Laura Hangard, 2016-05-09

Machine Learning Project

Testing SDSS classification of objects

1. Data and goal :

We perform an analysis on a SDSS dataset composed of two classes of objects : galaxies and stars. The objects in each classe are in roughly equal proportions, with 4803 galaxies and 1758 stars. Among them, we know that 1762 are stars misclassified by SDSS as galaxies. The idea here is to test the morphology classification performed by SDSS, using the same features they use, and adding new features to see if we can perform better on those misclassified objects.

For that we will use supervised classification techniques and evaluate their performances depending on the input features. The techniques used in this project are the Decision Tree, the Random Forest, the Support Vector Machine, and the boosting classification.

We preliminarily scale all the features to have a zero mean and a standard deviation of 1. We also split the dataset into a training sample containing 80% of the objects and a test sample containing the remaining 20%.

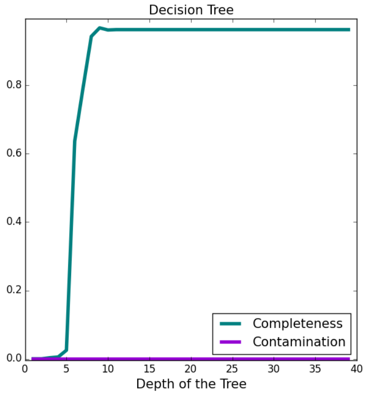
2. Confirmation of SDSS classification method on correctly classified objects :

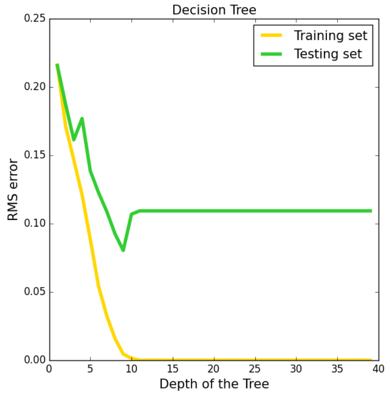
The method used by SDSS to decide if an object is a star or a galaxy is described in their documentation :

- if (psfMag - cmodelMag) > 0.145, the object is said to be “extended” and classified as a galaxy;
- otherwise, it is classified as a star.

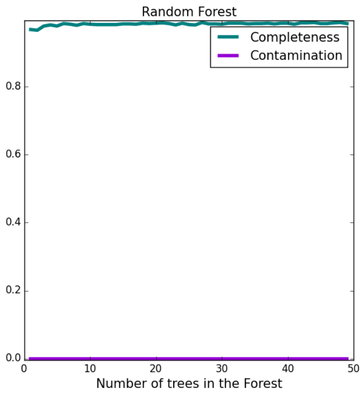
Let’s first test this classification method to see if we find the correct result, by applying the classification tools on the objects that are correctly classified by SDSS. The input features contain the ones used by SDSS to perform their classification (i.e. psfMag and cmodelMag). A first check of their criterion gives us the same classification except for 411 objects, which suggests that some other parameters are probably taken into account as well. This explains why the following results don’t show absolutely perfect classification.

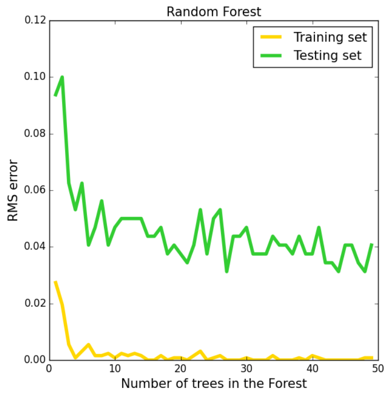
We first apply the Decision Tree tool that gives the following performances, from which we nicely deduce that the optimal depth of the tree is 10, which is what we thus apply onwards.

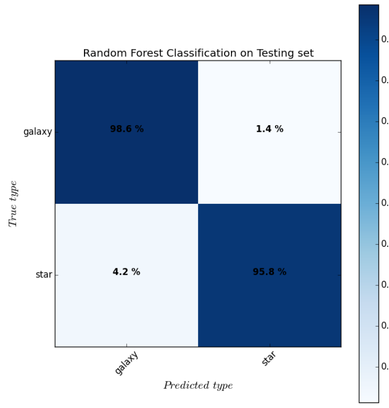




The Random Forest gives nice result as well, as illustrated in the following plots, with a completeness of almost 1 and a contamination of 0, as well as a low RMS error of 0.04. We deduce that no noticeable improvement happens for a forest of more than 10 trees, so we chose this value. The right plot shows that almost all the stars and galaxies are well classified by the model, with 4.2% of misclassified stars and 1.4% of misclassified galaxies.

