Supplementary Information

Dependence-Aware Forecast Combination under Temporal Distribution Shift: European Insurer Solvency-Ratio Deterioration

Aleksander Wojtowicz, Ewa Rak, Rafał Rak, Anna Denkowska, Stanisław Wanat

S1 Additional data and model details

Table S1 reports the candidate variables and diagnostic series. Macro levels and the year-level Herfindahl index are used only as shift diagnostics. The eight modelling candidates are log assets, the current solvency ratio, capital intensity, the prior one-year solvency-ratio change, asset growth, and country-year changes in the policy rate, inflation and GDP growth.

Table S1: Descriptive statistics of the labelled panel (9,239 firm-years, forecast origins 2017–2023).

Variable	Mean	Std. Dev.	Min.	Median	Max.
Log assets	13.078	2.164	7.312	12.906	19.835
Solvency ratio (%)	268.300	165.320	25.310	218.000	1873.000
Capital intensity	0.192	0.845	0.000	0.165	66.986
SR change, $t-1$ to t (pp)	1.163	81.956	-1363.340	0.300	1333.420
Asset growth	1.031	34.742	-0.999	0.040	1645.891
Δ interest rate (pp)	0.689	1.114	-1.750	0.000	5.000
Δ inflation (pp)	0.784	3.084	-10.300	0.200	14.900
Δ GDP growth (pp)	-0.179	6.030	-13.800	-0.650	21.700
HHI (diagnostic only)	108.111	8.658	97.490	108.245	120.744

The main classifiers are logistic regression, random forest, XGBoost and a linear support-vector machine with Platt scaling. Table S2 gives the fixed settings and the checks conducted on the validation year. Scalers and winsorisation bounds use model-fit observations only. Forest and XGBoost results were checked across five seeds in Window 1; the reported seed is 42.

Table S2: Primary hyperparameter configurations and validation checks. The final column reports Window 1 AUC after the validation-year check.

Model	Primary settings	Validation check / Window 1 AUC
Logistic regression	$C = 1$, L2, balanced class weights, max. iter. 8000	$C = 1$ retained; AUC 0.727
Random forest	250 trees, depth 8, min. leaf 5, balanced weights	depth 6 after validation check; AUC 0.717
XGBoost	depth 4, 250 trees, lr 0.05, subsample 0.8, colsample 0.8	depth 3 after validation check; AUC 0.707
Linear SVM	$C = 1$, balanced weights, Platt (sigmoid, 3-fold)	Fixed comparator; AUC 0.727
RBF SVM (diagnostic)	$C = 1$, $\gamma =$ scale, Platt as above	Diagnostic only; AUC 0.622; excluded from fusion

S2 Feature-selection and model-composition checks

The number of retained features is selected from $k \in \{3, 4, 5, 6, 8\}$ by expanding-year inner validation and the one-standard-error rule. This gives $k = 3$, although all eight candidates have the highest inner mean AUC (0.748). The three-feature specification contains solvency ratio, sr. ratio change ($t - 1$ to t) and δ inflation. Figure S1 reports the selector inclusion counts. The final feature list is stated directly because the inclusion counts alone do not describe the cut-off tie handling.

