Metallicity of galaxies using 5 band SDSS photometry and Supervised Machine Learning algorithms

1

*Corresponding author: email@my-email.com

Compiled September 16, 2022

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

Here we have used supervised ML algorithms to measure gas phase metallicity of galaxies, using SDSS five band photometry. We used spectroscopic estimates of metallicity as ground truth. We have divided data in to two data sets, one with spectroscopic redshift between 0.09 and 0.12 and r-band magnitude <18 and other with spectroscopic redshift between 0.2 and 0.25 and r-band magnitude between 15 and 25.Here we are using 'model' magnitudes not 'C' magnitudes.Initially features used are observed photometry and colors,than including derived quantities- stellar mass and photometric redshift too.We ran our data on Ada boost, RidgeCV, randomforest, support vector machine and extremely randomized trees. It is found that Extremely Randomised trees was best. Later we will show that although good amount of data is preferable but the results of this model are very reliable even if the data is contaminated or small and also, we calculated the contribution of experimental error in our prediction by making stimulated catalogs using guassian distribution. Further we also included spectroscopic data like 5 emission line measurements for much better prediction of metallicity. Metrices that we used are RMSE and R2 score.

2. DATA WRANGLING AND SOME CORRECTIONS

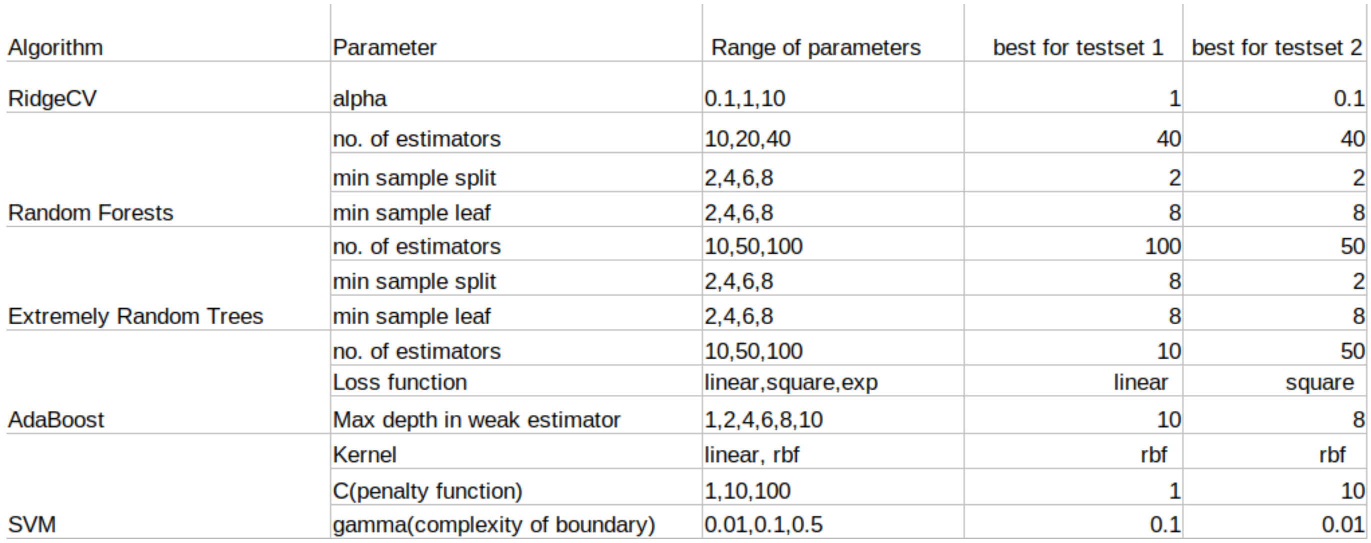
Now here we use Model Magnitude as they provide better colour estimation than C magnitude.Now removing galaxies with AGN flag and Redshift warning.Further removing objects with error in R band magnitude >0.15. Also we did dust correction in magnitude using maps of arXiv:astro-ph/9710327. Finally,adding ten independent colors and their squares.Here we don't need to do K correction as redshifts of data in both datasets differ by less than 0.1 . This will result in a total of 166246 objects. All this data and features has been downloaded from Casjobs Skyserver for Data Release 10.

3. MACHINE LEARNING ALGORITHMS AND HYPERPARAMETERIZATION

We have used five supervised ML algorithms: RidgeCV, Randomforest, Extremely Randomised Trees, Ada boost, Support vector machines and compared their Rmse values to get the best algorithm for further use. We have used GridsearchCV with 3 fold cross validation for hypertuning except for support vector machine for which we have used RandomsearchCV. We have calculated Rmse,R2 score and fraction of outliers $(Z_{true} - Z_{predicted} > 0.2dex)$ for both datasets. see Fig. 1.

