MSD 2019 Final Project

A replication and extension of Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data by David Muchlinski, David Siroky, et. al., October 22, 2015

Deniz Ulcay, Vatsala Swaroop, Ongun Uzay Macar (du2147, vs2671, oum2000)

2019-05-13 23:33:14

Contents

Brief Comments regarding the reproducibility of the original paper

i) In terms of variable importance, we can see the same variables in the plot as the 20 most important variables. However, the order of these variables are different than those reported in the original paper

ii) The F1 score variation plot couldn't be replicated because the author's paper did not make these training test split ratios clear. We do not know the seed they used to shuffle their data. We can definitely split our data as per different ratios but we do not get exactly similar characteristics in the resulting training - test set (basically war - peace ratio samples)

iii) The seperation plots we generated for author's replicated models and the ones presented in the original paper are somewhat similar for all of the models. In some runs, it is different due to the downsampling in Random Forests.

iv) F1 score is different frm the ranges in the paper. We get a much lower score, the highest reaching 0.4 for a model. However, the paper reports ranges of upwards 0.6.

v) For ROC, as the model is used to predict samples that it has seen during the training process, an AUC of 0.97 obtained this way. This is higher than 0.91 based on cross validation, reported in the paper. The same error has affected the receiver operating characteristic (ROC) curves and the separation plots for all the classifiers. The order of their performance is similar.

Analysis for the paper

We observed some glaring issues in the paper

- A lot of results/graphs could not be replicated and hence, we are not convinced that Random Forests was necessarily the best model possible.
- The results reported strangely did not use out-of-sample data.
- Different models were trained on different features. Yet, the author presented a comparison between them without accounting for this discrepancy in features.
- The original code given by the author did not contain replication for some important parts such as F1 scores.
- Also, the two datasets used by the author in the paper to show different countries in which civil war was reported. We don't know why this is, but we expected that similar countries should show civil war cases in both datasets.

Comparing our models / findings

Model	AUC
ReP Penalized LR F1	0.788
Rep Penalized LR CH	0.8
ReP Penalized LR HS	0.826
Rep Random forest	0.971
ReP Uncorrected LR F1	0.789
Rep Uncorrected LR CH	0.842
Rep Uncorrected LR HS	0.83
Rep Random Forest	0.971
Uncorrected LR F1	0.805
Random Forest F1	0.754
Uncorrected LR Ch	0.837
Penalized LR Ch	0.796
Random Forest Ch	0.929
Uncorrected LR HS	0.837
Penalized LR HS	0.834
Random Forest HS	0.912
Uncorrected LR AS	0.867
Penalized LR AS	0.812
Random Forest AS	0.943
Uncorrected LR AS smoted	0.838
Penalized LR smoted	0.778
Random Forest smoted	0.909
AS_DT	0.724
AS_Boost	0.944

- In terms of AUC scores, we observe that the random forest performs better than Logistic Regression Models in most cases.
- The AUC score for Random forests with training test split is lower than that of the author's Random forest model which is tested on the training set itself. For logistic regression models though, this value is comparable to the one replicated in the paper. In fact, it is even higher in some cases.
- In terms of F1 score, it is generally poor for most of the models.
- Except in the penalized case, generally the AUC scores increase for more variables in the specification.
- For smote analysis - We tried variations for different nearest neighbours, upsampling and downsampling. While our ROC score is comparable to paper results, F1 score is quite poor. Since test set is still skewed and training set reasonably balanced, our model estimates far more cases of False positives on the test. Hence, the precision is low and that leads to low F1 score.We believe that ROC is a much more important metric in this case.
- We also noticed that while the ROC score for different settings in smote was somewhat similar, the confusion matrix generated varied considerably in terms if the number of false positives. For eg. with 20 nearest neighbours far more false positives were generated than with 5 nearest neighbours. We have commented out/not included the code for these multiple variations to avoid unnecessary clutter.
- Our experimental Boosted model performs reasonably well and matches Random forests in terms of performance. However, decision tree has a much lower AUC score. This is perhaps because it wasn't complicated enough to capture the prediction.