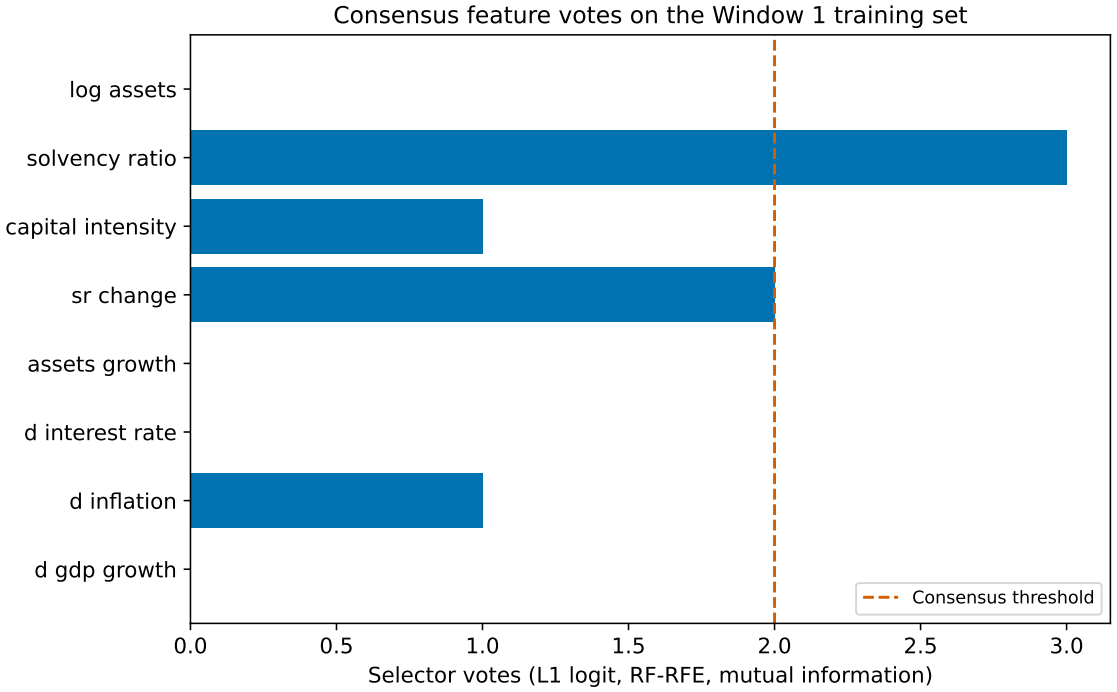


Figure S1: Selector inclusion counts on the Window 1 model-fit sample.

Feature-selection stability is separate from source-model contribution. The current solvency ratio and capital intensity are selected with frequencies 1.00 and 0.80. Mean Jaccard overlap is 0.569 compared with 0.744 under permuted labels. The higher overlap under permutation can arise from deterministic tie handling or fallback rules. These summaries do not establish informative stability. Figure S2 shows the bootstrap selection frequencies and the corresponding stability summary.

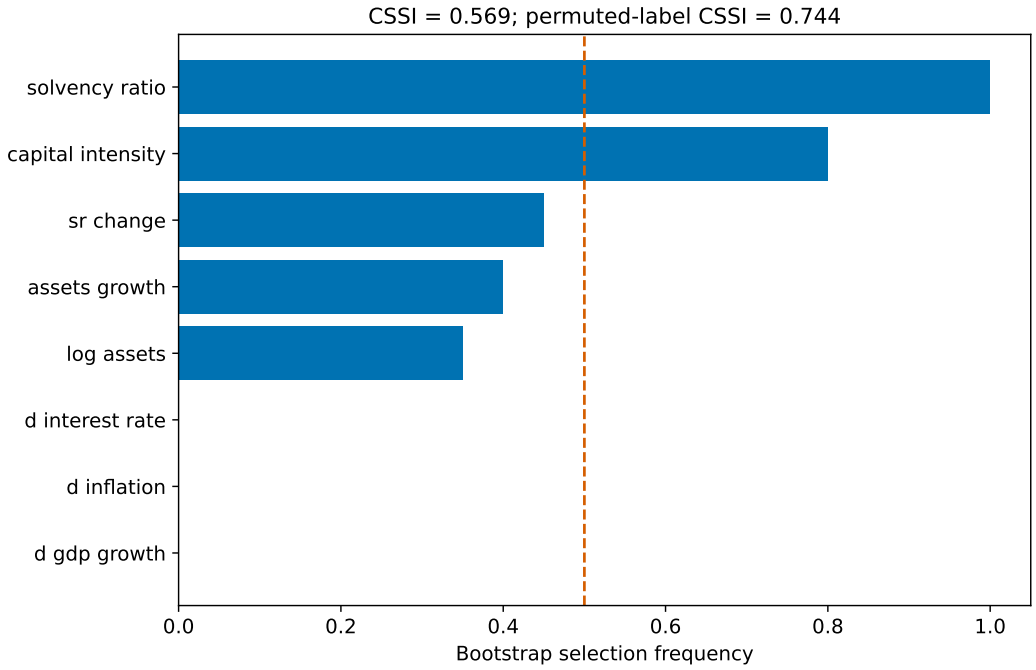


Figure S2: Bootstrap selection frequencies and the corresponding stability summary.

With all eight candidates, Window 1 Dempster AUC is 0.735 [0.715, 0.757], compared with forest AUC 0.728 [0.708, 0.748]. The clustered increment over the best single model is 0.008 [0.001, 0.014]. This result does not change the primary specification because the one-standard-error rule selected three features. Adding HHI to the selected features gives Dempster AUC 0.715 and logistic AUC 0.727.

The linear SVM reaches Window 1 AUC 0.727 after scaling and Platt calibration. The RBF SVM reaches 0.622 and is excluded from fusion. Including the linear SVM changes Dempster AUC from 0.718 to 0.720. A validation check selects forest depth 6 and boosting depth 3; their Window 1 AUC values are 0.717 and 0.707. Logistic $C = 1$ is retained.

S3 Additional classifier audit

Fixed-configuration CatBoost, LightGBM, Extra Trees and histogram gradient boosting were fitted on both temporal windows. They were evaluated with the selected three features and all eight candidates. Tables S3 and S4 report the results for Windows 1 and 2, respectively. These are secondary checks with fixed settings rather than a matched tuning competition.

Table S3: Additional tree models on Window 1. Values are AUC / PR-AUC.