4. SELECTION OF BEST ALGORITHM

Now we make two data sets, one with spectroscopic redshift between 0.09 and 0.12 and r-band magnitude <18 (24000 galaxies) and other with spectroscopic redshift between 0.2 and 0.25 and r-band magnitude between 15 and 25 (5000 galaxies). Features included are Model magnitudes of u,g,r,i,z bands, 10 independent colors and square of colors.
Now we run both datasets (80% training and 20% training) on all hyperparameterized algorithms. It is found that all ensemble methods perform very well.Extremely Randomized trees is the best,for dataset 1 and dataset 2 rmse were 0.078 and 0.104 with fraction of outliers $(|Z_{true} - Z_{predicted}| < 0.2dex)$ of 0.021 and 0.059 resp. Rmse after removal of outliers were 0.068 and 0.077 resp. Therefore we will use Extremely Randomised trees for further predictions. see fig.2 and fig.3 .



Algorithm	Parameter	Range of parameters	best for testset 1	best for testset 2
RidgeCV	alpha	0.1,1,10	1	0.1
	no. of estimators	10,20,40	40	40
	min sample split	2,4,6,8	2	2
Random Forests	min sample leaf	2,4,6,8	8	8
	no. of estimators	10,50,100	100	50
	min sample split	2,4,6,8	8	2
Extremely Random Trees	min sample leaf	2,4,6,8	8	8
	no. of estimators	10,50,100	10	50
	Loss function	linear,square,exp	linear	square
AdaBoost	Max depth in weak estimator	1,2,4,6,8,10	10	8
	Kernel	linear, rbf	rbf	rbf
	C(penalty function)	1,10,100	1	10
SVM	gamma(complexity of boundary)	0.01,0.1,0.5	0.1	0.01