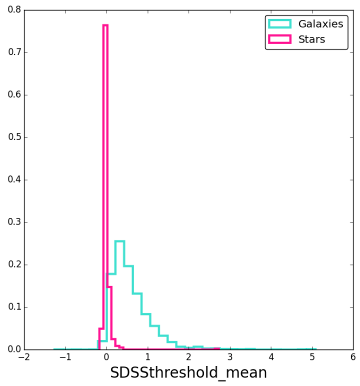


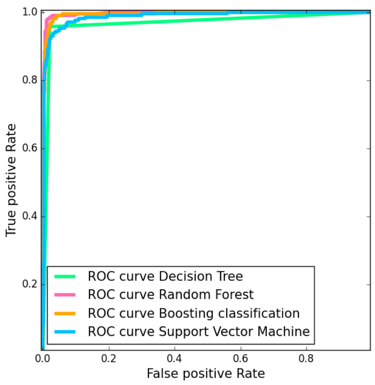




According to Random Forest, we confirm the importance in the classification model of the mean of the SDSS criterions in each bands. The nice performances of the different classification tools are expected since the class of the objects is determined thanks to the same features. The SDSS criterion is shown in the left figure below.

The final result of the performances of the different classification tool is shown in the right figure below representing the ROC curves. This plot indicates that Random Forest performs the best in this situation, with the lowest false positive rate for the highest true positive rate, followed by boosting classification.

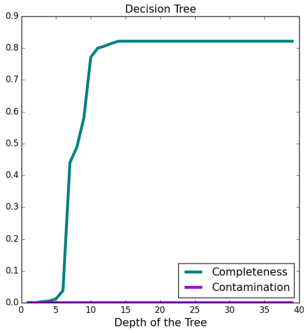


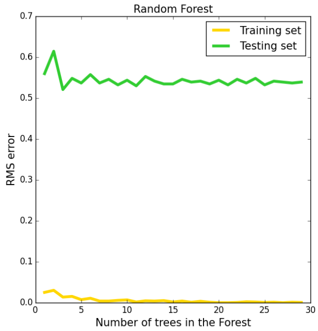


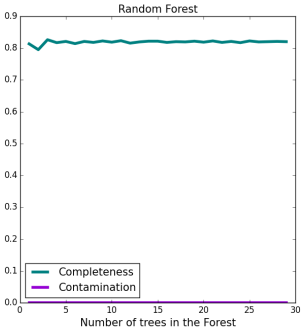
3. Testing SDSS classification with misclassified objects :

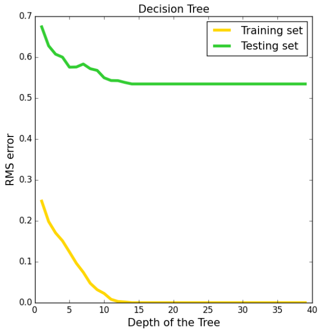
We now inject in the dataset the objects that are known to be stars misclassified by SDSS criterion as galaxies. In order to not confuse the supervised machine learning tools, we restore the correct class of these objects in our inputs.

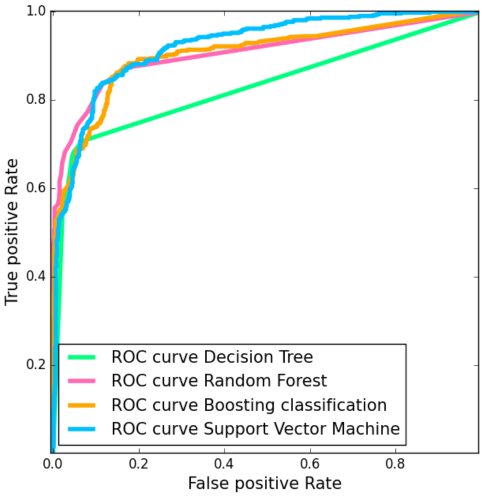
We first test how the classification tools are performing when the input features are the ones used for the SDSS criterion. As expected, we now get higher RMS errors of about 0.6 and lower completeness of 0.8. The ROC curves illustrate the lower performances, and the Random Forest tool, which remains among the best classifier, gets similar amount of misclassified galaxies, but a higher rate of misclassified stars (stars said to be galaxies) of 17.2%. So we confirm that the SDSS classification criterion is having a hard time classifying correctly the new objects that we added in this section.

