Discussion of *Supervised Learning and Rare Events*

Mirko Moscatelli

Bank of Italy

October 22, 2019

Overview

Summary

The paper uses supervised learning to detect reporting errors of the German securities holdings data at the Bundesbank. In particular, it deals with:

1. Improving the training dataset by using text mining on the written inquiries about implausible data points, and joining the result with those data points that changed during the initial and the final report of the data (i.e. after the data quality management process).
2. Feature engineering.
3. Overcoming the imbalanced learning (rare events) problem using a combination of asimmetric weights and sampling.
4. Improving the forecasting performance using a stacking approach that combines k-nearest Neighbors, Logistic Regression and Random Forest.

Overview

Tables

Adding a few tables could greatly help understanding the problem (and the paper):

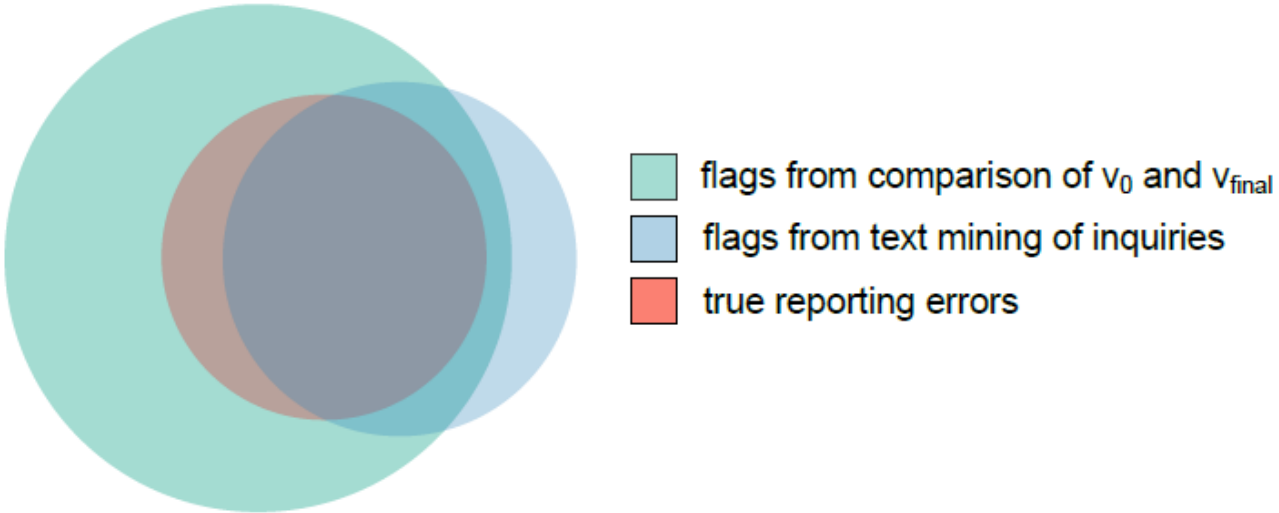
- Description of the training dataset (before and after rebalancing): number of observations, number of erroneous observations, summary statistics of the predictive variables.
- Variable importance for the three models (Nearest Neighbors, Logistic Regression, Random Forest), to understand the relevance of each predictive variable to their accuracy.
- Results of the stacking regression, to understand the relative importance of each model to the overall accuracy.
- Including also Nearest Neighbours and Random Forest in the final comparison.

Non-table stuff: number of neighbors in SMOTE, how feature engineering is done.

Training dataset

Since false negatives have a much larger cost and data is naturally scarce, it could be worthy to see if using the union dataset can provide additional useful information:

- Green points could identify errors that were corrected by the reporting institutions but didn't have a written inquiry on them by the supervisory authority.
- Blue points could identify suspicious points that, although were correct, were judged anomalous by a human analysis and could be errors in future observations.



Unsupervised learning

Why *supervised*?

- Large data scarcity for the minority class.
- As mentioned in the paper, for supervised learning *"no technique allows us to overcome the problem that because of the lower frequency of the minority class, we know little about the structure of the data of these observations"*.
- In these cases, it can be useful to use instead an *unsupervised* approach by learning the "underlying structure" of non-erroneous observations and classifying as anomalous the observations that deviate from that structure.

Oversampling and cross-validation

- Using oversampling *before* cross-validation (CV) is not a good idea[1].
- In order to correctly perform CV, all the steps involved in the building of the prediction model must be performed using only the training folds.
- In particular, it is important that oversampling is not performed on the entire dataset but only on the training folds.
- Otherwise, the procedure will produce unreliable and overoptimistic cross-validated estimates of the performance of the model (the problem is that minority observations that are similar only because of the oversampling procedure are included both when building the prediction model and when evaluating its performance).

[1] See for example *"Joint use of over-and under-sampling techniques and cross-validation for the development and assessment of prediction models"* (2015) and *"Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches"* (2018)

Focus of the paper

- Interesting paper that deals with a concrete problem, the prediction of reporting errors.
- Not sure why the focus is on a specific and already widely studied part, the rare events, while there is much more (data collection, feature engineering, model construction and evaluation).
- Perhaps focusing on the prediction of reporting errors instead of the very general rare events problem (that of course would still have plenty of room) could help to better frame the paper.