Stellar formation rates in galaxies using Machine Learning models

Michele delli Veneri[1], Stefano Cavuoti[1,2,3], Massimo Brescia[2], Giuseppe Riccio[2] and Giuseppe Longo[1,3]

1- Universitá degli Studi Federico II - Dipartimento di Fisica "E. Pancini", via Cintia 6, I-80135 Napoli, Italia

2- INAF - Osservatorio Astronomico di Capodimonte, via Moiariello 16, I-80131 Napoli, Italia

3- INFN - Napoli Unit, via Cintia 6, I-80135 Napoli, Italia

Abstract. Global Stellar Formation Rates or SFRs are crucial to constrain theories of galaxy formation and evolution. SFR's are usually estimated via spectroscopic observations which require too much previous telescope time and therefore cannot match the needs of modern precision cosmology. We therefore propose a novel method to estimate SFRs for large samples of galaxies using a variety of supervised ML models.

1 Introduction

One of the main parameters used to understand galaxy formation and evolution is the global Stellar Formation Rate (SFR) [1] which measures the luminosity-weighted average across local variations in star formation history and physical conditions within a given galaxy. The identification of proper SFR indicators and their calibration has been at the center of intensive debate for the last thirty years and many problems must yet be solved. However, in a broad oversimplification, global SFR indicators are measures of luminosity, either monochromatic or integrated over some wavelength range chosen in order to be sensitive to the short-lived massive stars. There is a large variety of such indicators spanning from the UV/optical/near-IR range ($\sim$ 0.1-5 μm) which probe the direct stellar light emerging from young stars, to the mid/far-IR ($\sim$ 5-1000 μm) probing the stellar light reprocessed by dust. Other indicators rely on the ionizing photon rate, as traced by the gas ionized by massive stars [2, 3]; hydrogen recombination lines, forbidden metal lines, and in the millimeter range, the free-free (Bremsstrahlung) emission. Finally, other indicators are, at least in principle, derived in the X ray domain, from X-ray binaries, massive stars, and supernovae; the non-thermal synchrotron emission produced by accelerated charged particles, which, following the suggestions in [4], can be calibrated as an SFR indicator.

It is clear, however, that the correct evaluation of SFR from optical/FIR broad band data is a highly non trivial task due to the complex and still largely poorly understood correlation existing between the SFR and the broad band photometric properties integrated over the whole galaxy. From the mathematical point of view, the estimation of SFR is a typical regression problem and, in a recent work [5] it has been shown how machine learning methods can be effectively used to estimate stellar formation rates
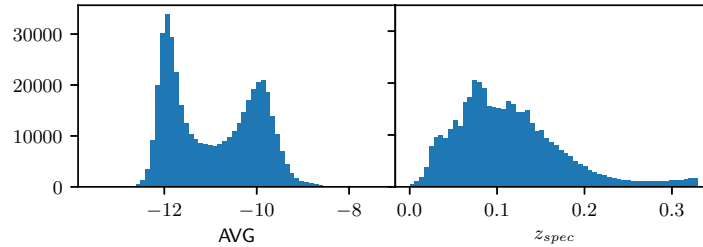
Fig. 1: *Left panel:* distribution of AVG; *right panel:* Spectroscopic redshift distribution of the training sample.

for objects in large panchromatic sky surveys. In this work we make use of the SFR's derived in [6] for a subsample of galaxies in the SDSS-DR7. While referring to the original paper for a detailed description, we shall just remind that Brinchman [6] used the Charlot & Longhetti model [7] and that the method to obtain SFR estimates for other classes of galaxies (AGN, Composite, low S/N SF, low S/N LINER and Unclassifiable) is nearly identical to that described in [6].

2 Data

In order to compare the results of our work with those obtained by [5] we used the Data Release 7 of the Sloan Digital Sky Survey (hereafter SDSS-DR7) [8]. In fact this data release was used in [6] to derive reliable star formation rates for a subsample of $\sim 10^6$ galaxies with good spectroscopy. To create our Knowledge Base (or KB), we extracted from the *SpecPhoto* view of the SDSS-DR7 all objects whose PlateID, FiberID and MJD matched the objects in the the Brinchmann et al. catalogue [6]. From the former we extracted the *psfMag*, *fiberMag*, *petroMag*, *modelMag*, *expMag* and *deVMag* magnitudes in the u, g, r, i, and z bands and the *spectroscopic redshift*; from the latter, the targets i.e. the average total star formation rates (*AVG*)[1]. Then the following constrains were applied:

