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)

2019-05-15 02:31:59

Contents

Comments About Reproducibility

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 scores are different from the ranges presented in the original paper. We derived a substantially 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 Summary

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.

Performance of Our Implemented Models

Model	AUC	F1
REP Uncorrected LR FL	0.789	0.01709402
REP Uncorrected LR CH	0.842	NaN
REP Uncorrected LR HS	0.83	0.01680672
REP Penalized LR FL	0.788	NaN
REP Penalized LR CH	0.8	NaN
REP Penalized LR HS	0.826	0.01694915
REP Random Forest AS	0.971	0.11582626
Uncorrected LR FL	0.805	NaN
Penalized LR FL	0.805	NaN
Random Forest FL	0.754	0.2068966
Uncorrected LR CH	0.837	NaN
Penalized LR CH	0.796	0.09859155
Random Forest CH	0.929	0.42105263
Uncorrected LR HS	0.837	0.07407407
Penalized LR HS	0.834	0.08333333
Random Forest HS	0.912	NaN
Uncorrected LR AS	0.867	0.09375
Penalized LR AS	0.812	NaN
Random Forest AS	0.943	NaN
Uncorrected LR AS SMOTED	0.782	0.08612440
Penalized LR AS SMOTED	0.758	0.05263158
Random Forest AS SMOTED	0.905	0.18947368
Decision Trees AS	0.724	0.1481481
Boosted Classification Trees AS	0.936	0.0800000

NOTE: NaN in F1-scores indicate that True Positives + False Negatives = 0 has occurred.

Closing Remarks & Statements & Findings

- First and foremost, we have observed that the Random Forest models perform better than Logistic Regression Models in most cases when the AUC score is considered.
- The derived AUC scores for Random Forest models when a *training-test split* applied is lower than that of the author's reported Random Forest model, which was only tested on the training set, and hence *in-sample* data. Our replication indicated an AUC of about 0.97 whereas the authors have reported it to be 0.91. If the ROC curve for this model on the original paper is examined, one can easily infer that their value might be an understatement.
- For logistic regression models, this AUC-scores are generally comparable to the ones reported in the original paper. At times we have even observed that our findings indicated better performances than reported ones.
- The AUC scores increase for more variables in the specification except in the penalized logistic regression case.

- The F1-score, as seen on the below table, was generally poor for most cases. Both the training and test sets are heavily skewed due to nature of the data, and hence our models estimates far more cases of False Negatives on the test. The low precision then eventually leads to a low F1 score. We believe that ROC is a more representable, if not reliable, metric in this case. In the future versions of this experiment, one might try out different upsampling and downsampling methods to see how high they can get their F1-scores.
- Although the ROC curves and the accompanying AUC scores reflect that random forests do better than logistic regression models, even when analyzed with fair methods as described, the low and NaN F1 scores (TP + FN = 0) indicate that when predicting civil war onsets random forest models and logistic regression models do comparable in *most cases*. Or if we are to be more precise, both are unsuccesfull in terms of predictive power in the general case. Hence, we might easily disagree with the general argument of the original paper. This is not to say that we disregard the possibility of better models; we believe that the paper has good intentions in pushing the statistical/computational political science in the right direction.
- Furthermore, it is valuable to note that in terms of F1 score, it is observed that the Random Forest model trained on **Collier and Hoeffler (2004)** specification had the best classification accuracy, correctly identifying 8 out 23 civil war onsets. It seems that although AUC scores increased as number of variables/features increase, the F1 scores show a reverse pattern.
- We tried variations for different nearest neighbours, upsampling and downsampling while we implemented our SMOTE algorithm. Although the SMOTE has increased the F1 scores of models with author specification (2016), it also produced a lot more false positives into the confusion matrices. Perhaps with such unique and crucial data, synthetic data generation doesn't work so well.
- When the experimental ML models we have implemented is taken into consideration, Boosted Classification Trees w/ Adaptive Boosting performs reasonably well and matches Random Forests in terms of ROC and AUC-score performance. However, the custom Decision Tree model had a substantially lower AUC score. This might be an indication that the decision tree model wasn't able to capture the complex relationships between the features and the dependent variable in the data. Then again, the F1-score for the Decision Trees model is actually higher than the Boosted Classification Trees model, but this comes at the cost of more false positives. Considering that false positives correspond to predicted civil war onsets in our experiment, this might again be another indication that F1-scores are not really reliable in this specific problem.