Protocol (26.11.2025) –

Credit Risk Modeling ILAB OeNB

By Anastasia Pryshchepa, Leonid Kuzmishin,
Tristan Leiter, Martin Yago Tortajada

WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

OESTERREICHISCHE NATIONALBANK
EUROSYSTEM

Protocol (26.11.2025)

Summary

1) Balance sheet data *(slide 3)*

- Apply the filters:
  - a) $f_{10} \geq 0$
  - *b)* $(f_2 + f_3) \leq f_1$
  - *c)* $(f_4 + f_5) \leq f_3$
  - *d)* $(f_6 + f_{11}) \leq f_1$

  , where we allow for a rounding error of 2 decimal places).

2) Feature selection *(slides 11 – 14)*

- Add descriptive statistics (sectors, firm size, how many firms, total # of firms, etc.) – useful insights for stratified sampling.
- Add univariate regressions (AuC-RoC in the univariate case; again for each sector).
- Is the Informational Value useful for categorical variables with only binary encoding?
- Repeat the multicollinearity plot with the standardized/log-transformed features. Be careful with conclusions from linear correlations when we can likely assume that the dependency is non-linear.

➢ Don‘t remove any features. In ML we do not have to pursue a parsimonious model.

Protocol (26.11.2025)

Summary

3) Data sampling *(slide 15)*

- Make sure to incorporate out of sample and out of time (not the same problem). Do not use each firm in the training and test set, rather split on different firms (overfitting concerns with less generalization on new, unseen data).
- Preserve the default ratio across sectors and account for firm differences. Split by sectors as well.
- Think of the evolution over time. Take the firms with the full time dimension in each split.

4) Loss Measure *(slides 16 - 18)*

- Focus on ROC and AUC.
- AUCs per sector (multivariate case). Pursue robust models without leaving out any features (even with high linear dependencies).
- Read on proper scoring rules (Brier-score, statistical axioms).
- Incorporate the insights from papers (with a focus on proper scoring rules).

5) Outlook

- First focus on the plain positions. Think about reasonable standardization methods, scaling, log-transformation, using empirical cdfs, robust scaling. We could divide by the total assets or replace the raw balance-sheet data with the empirical quantiles from the cdf.
- Bayesian optimization is fine. Use all existing packages -> but always know what you are doing..
- AutoML is also fine to use.
- No upsampling, data is numerically stable.