Recovery of redshift data - DESI

Sergio Daivd Lobo Bolaño

February 1, 2019

Abstract

DESI (Dark Energy Spectroscopic Instrument) uses a simulated survey of mock galaxies to compare their redshifts with the -also simulated- redshift measurements of DESI. 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.

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.

1.1 General Objective

- Recover the true redshifts of Bright Galaxies using observational variables from the simulated DESI measurements as an input to kernel methods of machine learning, to improve the accuracy of DESI when used in the real world

1.2 Specific Objectives

- Characterize the dataset using a simple statistical and graphical procedure to select the set of meaningful features as input to the machine learning algorithms
- Determine the set of computational parameters such as memory requirement, number of processors per node and size of dataset that performs the best on the cluster of the university restricted to constraints of resources, execution time and waiting time in the queue

- Train and tune the hyper-parameter of the models by using grid-search and cross-validation
- Test and select the best model based on performance on unseen data and model simplicity

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.

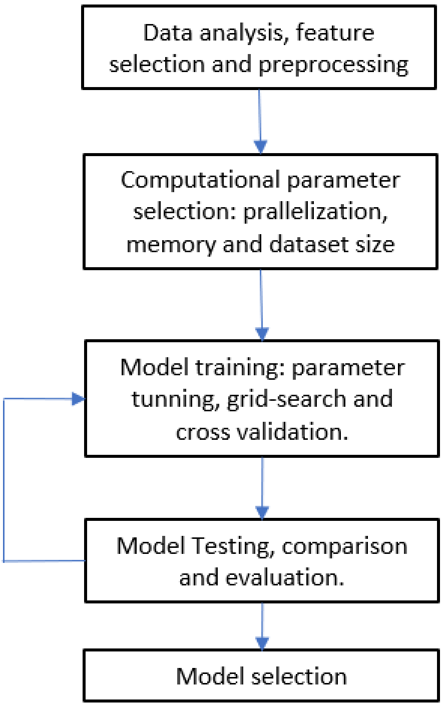

Data analysis, feature selection and preprocessing
Computational parameter selection: prallelization, memory and dataset size
Model training: parameter tunning, grid-search and cross validation.
Model Testing, comparison and evaluation.
Model selection

Figure 1: General methodology for the project.

1. Machine learning - training/testing

3 Data Overview

The true redshift and the target redshif of BGS galaxies should in principle be equal, however, due to propagating error and missclasification in the data pipeline, the predicted redshift of the objects may have a deviation from its true value. Figure 2 shows this relation, were a pattern between the two variables exist and is the reason to use machine learing.

To explain the relation between the redshifts in Figure 2, we used the Flux G, R, Z, W1 and W2 as input variables from the dataset based on the effects of this variables on the variable α defined as

$$\alpha = \frac{TRUEZ}{TARZ}, \tag{1}$$

as Figure 3 shows.