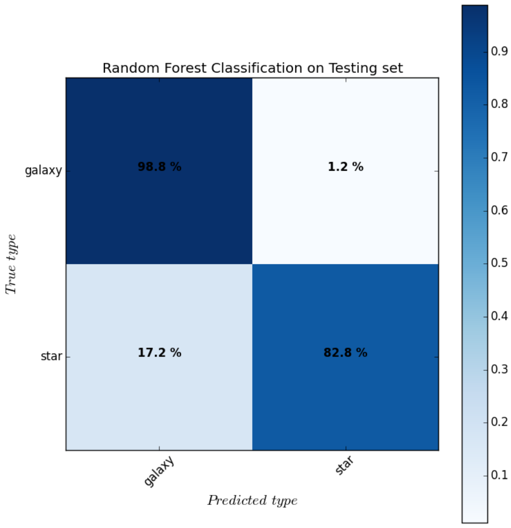






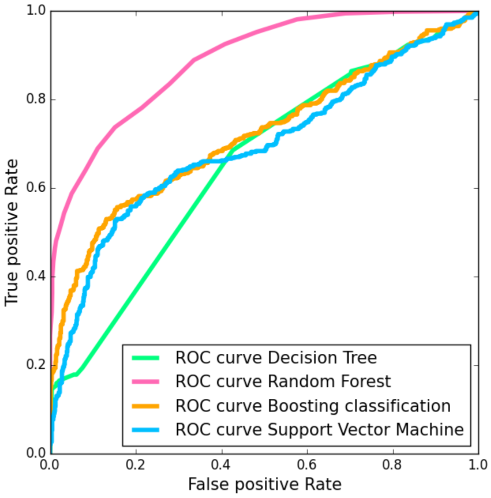


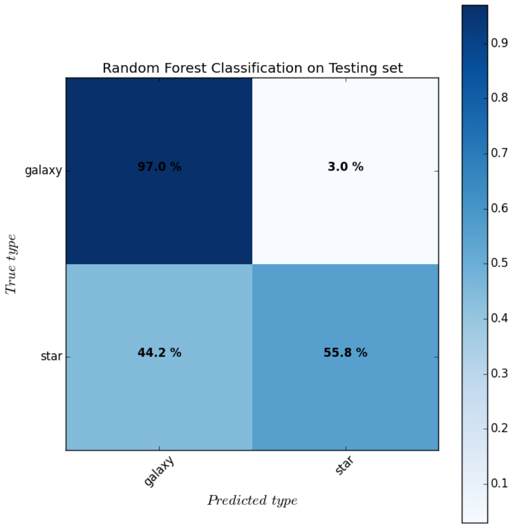




Note : The level of contamination in the plots remains ~ zero, probably because the tools chose by default the “positive” value being the class “star”. As it rarely happens that a galaxy is misclassified as a star, we stay with a false positive rate very low. The contamination in our case comes from the false negative (star misclassified as galaxy), which is not what is shown in the plots.

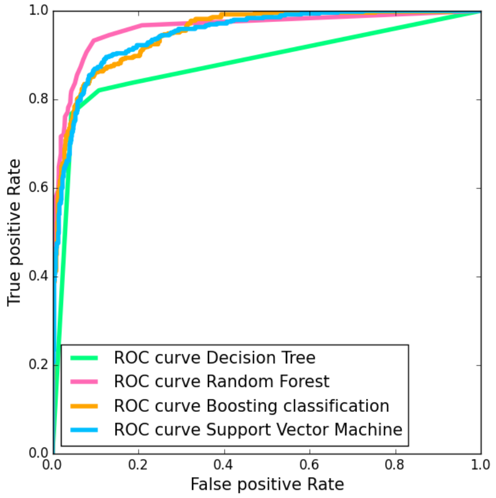
Let’s now test if we can perform a better classification using other features than the ones used by SDSS in their criterion. Instead of how “extended” the objects are, we now use their magnitudes, colours, radius, and a flag of clean photometry. The ROC curves below show that Random Forest remains the tool having the best performance on this classification. But its result is worse than previously, with 44.2% of stars misclassified as galaxies.
The most important feature in the model, according to Random Forest, is the radius of the objects.

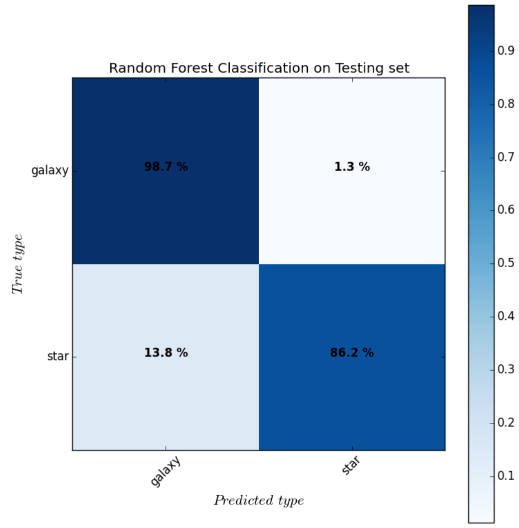




Finally, we mix both the features used by SDSS classification method, and our new tested features, to see is the SDSS criterion could be improved by the use of additional features. The result is shown below, where we can see a slight amelioration of the misclassified stars rate (13.8%), as well as better ROC curve for Random Forest than the one obtained with only SDSS criterion features.

Though, according to Random Forest, the most important feature in the model remains by far the threshold criterion of SDSS.





4. Conclusion :

We conclude that the criterion used by SDSS to classify their objects is the best feature of the models produced by supervised classification, but the use of additional features (for exemple the radius) could improve slightly the quality of the classification. Nevertheless, this would also imply a more complicated model of classification, which is a factor to be taken into account when considering a dataset as large as SDSS catalogue of objects to classify.