Recovery of DESI BGS redshift measurements using Machine Learning

Sergio Daivd Lobo Bolaño

February 8, 2019

Abstract

DESI (Dark Energy Spectroscopic Instrument) uses a simulated survey of mock galaxies to compare their redshifts with its -also simulated- redshift measurement. In this monograph we apply machine learning (ML) methods to the simulation data of DESI to recover the true redshift measurements of the Bright Galaxy Sample using observational variables as input, obtaining a $r^2 = 0.98$ using kernel and ensemble methods.

1 Introduction

The redshift measurements of DESI present differences with respect to the true redshifts values of the mock galaxies used in simulations. It is necessary to correct these measurements so that the instrument can work properly when tested in the real world. One possible way is to use machine learning to recover the true redshift from the mock galaxies from observational variables as input. Therefore, this monograph has as main objective to recover the true redshifts of Bright Galaxies using observational variables from the simulated DESI measurements as an input to machine learning models, to improve the accuracy of DESI when used in the real world

To accomplish this, first we pre-process the data and select the variables as input for the ML models, these variables are the fluxes from the different observation bands, which are the $g, r, z, WISE1$, and $WISE2$ bands, as well as the redshift output of the instrument's pipeline. The true redshift from the mock galaxies is used as output. We train and test support vector regression (SVR) and kernel ridge regression (KRR) models using cross-validation and grid search, later we use an ensemble model of different KRRs.

2 Methodology

First, we show a description of the data pre-processing, then the selection of computer parameters and finally, a description of the methodology for training and testing the models. Figure 1 shows the general methodology for the project, each step will be treated below.

The total dataset of size 889.336 was divided into training-test subsets of relative size 75 - 25. The training subset was used for training and selecting hyper-parameters by using grid search and cross validation. Figure 2 shows the data subgrouping.

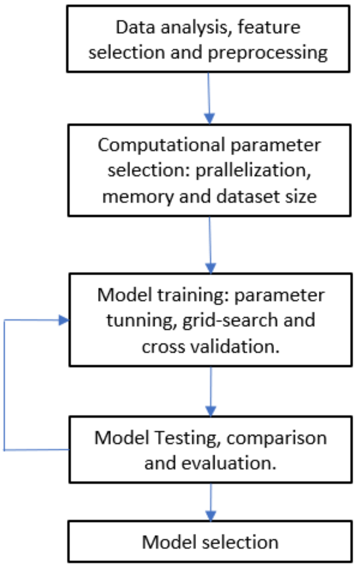

Data analysis, feature selection and preprocessing
Computational parameter selection: prallelization, memory and dataset size
Model training: parameter tunning, grid-search and cross validation.
Model Testing, comparison and evaluation.
Model selection

Figure 1: General methodology for the project.

During training, the development-evaluation stage was made on a even smaller subset inside the training set of size 100.000, due to computational resources. Therefore, to avoid sub-training the models we trained them in three different ”sub-training” sets of size 100.000 each, and after that, we created an ensemble model with the combined responses of each one. In the Results section we explain this further.

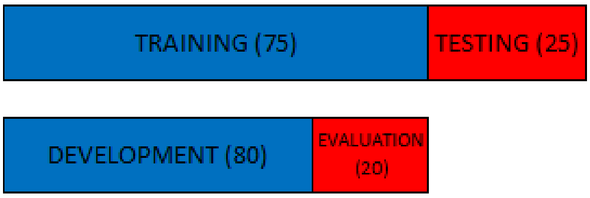


Figure 2: Division of the dataset in train and test. The train set is divided in development-evaluation subsets.

3 Data Overview

The true redshift (from the mock galaxies) and the target redshift (as measured by DESI) of BGS galaxies should in principle be equal, however, due to propagating error and misclassification in the data pipeline, the predicted redshift of the objects may have a deviation from its true value. Figure 3 shows relation between the true redshift (TRUEZ) and the target redshift (TARZ).