Fig. 1. Hyperparameterization of algorithms for both datasets

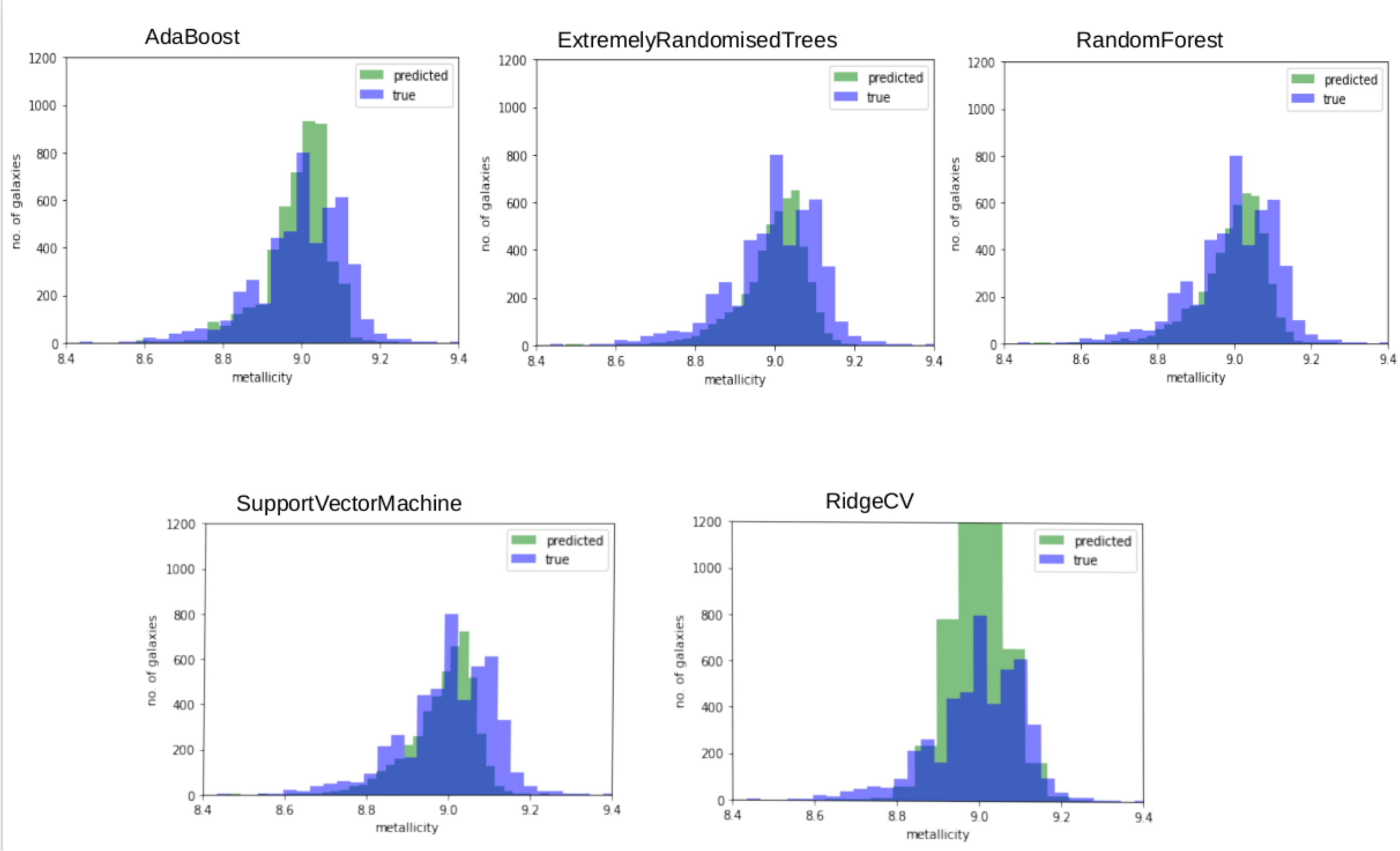
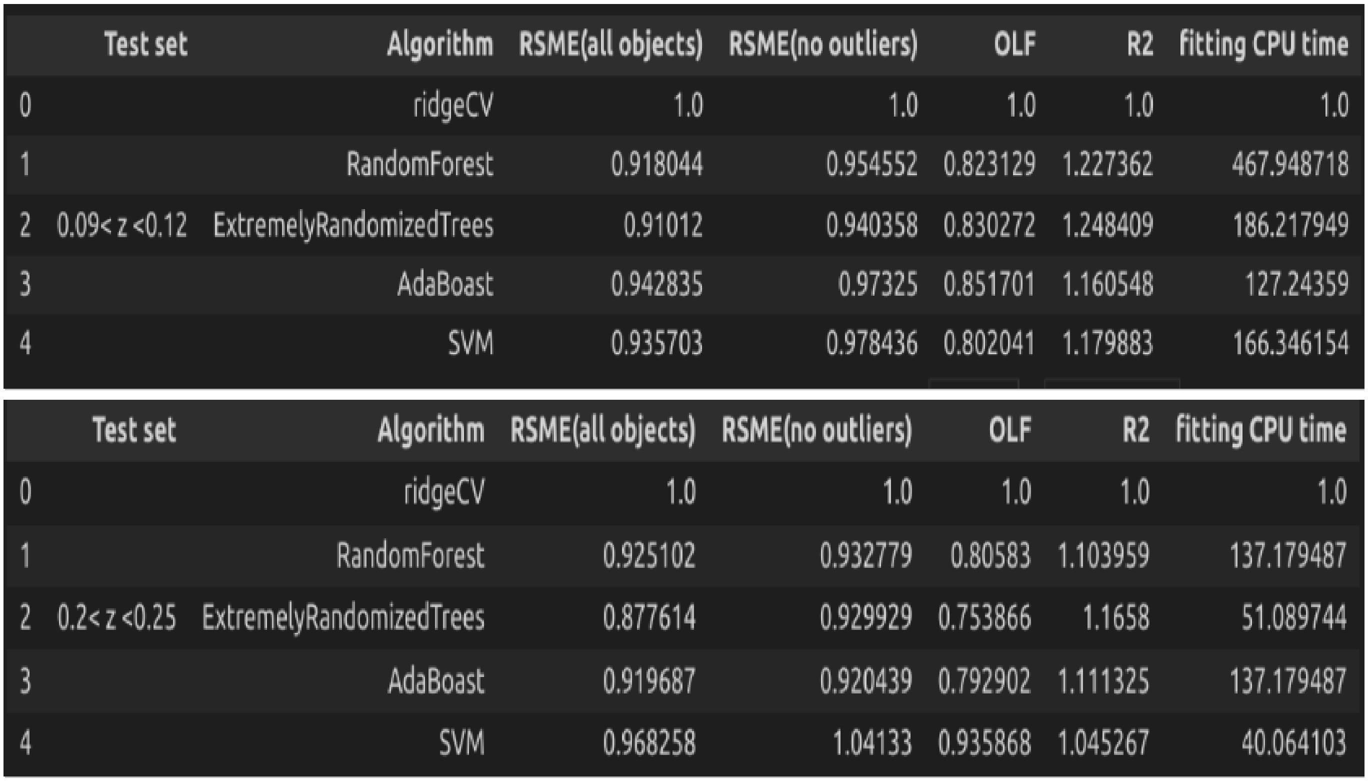


