photometric observations. Our work focuses on the latter method. We leverage on the existing plethora of photometric information of about 80,000 Bight Galaxies in the redshift range 0 <z <0.8 from the Dark Energy Spectroscopic Instrument Legacy Imaging Survey to train and tune Machine Learning Models with a sub sample with known spectroscopic redshift (true value) and photometric measurements and

* Released on April 24th, 2021

use them to make predictions on another sub-sample for which only the photometric information is known.

2. DATA

The Dark Energy Spectroscopic Instrument Bright Galaxy Survey (DESI-BGS) will be a flux-limited r-band selected sample of 10 million galaxies whose target selection will be done on the Legacy Surveys (LS) imaging. The DESI Legacy Imaging Surveys are a combination of three public projects (the Dark Energy Camera Legacy Survey, the Beijing-Arizona Sky Survey, and the Mayall zband Legacy Survey) that will jointly image about 14,000 deg^2 of the extragalactic sky visible from the northern hemisphere in three optical bands (g, r, and z). The survey catalog includes photometry from the g-r-z optical bands and from four mid-infrared bands (at 3.4µm, 4.6µm, 12µm and 22µm) observed by the Wide-field Infrared Survey Explorer (WISE) satellite. In this project we picked a subset of 80,952 objects with total 22 features - 20 numerical (g,r,z,w1,w2 magnitudes, g,r,z fibre magnitudes, delta chi square, extinction, sersic index, colors, shape) and 2 categorical (1. light profile morphology - radial, sersic, exponential, devoculers or point spread function and 2. sky survey region - north or south). We did some quality cuts and removed objects with any missing feature (nan,infinity values), spectral type 'STAR' and redshifts below 0 (mostly stars) and above 0.8 (mostly Quasars) from the analysis. After these selection cuts we had total 76609 objects.

3. METHODS AND METRICS

We deployed total 5 machine learning algorithms for our analysis.

1. **Random Forest Regression**
   Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. We used scikit-learn *RandomForestRegressor* package in ensemble module. We tested this algorithm with two weighting options - uniform and distance.

2. **K-nearest Neighbor Regression**
   KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood. We used scikit-learn *KNeighborsRegressor* package in neighbors module.

3. **Gradient Boosting**
   Gradient boosting is another class of ensemble machine learning algorithms that can be used for both classification or regression predictive modeling problems. Ensembles are constructed from decision tree models. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. This type of ensemble machine learning model referred to as boosting. Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. This gives the technique its name, "gradient boosting," as the loss gradient is minimized as the model is fit. We tested two types of gradient boosting algorithms :

(a) XGBoost Extreme Gradient Boosting is known for its computational efficiency and often better model performance.

(b) CatBoost The primary benefit of Category Gradient Boosting, in addition to computational speed improvements, is support for categorical input variables.

4. **Neural Networks**
This machine learning algorithm structure is inspired by the human brain, mimicking the way that biological neurons signal to one another. It is comprised of layers of nodes - an input layer, one or more hidden layers, and an output layer. Each node connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network. We deployed two Neural Networks in the analysis :

   (a) Multi-Layer Perceptron (MLP) Regression
   It is a class of feedforward artificial neural networks (ANN). Its multiple layers and non-linear activation distinguish it from a linear perceptron. It can distinguish data that is not linearly separable i.e. it can learn a non-linear function approximator. It utilizes supervised learning technique called backpropagation for training. We used a single hidden layer sequential model.

   (b) Keras based ANN
   Keras has a high level API compared to Pytorch that is built over Tensorflow and hence making it easier to use and implement. We used a 2 hidden layers sequential model with Rectified Linear Unit (Relu) activation function and l-2 kernel regularizer.

5. **Gaussian Process Regression**
It is a non-parametric, Bayesian approach to regression having the ability to provide uncertainty measurements on the predictions. Unlike many popular supervised machine learning algorithms that learn exact values for every parameter in a function, the Bayesian approach infers a probability distribution over all possible values. We tested this method with Gaussian likelihood, interpolated kernel (KISS) for calculating marginal likelihood and posterior mean and LOVE algorithm for posterior sampling and covariance matrix calculations.