1. for SFRs we required a successful estimation (*flag = 0* in [6]);

2. for redshifts we required a successful spectroscopic estimation (zWarning = 0 in [8])

A sample of 572,285 galaxies satisfied those criteria. Then we eliminated from the sample all objects with a "*Null*" feature and truncated all magnitude distributions to physically meaningful values, obtaining a KB of 570,908 galaxies. For each magnitude type, we produced the *u - g*, *g - r*, *r - i*, and *i - z* colors, reaching a total of 55 features for our KB (54 photometric plus the AVG as a target). The distribution of spectroscopic redshifts and AVGs for the KB is shown in Fig. 1 (right and left panel, respectively) . To study the relationship between redshift and star formation rate, we cross-matched the

---

[1]"Magnitudes" m are the conventional astronomical equivalent of the fluxes f_λ. More precisely $m = -2.5 Log f_\lambda + q_0$, where q_0 is a zero point.
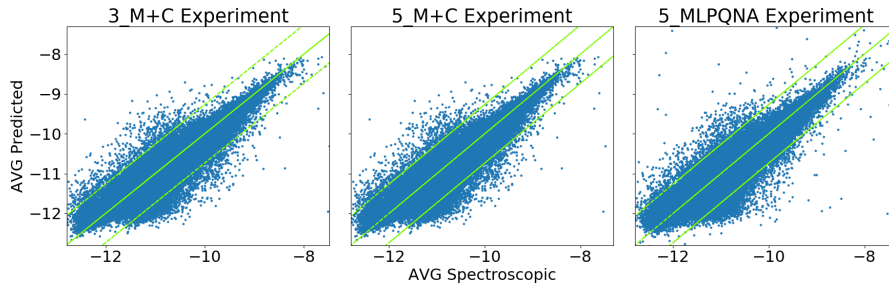
Fig. 2: Spectroscopic AVG VS Predicted AVG.The central line gives the bisector, while the other two define the 3σ regions. Objects outside these **green lines** must be considered as "outliers".

sample of 570,908 galaxies with our catalogue of photometric redshifts for the SDSS-DR9 galaxies [9], retaining 436,490 galaxies out of the 570,908. Thus integrating the features list with: the spectroscopic redshift *zSpec* and the relative photometric redshift estimations *zPhot* derived from [9], we reached a 56-dimensional parameter space (hereafter DR7-KB). Furthermore we created a second database based on the first one (DR9-KB) using the more accurate photometric data from the SDSS-DR9 [10].

3 Algorithms

In the present work we make use of two Machine Learning (ML) methods, respectively Random Forest (RF) [11] and Multi Layer Perceptron trained by the Quasi Newton Algorithm (MLPQNA) [12]. In order to optimize their performances, we apply the k-fold cross validation [13] and two feature selection methods. The latter are respectively, the forward selection [14] based on the feature importance embedded in the RF and PHi-LAB (Parameter Handling including Lasso Analysis as Benchmark). PHiLAB (Brescia et al. in prep.) is a method designed to solve the known problem of the all-relevant feature selection, i.e. the identification of the exact parameter space that could be considered as relevant in the context of a problem solution analysis. By performing an exhaustive selection of all relevant features, it enables a deeper understanding of the physical phenomena governing the specific problem (in this case the star forming rate prediction). The importance of any feature of the given parameter space is evaluated as the accuracy loss induced by a random perturbation of the original feature among the available data entries. For each feature the relative measure of significance is then compared to the statistical score obtained by evaluating the regression estimation, by means of the known LASSO statical method [15], based on the L1-norm regularization. The combination of the features supporting the optimal solution of Lasso regression with those out-coming from the random noising mechanism, is able to furnish a robust and complete set of features composing the best parameter space for the current problem [15].

4   Experiments

We performed a series of experiments divided in five groups (as listed in Table 1). In the first group, we used the full DR7-KB (54 or 55 features depending on the experiment) and the k-fold cross validation in order to compare our results with the ones obtained in [5]; in the second, we used the same KB but, instead of the k-fold cross validation, we used the traditional train-validation-test approach [16] in order to evaluate the cross validation effect on the performances of our models. In the third group, we replicated the second group experiments by replacing the full set of features with the most informative ones selected by the PHiLAB method. In the fourth group we replicated the second group experiments on the DR9-KB, in order to identify the best model to be used to produce the final SFR catalogue as well to compare the effects of different photometry. In the fifth group, we replicated the third group of experiments on the DR9 KB, using only the most informative features selected by the PHiLAB method. The experiment ID (see Table 1) is composed of a number and an acronym; the number specifies the group while the following acronym indicates the set of features:

