Abdiel Fernandez Alvarez

Kaggle competition 2

IFT 6390 - 2021

Report

Remote sensing is one of the more powerful tools for monitoring, planning, and managing on a large scale. The automation in detecting “Regions of Interest'' (ROI) has been highly enhanced by artificial intelligence. The aim of the competition is to perform a classification of surfaces as “cropland” or “non-cropland'', based on some satellite, radar and meteorological data. Two models were trained for classification, “Random Forest '' (RF) and “XGBoost” (XGB). The RF was implemented with an exploratory search grid to find prediction a baseline. A RGB was trained with the objective of enhancing the prediction score and the generalization capacity of the RF model. To increase the confidence, prediction scores of models were taken from the cross-validation method. Surprisingly, after several attempts of parameter exploration the results from XGB didn’t represent an improvement on RF, which last as a better estimator for this problem.

*Dataset description*

The given dataset is composed of two sections, training and test(non\_labeled). The training set contained 62000 samples, 216 features with labels and the test was composed of 1200 samples without labels. A segmentation on training(80%)/validation(20%) was performed on the training set. The distribution of classes turned out in 66% class 1 and 33% class 0 in both data sets. To reduce possible noise in the data, we applied a correction of missing values by replacing missings for the means of columns on the original dataset, before segmentation.

Regarding the features, it is well known in remote sensing literature that slope and elevation are both good estimators for crop regions, however, the considerations of these features with monthly frequency can be seen as redundancy (slope and elevation doesn’t change in a year). In addition, we confirm that slope and elevations vary from month to month in the dataset, which could be considered as an inconsistency. The topological information was summarized in two new columns (ElevationMean, SlopeMean), replacing the month values (12 columns Slope, 12 columns Elevation) for its “annual average”. Those two columns were found between the three more important recurrently by all models.

*Methodology:*

A RF algorithm was applied, with a preliminary research grid (see below), as the first attempt to understand the data and establish a prediction baseline. The selection of RF was made based on several reasons: it’s appropriate for classification, has the capacity to deal with data in different scales; every tree considers just a few features so it’s not affected by the curse of dimensionality and in general it’s capable to avoid the problem of overfitting since output classes are chosen by majority vote or averaging(3). The exploratory search grid :

param\_grid = { 'n\_estimators': np.linspace(50, 400, 5).astype(int), 'max\_depth': [None] + list(np.linspace(3,20,5).astype(int)), 'max\_features': [None] + list(np.arange(0.5, 1, 0.1)), 'bootstrap': [True, False]}

From this research was extracted the model parameters of the best estimator, the list of feature importance and the first prediction on the test set (test\_no\_labels).

Since the result coming from RF was very prometizing, it was decided to try to improve the RF performance by applying a boosting method. An XGBoost algorithm was applied (XGBClassifier from xgboost) with a search grid as well. To justify the consistency of the classifier found with XGBoost, the best classifier was fitted to the data several times and scored on the performance of classifying the validation set. Each iteration was made in a new random segmentation of the dataset keeping the proportion (66%training/33%Validation) and the score of the algorithm was taken from the average of 3 rounds of cross validation (StratifiedKFold from sklearn.model\_selection) every time.

The predictions with XGBoost don't perform better than RF in the non\_labeled dataset. For this reason improvements were tried for the RF research in order to achieve a better score. A more large research grid was used to train and selec a new configuration for RF that better predicts on the on-labeled dataset:

param\_grid = {'n\_estimators':[200,1000,20],'max\_depth': [None] + list(np.linspace(3,20,10).astype(int)), 'max\_features': ['auto', 'sqrt', None] + list(np.arange(0.5, 1, 0.1)), 'max\_leaf\_nodes': [None] + list(np.linspace(5, 50,10).astype(int)), 'min\_samples\_split': [2, 5, 10],'bootstrap': [True,False]}

A new better estimator was found:

*We can divide evaluation metrics into three useful groups; they are:*

1. *Threshold Metrics*
2. *Ranking Metrics*
3. *Probability Metrics.*

*This division is useful because the top metrics used by practitioners for classifiers generally, and specifically imbalanced classification, fit into the taxonomy neatly.*

*Several machine learning researchers have identified three families of evaluation metrics used in the context of classification. These are the threshold metrics (e.g., accuracy and F-measure), the ranking methods and metrics (e.g., receiver operating characteristics (ROC) analysis and AUC), and the probabilistic metrics (e.g., root-mean-squared error)(7).*