Fig. 2. These histograms are of objects with test set 1 (0.09<z<0.12 containing approx.5000 galaxies). They show metallicity(predicted and true) VS no. of galaxies graph

5. INCLUDING DERIVED QUANTITIES AS WELL

Now as we have got our best performing algorithm, we can try including features like stellar mass and photometric redshift as well for much better prediction of metallicity. As its already known to have a mass-metallicity relationship(https://iopscience.iop.org/article/10.1086/423264/pdf),adding fluxes and colors will reduce the bias and variance



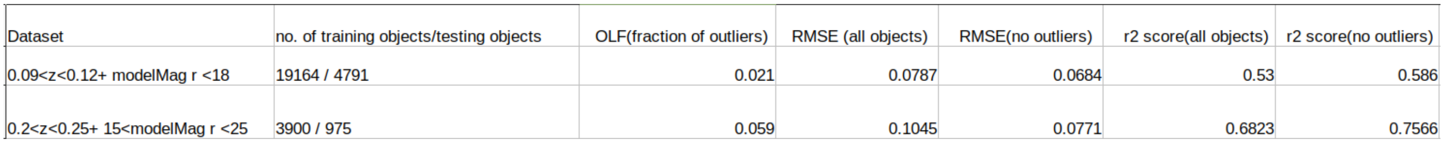
	Test set	Algorithm	RSME(all objects)	RSME(no outliers)	OLF	R2	fitting CPU time
0		ridgeCV	1.0	1.0	1.0	1.0	1.0
1		RandomForest	0.918044	0.954552	0.823129	1.227362	467.948718
2	0.09< z <0.12	ExtremelyRandomizedTrees	0.91012	0.940358	0.830272	1.248409	186.217949
3		AdaBoast	0.942835	0.97325	0.851701	1.160548	127.24359
4		SVM	0.935703	0.978436	0.802041	1.179883	166.346154

	Test set	Algorithm	RSME(all objects)	RSME(no outliers)	OLF	R2	fitting CPU time
0		ridgeCV	1.0	1.0	1.0	1.0	1.0
1		RandomForest	0.925102	0.932779	0.80583	1.103959	137.179487
2	0.2< z <0.25	ExtremelyRandomizedTrees	0.877614	0.929929	0.753866	1.1658	51.089744
3		AdaBoast	0.919687	0.920439	0.792902	1.111325	137.179487
4		SVM	0.968258	1.04133	0.935868	1.045267	40.064103

Fig. 3. Comparing optimized algorithms at low and high redshifts. These results are normalized about performance of RidgeCV. features used here are only magnitudes of bands and colors and square of colors

and lead to better regression model. But they might not increase overfitting as ETR are not prone to overfitting (also there was not much difference between taining data and testing data accuracy). As a result we see that we got rmse of 0.078 and 0.104 and fraction of outliers 0.021 and 0.059 on dataset 1 and dataset resp. And rmse of 0.068 and 0.077 resp. after removing outliers. see Fig.4 and Fig.5 .

In Fig.6 we have also calculated the ranking of features according to their contribution in predicting metallicity, using ExtraTR algorithm. We can see that mass is most important feature,as a result of known mass-metallicity relation. we can also see the variation in RMSE,r2 score, and OLF as a result of successively adding features according to their ranking.



Dataset	no. of training objects/testing objects	OLF(fraction of outliers)	RMSE (all objects)	RMSE(no outliers)	r2 score(all objects)	r2 score(no outliers)
0.09<z<0.12+ modelMag r <18	19164 / 4791	0.021	0.0787	0.0684	0.53	0.586
0.2<z<0.25+ 15<modelMag r <25	3900 / 975	0.059	0.1045	0.0771	0.6823	0.7566

Fig. 4. These results are for extremely randomized trees with features as magnitudes of bands, colors, square of colors, stellar mass and photometric redshift.

6. MORE SLICED AND RESTRICTED DATA

The most relevant comparison of metallicity measurements from photometry is the at https://ui.adsabs.harvard.edu/abs/2013ApJ.. To test it, we apply the exact same selection criteria to the SDSS data set as those applied by the above authors. The most significant cuts come from selecting the objects with r mag < 17.77, with redshift between 0.03 and 0.3, and H/H flux > 2.5, which resultas in 116391 objects. We see that by applying ETR this data we get RMSE value: 0.0976. If we apply less restrictive data like forfeiting the cuts based on Halpha and Hbeta fluxes which results in 116539 objects, we get RMSE value: 0.0971

Therefore we see that in this case a larger sample is more valuable than cleaned data as RMSE increases in later case.