- M+C: *Magnitudes + Colors*. We used as input features the combination of magnitudes and colors described in section 2 and as a target the *AVG*;

- ZSPEC: *Spectroscopic Redshifts*. We added to the M+C parameter space the spectroscopic redshift (*zSpec*) to check if a correlation between SFR and spectroscopic redshifts exists;

- ZPHOT: *Photometric Redshifts*. We added to the M+C parameter space the photometric redshifts estimates (*zPhot*) to check if a correlation between SFR and photometric redshifts exists;

- MLPQNA. We used the same KB used for the ZSPEC experiment in the third group and for the M+C experiment in the fifth group but, instead of the RF, we used the MLPQNA described in [17].

5   Results, Conclusions and Future Works

The statistical results of the experiments described above are summarized in Table 1. In Fig. 2 we plot the spectroscopic AVG vs our output for three experiments: 3_M+C (best experiment on DR7-KB), 5_M+C (same as previous experiment but applied to the DR9-KB) and, finally, 5_MLPQNA (best experiment on DR9-KB). In Fig. 3 we show the feature importance for those three experiments. In particular, by focusing the attention on RMSE and $\eta_\%$ (percentage of outliers having estimates outside the ($> 3\sigma$) boundaries) quantities, group 1 experiments show that the k-fold cross validation technique does not improve the performance on such KB. From group 2 and 3 it appears evident that the information about photometric redshifts does not improve the SFR estimation performance. In fact, this feature is always relegated at the bottom of the feature importance ranking list. This could be explained by keeping in mind that, due to obvious selection effects, in this work only low redshift galaxies ($z_{spec} < 0.33$) are been evaluated (see Fig. 1) and thus the photometric redshift does not carry enough information to boost the performance. On the other hand, the introduction of the spectroscopic

Table 1: Results of the experiments; in column 1 we have the ID, in column 2 the size of the parameter space, in column 3 the used model, in column 4 the KB, in column 5 the root mean squared error, in column 6 the median, in column 7 the standard deviation and finally in column 8 the percentage ($\eta_\%$) of outliers having estimates outside the ($> 3\sigma$) boundaries.

ID	feat.	Model	KB	RMSE	Median	σ	$\eta_\%$
1_M+C	54	RF	DR7	0.247	−0.022	0.247	2.08
1_ZSPEC	55	RF	DR7	0.229	−0.019	0.229	2.26
1_ZPHOT	55	RF	DR7	0.247	−0.022	0.247	2.06
2_M+C	54	RF	DR7	0.245	−0.022	0.245	2.07
2_ZPHOT	55	RF	DR7	0.245	−0.022	0.245	2.05
3_M+C	32	RF	DR7	0.245	−0.022	0.245	2.12
3_ZPHOT	33	RF	DR7	0.245	−0.022	0.245	2.10
3_ZSPEC	33	RF	DR7	0.227	−0.020	0.227	2.32
3_MLPQNA	33	MLPQNA	DR7	0.218	−0.015	0.218	2.26
4_M+C	54	RF	DR9	0.261	−0.020	0.261	1.95
4_ZPHOT	55	RF	DR9	0.262	−0.019	0.262	1.89
5_M+C	27	RF	DR9	0.261	−0.019	0.261	1.94
5_ZPHOT	28	RF	DR9	0.262	−0.019	0.262	1.93
5_MLPQNA	27	MLPQNA	DR9	0.254	−0.015	0.254	2.05

redshift in our KB, significantly improved the performance of our models but, due to the small percentage of the SDSS-DR9 that has measured spectroscopic redshifts which are not already contained in the KB, this information is of little or no practical utility in producing a wider catalogue of SFRs. From Table 1, by looking at group four and five, it is evident that our performance on the DR9-KB is worst than that on the DR7-KB; this can be traced to the changes in photometry between the two data releases (changes reflected by the differences in the feature ranking as shown in Fig. 3). Regardless of the general improvement of the DR9 photometry over the DR7, the AVG estimation depends from the DR7 photometry [6] and thus it carries over all its intrinsic errors later corrected in the DR9. At this stage of our work, we believe that the loss in performance is acceptable ($\sim 6\%$). At the moment we are performing also other experiments using different ML models and different types of SFR estimators (i.e. *specific SFR* [6]). This will enable the possibility of a direct comparison between our experiments and those described in [5]. We wish however to emphasize that these results are the best estimates of SFRs obtained so far with ML based methods and that they can therefore provide a viable alternative to the spectroscopic approach.