*Result and discussion.*

From the preliminary study with the RF and research grid, the best estimator performed a score of 0,87 on the validation set and baseline of 0,97831 (Kaggle: CropHarvest - crop vs. non-crop) on the test\_set without labels. The best classifier was composed as follow:

'n\_estimators': 110, 'min\_samples\_split': 5, 'max\_leaf\_nodes': 43, 'max\_features': 'auto', 'max\_depth': 16, 'bootstrap': True

The ROC curve was one of the criteria to evaluate the quality of predictions (on the validation set). From Fig1a. we can see how the relation between true positives and false positives it’s favorable on this prediction. The traditionally used F1 score (F1 = ) is equally elevated.

The XGBoost algorithm was implemented with a research grid as well. The best parameters from the found best estimator are:

'n\_estimators': 108, 'min\_samples\_split': 5, 'max\_leaf\_nodes': 17, 'max\_features': 'auto', 'max\_depth': 4, 'bootstrap': True.

This setting gave a score of 0,92 (Fig 1 b.) on the validation set, however, once uploaded on Kaggle competition system, the score (0.94230) of prediction in the non labeled dataset was lower than the RF score. Seems to be that XGBoost it’s not making better the generalization capacity of the RF model.

|  |  |
| --- | --- |
| a) | b) |
| Fig 1: ROC curve. | |

Segunt RF

Cross Validation[0.84032258 0.8016129 0.82580645 0.83870968 0.83548387 0.80322581

0.83064516 0.81774194 0.80645161 0.82258065 0.83225806 0.81290323

0.82580645 0.82096774 0.85 0.83709677 0.82741935 0.82741935

0.82096774 0.82580645]

# CONCLUSIÓN

* The RF model it’s able to better explain the data, even when some standard metric like ROC shows more favorable for XGBoost.
* XGBoost makes good scores on the validation set but it's not a good estimator to generalize and make predictions on unseen data.

# Bibliografía:

# <https://machinelearningmastery.com/random-forest-ensembles-with-xgboost/>

# <https://towardsdatascience.com/xgboost-fine-tune-and-optimize-your-model-23d996fab663>

1. <https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/>
2. <https://www.askpython.com/python/examples/k-fold-cross-validation>
3. <https://machinelearningmastery.com/how-to-calculate-precision-recall-f1-and-more-for-deep-learning-models/>
4. <https://machinelearningmastery.com/avoid-overfitting-by-early-stopping-with-xgboost-in-python/>
5. ​​<https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/>

Apendice.

Apendice 1: List of feature priority of best estimator from RF search grid.

|  |  |
| --- | --- |
| Feature | importance |
| ERA5\_temperature\_2m\_oct  ERA5\_temperature\_2m\_jan  topo\_elevation\_aug,  ERA5\_temperature\_2m\_dec,  Topo\_elevation\_nov,  ERA5\_temperature\_2m\_feb,  Topo\_elevation\_jul,  topo\_elevation\_mar  ERA5\_total\_precipitation\_sep,  ERA5\_temperature\_2m\_nov,  ERA5\_temperature\_2m\_sep,  Topo\_elevation\_feb,  Topo\_elevation\_apr,  ERA5\_temperature\_2m\_mar,  Topo\_elevation\_jun,  Topo\_elevation\_jan,  topo\_elevation\_oct,  Topo\_elevation\_sep,  ERA5\_total\_precipitation\_aug,  ERA5\_total\_precipitation\_nov,  ERA5\_temperature\_2m\_aug,  ERA5\_total\_precipitation\_jul,  Topo\_elevation\_dec,  ERA5\_temperature\_2m\_apr,  Topo\_elevation\_may  ERA5\_total\_precipitation\_mar,  S2\_B3\_mar,  topo\_slope\_apr, | 0.042420556546751205  0.03997051506103911  0.03799570339999167  0.03700179409427839  0.03299495499771793  0.03190773746300688  0.03157447791592932  0.02842906457066808  0.02774645763273691  0.02726401655579359  0.026634488902986257  0.02570901687846867  0.0230376581951413  0.022345260876014027  0.018477366272203585  0.01766505549452702  0.017004517547519433  0.016498790870317623  0.016436360799885263  0.015806705872393766  0.01331562152956314  0.013088099201785967  0.012857737421517931  0.012617980084716044  0.012277800331249998  0.012195990379210892  0.011135704561710154  0.011039795272112374 |