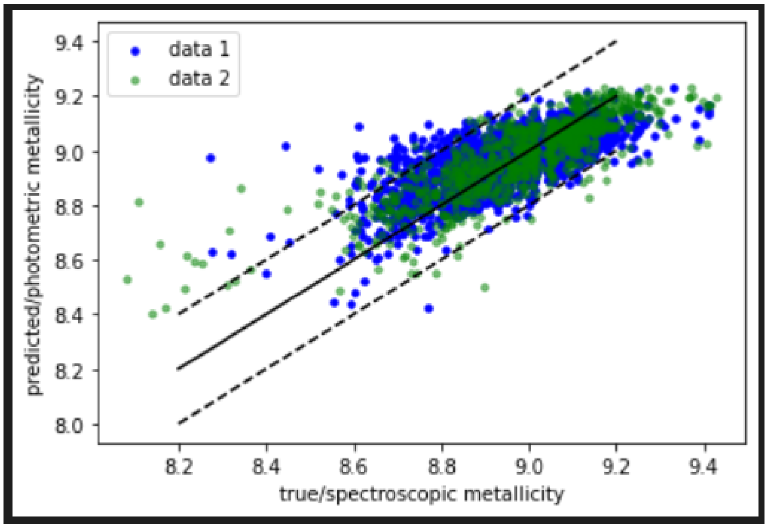


Fig. 5. solid line: 1:1 correspondence and dashed line: 0.2 dex deviation .Objects outside dashed lines are considered as outliers. We are able to find metallicity better than 0.1 dex for 85% and 79% objects for data set 1 and 2 resp. And metallicity of only 2.1% and 5.9% of objects differ more than 0.2 dex from true/spectroscopic metallicity for data set 1 and 2 resp.