References

[1] P. Madau and M. Dickinson. Cosmic Star-Formation History. ARA&A, 52:415–486, 2014.

[2] D. Calzetti, J. Harris, J. S. Gallagher, III, and al. The Ionized Gas in Local Starburst Galaxies: Global and Small-Scale Feedback from Star Formation. AJ, 127:1405–1430, March 2004.

[3] S. Hong, D. Calzetti, M. A. Dopita, and et al. Blair. Large-scale Shock-ionized and Photoionized Gas in M83: The Impact of Star Formation. ApJ, 731:45, April 2011.
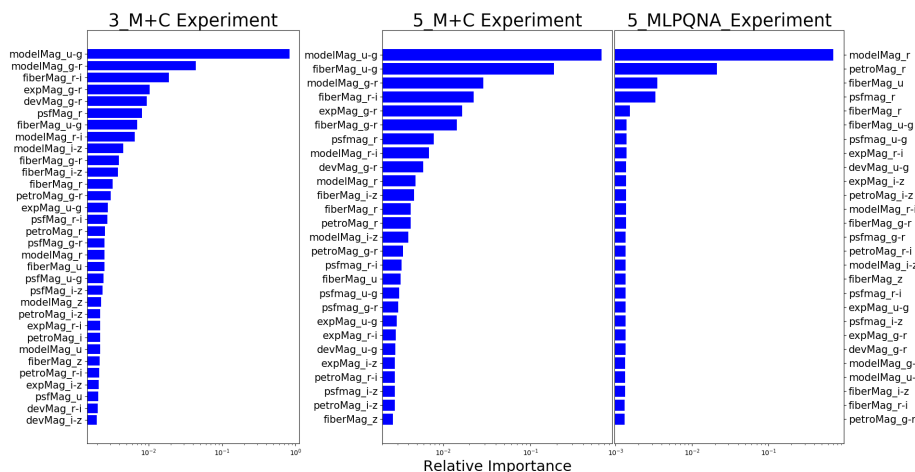
Fig. 3: Feature Importance Histogram

[4] J. J. Condon. Radio emission from normal galaxies. *ARA&A*, 30:575–611, 1992.

[5] K. Stensbo-Smidt, F. Gieseke, C. Igel, A. Zirm, and K.S. Pedersen. Sacrificing information for the greater good: How to select phot. bands for optimal accuracy. *MNRAS*, 464(3):2577–2596, 2016.

[6] J. Brinchmann, S. Charlot, S.D.M. White, and Tremonti et al. The physical properties of star-forming galaxies in the low-redshift Universe. *MNRAS*, 351(4):1151–1179, 2004.

[7] S. Charlot and M. Longhetti. Nebular emission from star-forming galaxies. MNRAS, 323:887–903, May 2001.

[8] K.N. Abazajian, J.K. Adelman-Mccarthy, M.A. Ageros, S.S. Allam, and C. et al. Allende Prieto. The seventh data release of the sloan digital sky survey. *ApJSS*, 182(2):543–558, 2009.

[9] M. Brescia, S. Cavuoti, G. Longo, and V. De Stefano. A catalogue of photometric redshifts for the sdss-dr9 galaxies (research note). *A&A*, 568, 2014.

[10] C. P. Ahn, R. Alexandroff, and Allende Prieto et al. The Ninth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Baryon Oscillation Spectroscopic Survey. *ApjSS*, 203:21, December 2012.

[11] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[12] S. Cavuoti, M. Brescia, V. De Stefano, and G. Longo. Photometric redshift estimation based on data mining with photoraptor. *Experimental Astronomy*, 39(1):45–71, 2015.

[13] Ron Kohavi. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, 1995.

[14] K.Z. Mao. Orthogonal forward selection and backward elimination algorithms for feature subset selection. *IEEE Tran.on Systems, Man, and Cybernetics, Part B*, 34(1):629–634, 2004. cited By 68.

[15] Satoshi Hara and Takanori Maehara. Enumerate Lasso Solutions for Feature Selection. In *Proceedings of the Thirty-First AAAI Conference*, pages 1985–1991, 2017.

[16] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[17] M. Brescia, S. Cavuoti, R. D'Abrusco, G. Longo, and A. Mercurio. Photometric redshifts for quasars in multi-band surveys. *ApJ*, 772(2), 2013.