We used 4 metrics for the analysis and in quantifying the performance and accuracy.

1. **Normalized Median Absolute Deviation (NMAD)** : It is a robust estimator of scatter as it is not sensitive to outliers. It is defined as

   $$\sigma_{NMAD} = 1.48 \times median\ \frac{|z_{phot} - z_{spec}|}{1+z_{spec}}$$

2. **Bias**

   $$\text{Bias} = median\ \frac{z_{phot} - z_{spec}}{1+z_{spec}}$$

3. **Outliers**

   $$\text{Outliers} = \frac{z_{phot} - z_{spec}}{1+z_{spec}} > 0.15$$

4. **Root Mean Squared Error (RMSE)**

$$\text{RMSE} = \sqrt{\frac{1}{n}\Sigma(\frac{z_{phot}-z_{spec}}{1+z_{spec}})^2}$$

4. RESULTS AND DISCUSSION

In diagnostic 1 for scaling relations with training sample size used to train the models, Keras based ANN gave the best performance with rapidly declining exponential relation requiring least training sample size to constrain the metrics NMAD and Outlier fractions together.

In diagnostic 2 of different features test, we found that Random Forest and CatBoost Regression gave best performance when all the features were included in the training data set. Both weighted and unweighted KNN algorithms performed best with only Color + Magnitude + Half Light Radius information in the training. XGBoost and Multi-Layer Perceptron Neural Network gave mixed performance with least NMAD when all the 22 features and only Color + Magnitude + Half Light Radius were included for both respectively and least outliers with Color + Magnitude + Categorical information in the features.

In diagnostic 3, we tested two different normalization of training sample before feeding it to the regression model, standard scalar and minmax scalar.Standard scalar standardizes the features by removing the mean and scaling to unit variance. Minmax scalar transforms features by scaling each feature to a given range. We used default range of (0,1). We did not find any significant deviations in the performance between the two scalars.

In diagnostic 4 with different manual hyper-parameter optimizations, we found both random forest regressor and Keras based ANN giving best performance with least outlier fractions (RF-0.672%, Keras-0.659%) and least NMAD (RF- 0.0255, Keras- 0.0267). Though it should be noted that there is scope for further optimizations in each method using optimization routines like Exhaustive Grid Search and Randomized Parameter Optimization that could perform extensive cross validation for multiple parameters at the same time.

Table 1. Outlier Fractions and NMAD metric for various ML Methods Used

ML method	Metrics	Col. & Mag.	Col., Mag. & HLR	Col., Mag. & Cat.	All features
Random Forest	NMAD	0.0283	0.0271	0.0274	0.0250
	Outliers	1.165 %	1.133 %	1.017 %	0.863 %
Weighted KNN	NMAD	0.0286	0.0281	0.0290	0.0320
	Outliers	1.101 %	0.991 %	1.191 %	1.088 %
XGBoost	NMAD	0.0359	0.0353	0.0353	0.0328
	Outliers	1.449 %	1.223 %	1.178 %	1.204 %
CatBoost	NMAD	0.0299	0.0290	0.0273	0.0248
	Outliers	1.281 %	1.185 %	1.114 %	0.927 %
MLP	NMAD	0.0329	0.0300	0.0319	0.0308
	Outliers	1.185 %	1.140 %	0.933 %	1.062 %