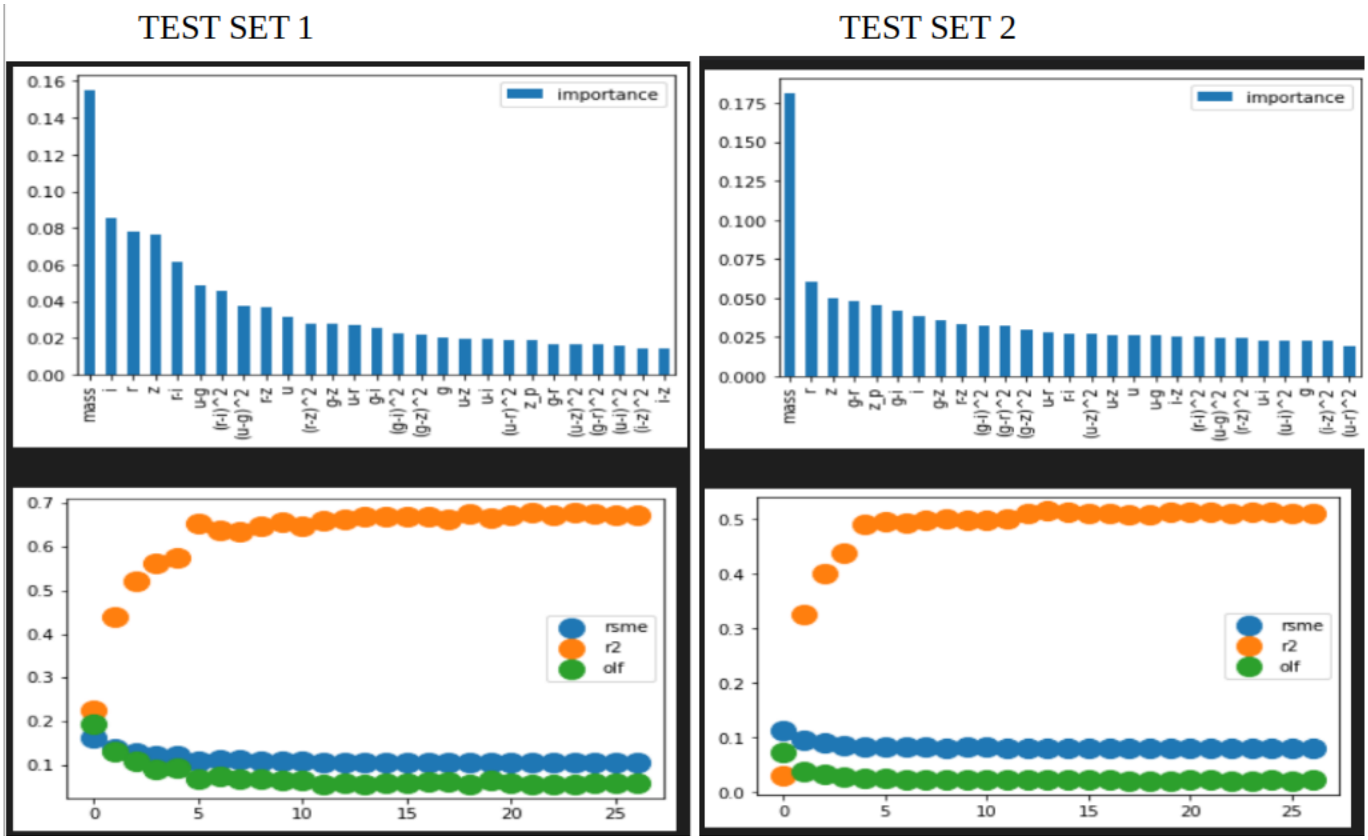


Fig. 6. These results are computed By fitting datasets in ETR. In graph below feaures are added consecutively according to their Normalized rankings in above graph.Mass is the most important and including information about luminosity, colors and square colors is essential to tighten the constraints on metallicity.

The most notable difference between the two methodologies is the fact that we don't need to apply any K-correction, since we use redshift as one of the features of our algorithm.
It is also interesting to break down the performance of the algorithm by redshift, and number of objects. We divide our

sample with width z of 0.03 between 0.03 and 0.27, and train ERT algorithm separately on each slice. that the average RMSE in the results is dominated by the objects at redshifts 0.03 < z < 0.06, which constitute a third of the sample and exhibit higher scatter. In all other slices, including those only populated by a few thousands objects, the number of outliers and RMSE are actually considerably lower, while they increase again in the last slices as a result of the excessively small sample size. See Fig. 7 .