In order to plot the relation between TRUEZ, TARZ and other variables, we defined a new variable α as

$$\alpha = \frac{TRUEZ}{TARZ}. \quad (1)$$

Therefore, α will take values of 1 when $TRUEZ = TARZ$ and other if their are different, the larger (or smaller) *alpha*, the larger the difference between the two redshifts. We plot in

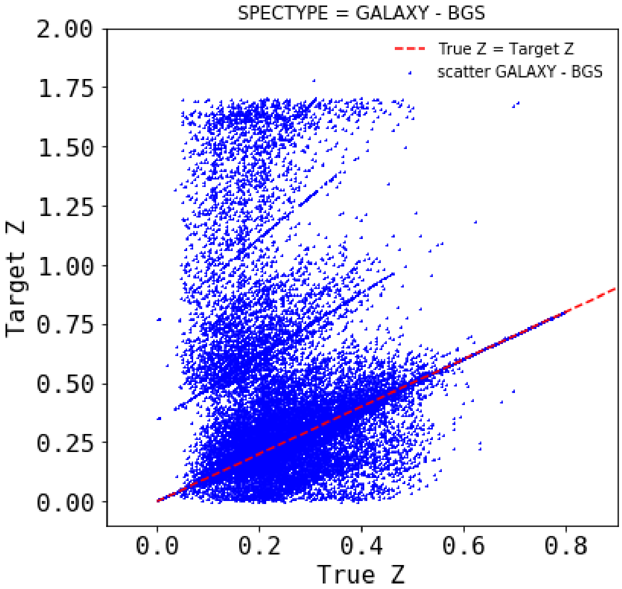
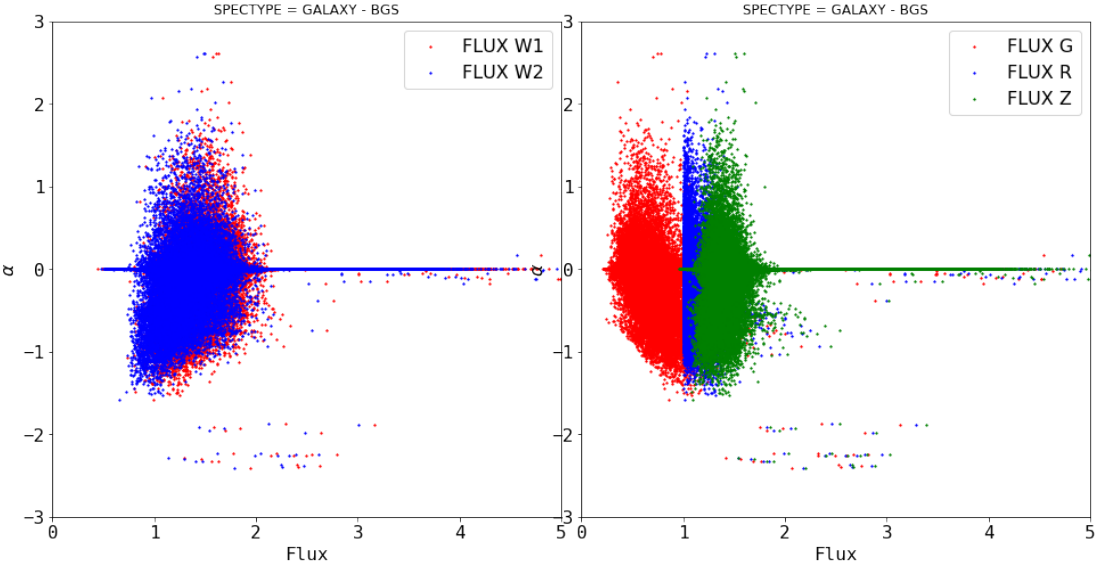


Figure 3: True and Target redsfhit relationship for BGS galaxies in the dataset. The red line indicates the line of calibration where the predicted redshift (target) is equal to the redshift of the galaxy (true)

Figure 4 the Flux G, R, Z, W1 and W2 to see if there was a pattern between this variables and α.



(a)

(b)

Figure 4: Relation between fluxes, TRUEZ and TARZ. (a) α as a function of W1 and W2 fluxes. (b)α as a function of G, R and Z fluxes. $\log_{10}$ values presented.

From Figure 4 we see that ther is a relation betwen the FLUXES variables and the redshifts encoded in α, for this reason, we choose as input variables for the machine learning model: $\log_{10}(FLUX_G)$, $\log_{10}(FLUX_R)$, $\log_{10}(FLUX_Z)$, $\log_{10}(FLUX_{W1})$, $\log_{10}(FLUX_{W2})$ and $TARZ$. For the predicted variable we chose $TRUEZ$, so our model takes in observational information *and* DESI's predicted redshift to *recover* the **true redshift** of the object.