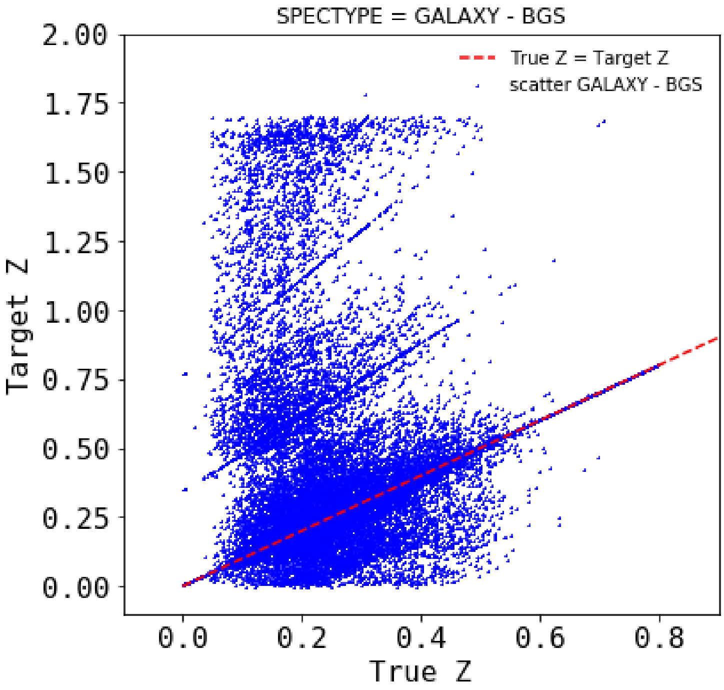
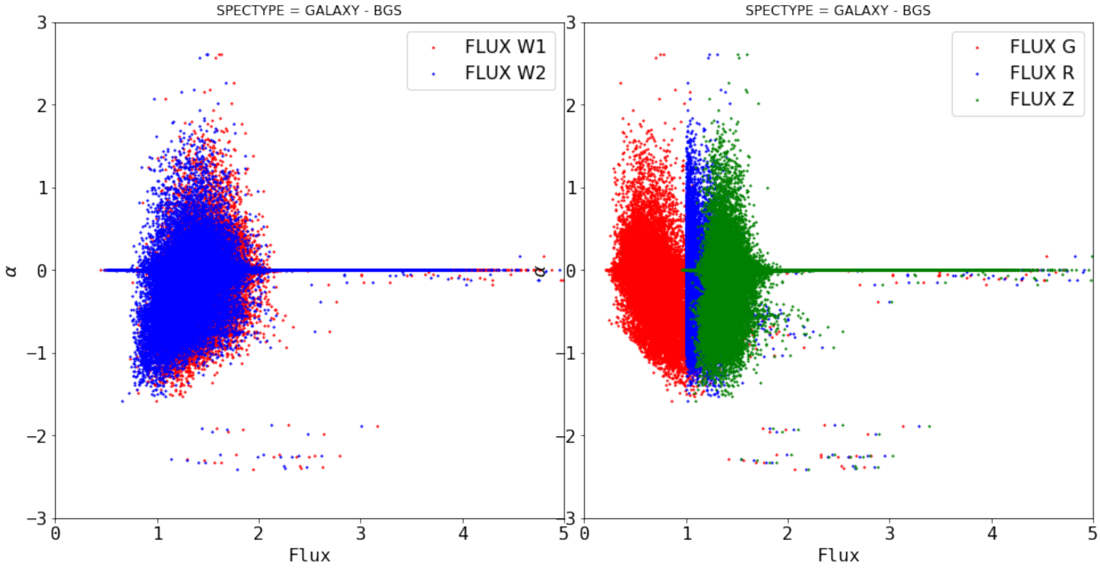


Figure 2: True and Target redsfhit relationship for BGS galaxies in the dataset. The red line indicates the line of calibration where the predicted redshift (target) is equal to the redshift of the galaxy (true)



(a)

(b)

Figure 3: 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.

Therefore, the selected variables for the machine learning model were: $\log_{10}(FLUX_G)$, $\log_{10}(FLUX_R)$, $\log_{10}(FLUX_Z)$, $\log_{10}(FLUX_{W1})$, $\log_{10}(FLUX_{W2})$ 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 4 shows.

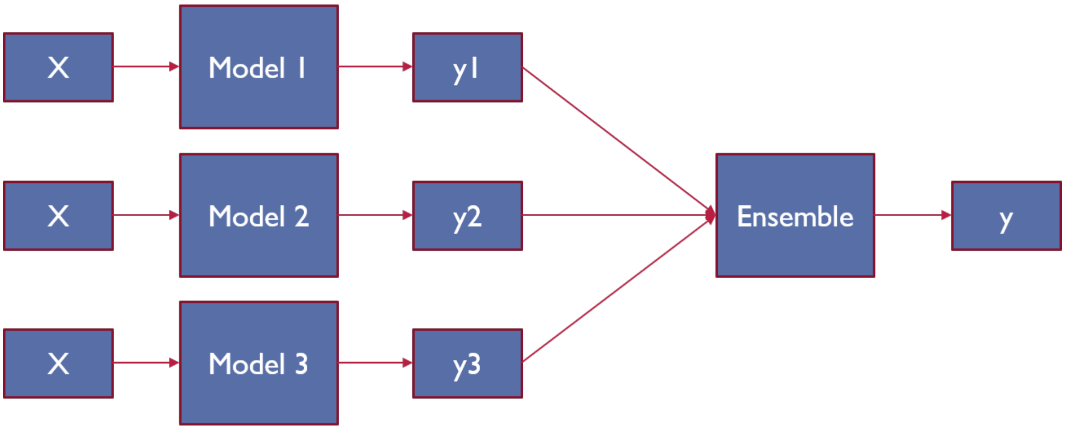


Figure 4: Graphic description of the ensemble model. It is formed based on the trained KRR models, each model previously trained on a random subset of size 100.000

Table 1 shows the coefficient of determination, r^2 for each regression model on the testing(unseen) subset.

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