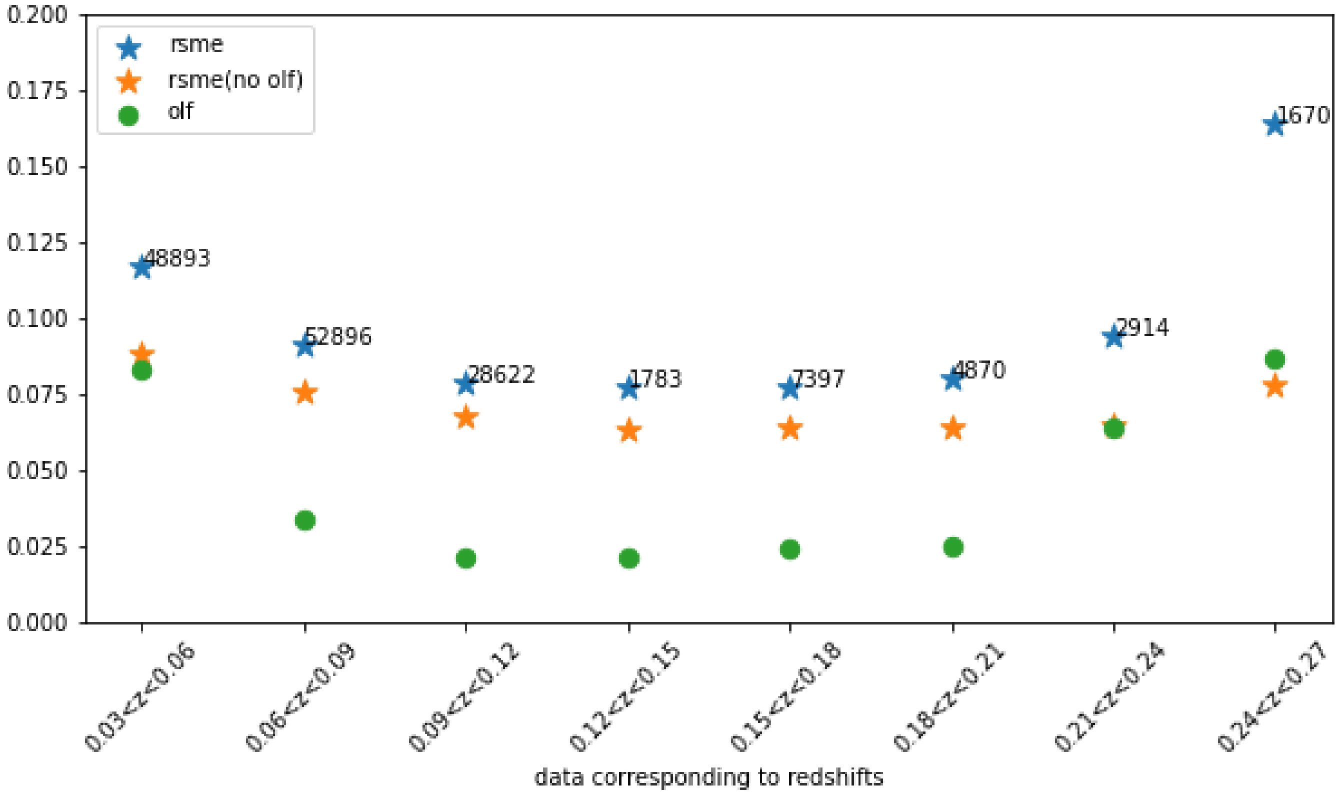


Fig. 7. divided in eight slices of uniform width z = 0.03 between z = 0.03 and z = 0.027. Sample sizes are written at each point. The objects in the first slice exhibit the highest fraction of outliers and the highest RMSE, with the exception of the very small sample in the highest redshift slice.

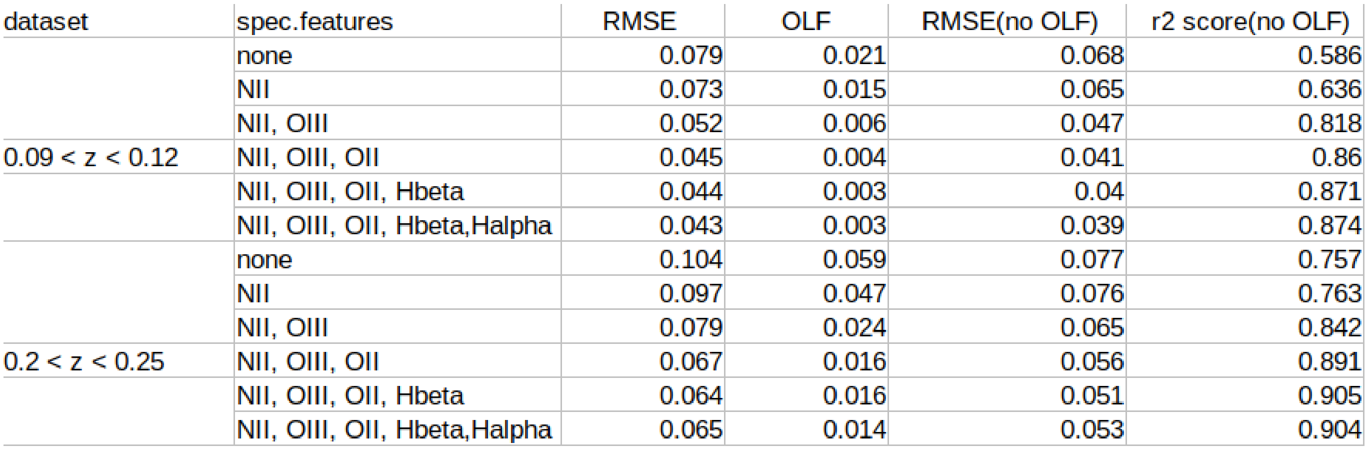
7. ERROR ESTIMATION

Now we want to know how much error in our prediction is due to the error in experimental observations (experimental uncertanities). For this we made 100 stimulated catalogs for each testset, using guassian distribution. We replaced each data point of test features with corresponding most probable value of guassian. We used observed values as mean and error in measurement as standard diviation for each band ,redshift and mass,but for mass standard diviation is used as half of difference between 84th and 16th percentile values. Then we ran ExtratreeRegressor on both stimulated catalogs. we get an rmse of 0.0725 and 0.0744 for dataset 1 and 2 resp. Therefore, after comparing it with previous results we see that error due to experimental uncertainties are 0.0040 and 0.0056 for set 1 and 2 which are pretty low.

8. COMBINING SPECTROSCOPY TOO

Here, we add to our data set five additional emission line measurements, available for the SDSS catalog: [OII] (doublet at 3726 and 3729 A), [OIII] (doublet at 4959 and 5007 A), [NII] (doublet at 6548 and 6584 A), H at 6563 A, and H at 4861 A. adding all the five emission lines does a very good job to measure metallicity, with a reduction in the root mean square error of 50when all lines are included.See Fig.8 . By ranking the features in order of importance, we observed that for both sample data sets, measurements of [OIII], [NII] and [OII] emission line fluxes were the most effective in increasing the accuracy of the metallicity measurement, and accounted for 90% of the total improvement, followed by Mass and then Halpha and Hbeta emission line fluxes although there were differences in the rankings between the two data sets. See Fig.9 .