4 Results

The dataset was split in a training set (75%) and a test set (25%). We trained three SVR and three KRR models on random subsets of the training set of size 100.000 each due to memory and training time problems, this subset was also subdivided in an 80-20 development - evaluation datasets for the application of grid search and cross-validation. The three KRR outperformed the three SVR models, then the KRR were ensemble together to produce a single oputput as Figure 5 shows. The results of the three models on the training development set is shown in Figure 6 and 7.

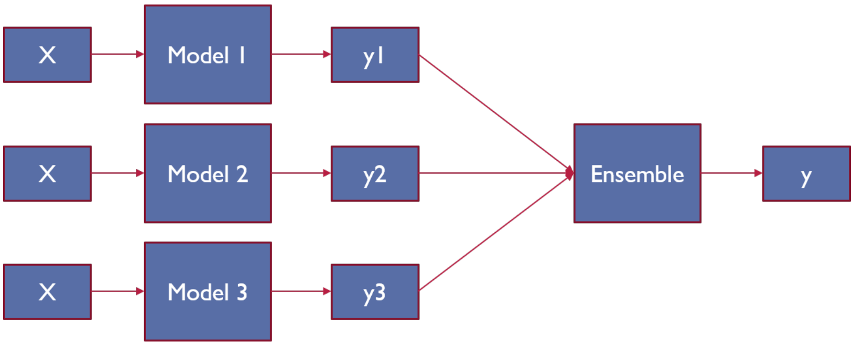


Figure 5: Graphic description of the ensemble model. It is formed based on the trained KRR models, each model previously trained on a random subset of size 100.000

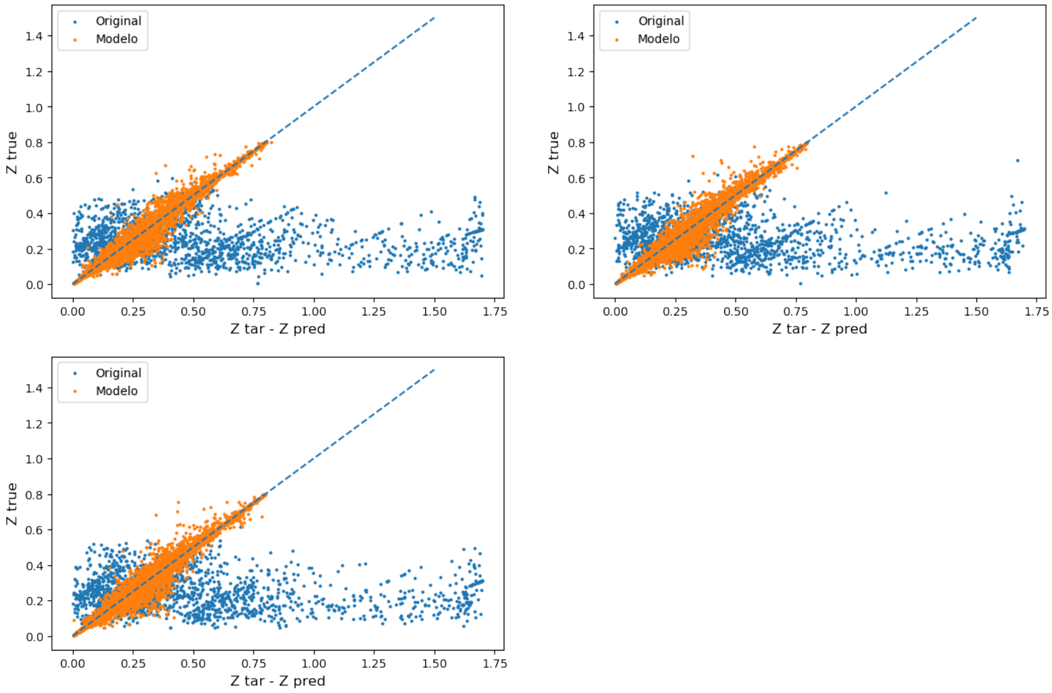


Figure 6: Results on the training set after training the KRR model.