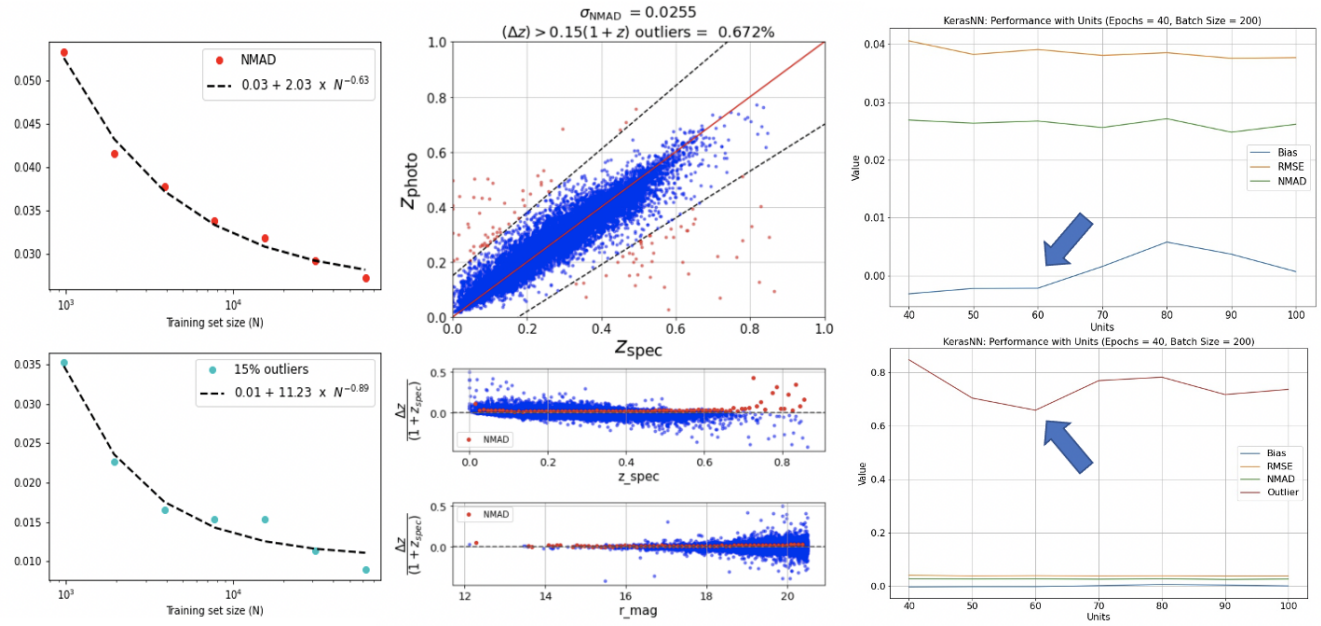

Figure 1. Example scaling relations, test-set predictions, error trends with true redshift and r-band magnitude and trends of Bias, RMSE, NMAD with hyperparamters

5. CONCLUSIONS AND FUTURE WORK

Though both random forest regressor and Keras based ANN gave best performance with least outlier fractions and NMAD metric and GPR gave worst performance with most number of outliers (∼1%) and highest NMAD, it should be noted that there is scope for further optimizations in each method using optimization routines like Exhaustive Grid Search that could perform extensive cross validation for multiple parameters at the same time and the power of obtaining confidence intervals from GPR should not be underestimated. In our future work, we would like to first optimize the hyper-parameters in our model using scikit learn routines and then re-perform the scaling and feature set analysis on the best hyper-parameter optimized models to see which method is giving us most accurate results with least computation time.

6. REFERENCES

1. DESI Imaging Team, Overview of the DESI Legacy Imaging Surveys, Arjun Dey et al. + 2019
2. DESI BGS Team, Preliminary Target Selection for the DESI Bright Galaxy Survey (BGS), Omar-Ruiz-Macias et al. + 2020
3. Rongpu Zhou + 2021, The Clustering of DESI-like Luminous Red Galaxies Using Photometric Redshifts
4. Brescia M. + 2014, A catalogue of photometric redshifts for the SDSS-DR9 galaxies
5. Carrasco Kind M., Brunner R. J + 2013, TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests
6. Almosallam et al. + 2016, GPz: Non-stationary sparse Gaussian processes for heteroscedastic uncertainty estimation in photometric redshifts