One interesting thing to note here is that a significant reduction in RMSE is observed when we add second emission line flux.For example when we add OIII after NII, RMSE reduces from 0.073 to 0.052 in dataset 1.This confirms the well-known results that line ratios are more effective tracers of metallicity than single emission lines.



dataset	spec.features	RMSE	OLF	RMSE(no OLF)	r2 score(no OLF)
0.09 < z < 0.12	none	0.079	0.021	0.068	0.586
	NII	0.073	0.015	0.065	0.636
	NII, OIII	0.052	0.006	0.047	0.818
	NII, OIII, OII	0.045	0.004	0.041	0.86
	NII, OIII, OII, Hbeta	0.044	0.003	0.04	0.871
	NII, OIII, OII, Hbeta,Halpha	0.043	0.003	0.039	0.874
0.2 < z < 0.25	none	0.104	0.059	0.077	0.757
	NII	0.097	0.047	0.076	0.763
	NII, OIII	0.079	0.024	0.065	0.842
	NII, OIII, OII	0.067	0.016	0.056	0.891
	NII, OIII, OII, Hbeta	0.064	0.016	0.051	0.905
	NII, OIII, OII, Hbeta,Halpha	0.065	0.014	0.053	0.904

Fig. 8. adding spectroscopic data to our initial photometric data for significantly much better prediction of metallicity. Also, a good reduction in rmse by adding OIII, proves the well-known results that line ratios are more effective tracers of metallicity than single emission lines.

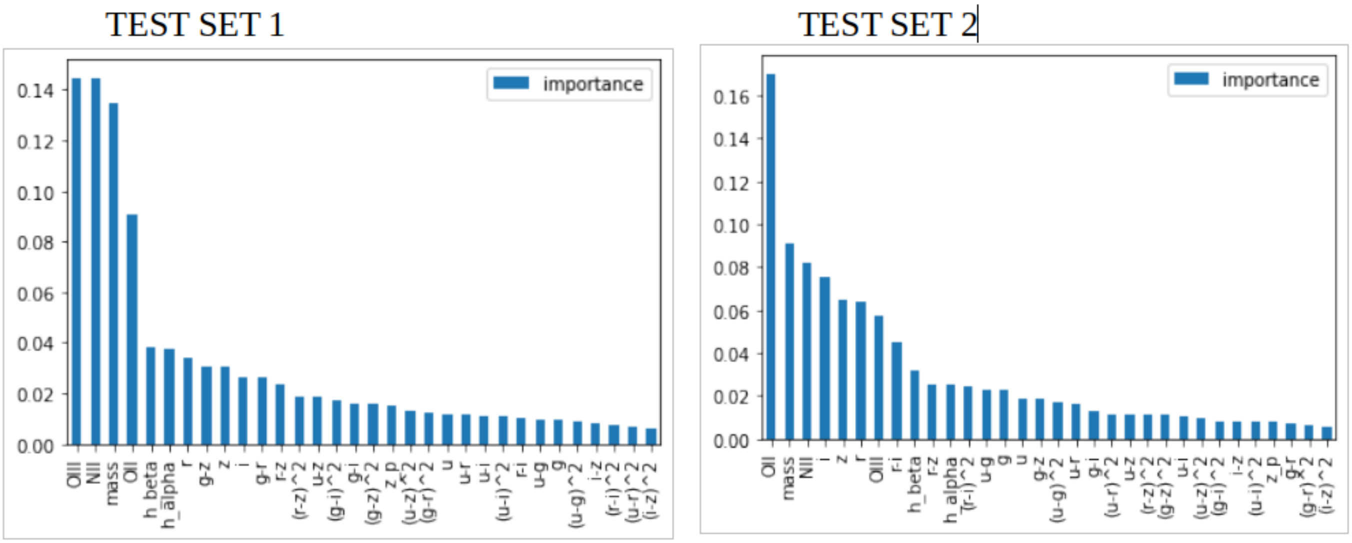


Fig. 9. Ranking of features according to their contribution for predicting metallicity, using Extra tree regressor. We see that spectroscopic data and mass occupies better ranks in both datasets

9. CONCLUSION

As a result we see that metallicity of galaxies can be calculated better than 0.1 dex using model magnitudes and colors of 5 Band data. ExtraTreeRegressor works best with least overfitting and high accuracy. using derived quantities like mass is very useful due to mass-metallicity relationship and then magnitudes and colors help to tighten the correlation. Then we also saw that a large dataset is better than more sliced one with more restriction of redshift and emission lines. Than we calculated the contribution of experimental uncertainities in RMSE (by making stimulated catalog using guassian distribution), which came pretty low. Finally we added Spectroscopic data(5 emission line measurements) to our datasets to see that they were the most important features followed by mass. we also proved that ratio of lines are much better tracers of metallicity than single line as rmse drastically decreased after adding second emission line to our dataset.