Predicting bank distress in the UK with machine learning

Joel Suss (Bank of England, London School of Economics)
Henry Treitel (Bank of England)

Summary:
- Using regulatory data, we compare classical statistical models with machine learning techniques for predicting bank distress.
- Implement rigorous, double-block randomisation CV procedure to account for hierarchical nature of data (intra-firm & quarter correlation).
- Random forest (RF) best based on AUC and Brier Score.
- RF also best when varying the relative cost of false negatives (missing actual cases of distress) & false positives (wrongly predicting distress) for discrete decision thresholds.
- Investigate drivers of bank distress using Shapley values and regression, and H-statistic (for interaction strength).
- Explore simple ensembling techniques to demonstrate additional performance benefits.
- Robustness checks: different time horizons (1, 2, 3 and 8 months), rolling forecast CV, omitting pre Q3 2009 data.

Data:
- Outcome measure: Subjective supervisory assessments of firm risk (UK Financial Services Authority ARROW Total Probability score).
- 1-10 score, 8 or above considered high-risk and labelled distressed – i.e. converted to binary.

Cross-validation procedure

Distribution of Arrow scores

Mean absolute Shapley values

Ensemble performance

Shapley regression coefficients (standardized & exponentiated)