Table 1 shows the coefficient of determination, r^2 for each regression model on the testing(unseen) subset. The coefficient of determination estimated over $n_{samples}$ is defined as

$$r^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{samples}-1} (y_i - \bar{y}_i)^2}, \tag{2}$$

where $\hat{y}_i$ is the predicted value of the i-th sample and y_i is the corresponding true value.

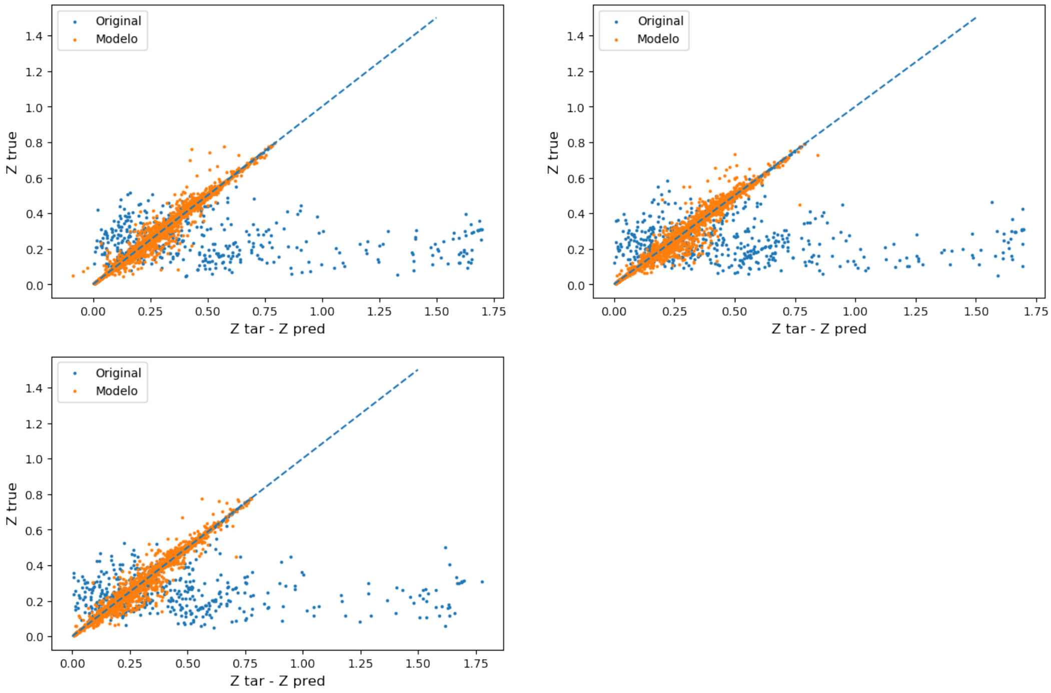


Figure 7: Results of each KRR model on the development set after training.

Model	r^2
Model 1	0.958
Model 2	0.956
Model 3	0.962
Model avg	0.962
Model max	0.949
Model w	0.983

Table 1: Results on the **test** set for the three KRR model and the ensembles. Model w corresponds to training a linear model to find the best weight coefficients of each of the three models.

5 Conclusions

The object of this monograph was to apply machine learning (ML) methods to data coming from DESI end-to-end simulations to correct (or recover) true redshift of Bright Galaxies using observational variables as input.

We found that the true redshift of the Bright Galaxy Sample (BGS) can be recovered using the variables FLUX_G, FLUX_R, FLUX_Z, FLUX_W1 and FLUX_W2 as input to kernel methods. Using an ensemble model consisting of three KRRs (Kernel Ridge Regression) trained on subsamples of size 100.000, we were capable of approximate the DESI instrument redshift measurements to its true value up to a coefficient of determination $r^2 = 0.98$. This model shows great potential to further enhancement, i.e. the fine-tuning of the model hyper-parameters, different classes of machine learning algorithms, as well as further use of ensemble methods with more than three ”subtraining” sets in order to reduce bias.

The trained models did not work when tested in other classes of galaxies, probably because of the different distributions of fluxes. However, we infer that the same methodology could be used starting by training the models in the corresponding datasets. We expect that by doing so, the results of this work can be replicated and applied to all the classes in DESI.