Feature set	CatBoost	LightGBM	Extra Trees	HistGBDT
Selected (3)	0.701/ 0.385	0.686/ 0.379	0.670/ 0.401	0.704/ 0.390
All candidates (8)	0.721/ 0.403	0.723/ 0.417	0.726/ 0.429	0.717/ 0.409

Table S4: Additional tree models on Window 2. Values are AUC / PR-AUC.

Feature set	CatBoost	LightGBM	Extra Trees	HistGBDT
Selected (3)	0.671/ 0.376	0.654/ 0.347	0.690/ 0.386	0.665/ 0.360
All candidates (8)	0.681/ 0.380	0.665/ 0.363	0.687/ 0.382	0.680/ 0.375

S4 Conditional statistical diagnostics

Out-of-time AUC intervals and pairwise differences use 1,000 bootstrap resamples of undertakings. Each draw retains all available observations for a sampled firm. The Giacomini–White and Clark–West statistics are reported as conditional firm-year diagnostics. They do not measure uncertainty across future calendar regimes because the main test contains only two forecast origins. The paired firm bootstrap is therefore the main uncertainty summary.

An unweighted logit on the selected features has Window 1 AUC 0.723 and Brier score 0.164, compared with 0.165 for the ratio-only probability model. The Clark–West diagnostic is $t = 5.40$ ($p < 0.001$). For raw Brier loss, the Giacomini–White diagnostic for Dempster and logistic regression is $t = -1.41$ ($p = 0.16$); for soft voting and logistic regression it is $t = -2.89$ ($p = 0.004$). These values are conditional on the observed firms and years.

In firm-clustered bootstrap samples, balanced logistic regression has the highest AUC in 63.4% of draws, the current ratio in 12.1%, the linear SVM in 20.7%, and Dempster fusion in 2.9%. Figure S3 shows the share of bootstrap samples in which each score has the highest AUC.

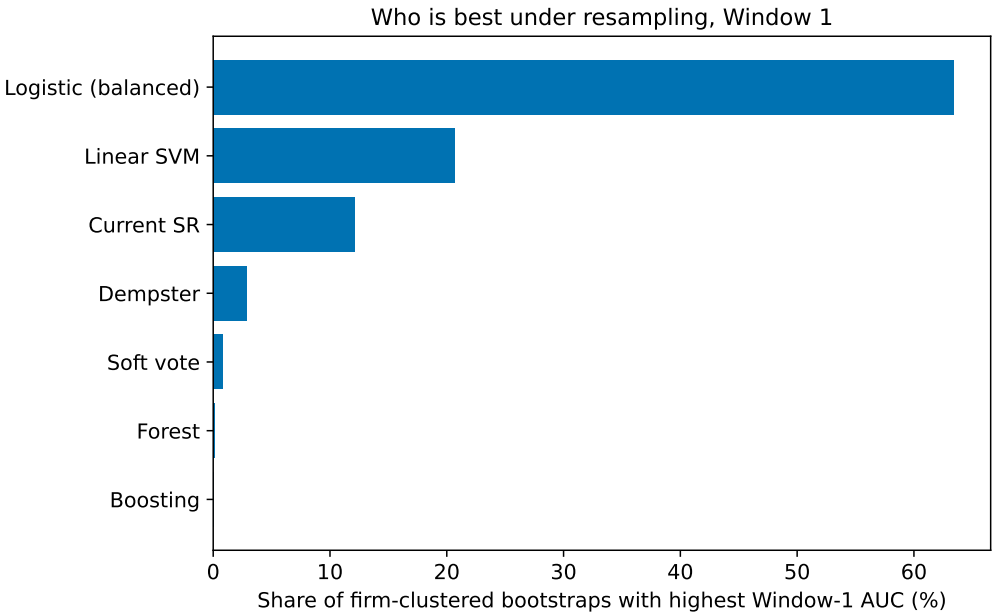


Figure S3: Share of firm-clustered Window 1 bootstrap samples in which each score has the highest AUC.

The model confidence set uses sequential max-t elimination with firm-block bootstrap rather than the full Hansen range statistic. A 10% model confidence set based on isotonic-calibrated Brier losses retains 3 members: logistic regression, current-ratio probability, linear SVM. Dempster fusion, soft voting, random forest and boosting are eliminated. This result is also conditional on the observed calendar origins.

S5 Synthetic dependence experiment

The simulation varies source-error correlation, prevalence shift and calibration scale. At correlations of at least 0.9, Dempster's mean Brier excess over the best source is 0.026, and Dempster is harmful in 100% of repetitions averaged over the tested drift and calibration cells. Table S5 reports the mean Brier difference from the best single source. A negative value favours the pool; a positive value favours the best source.

Table S5: Mean Brier difference from the best single source in the synthetic experiment.

Source-error correlation ρ	0	0.5	0.9	0.99
Equal pool	-0.0120	-0.0057	-0.0007	0.0001
Dempster rule	0.0117	0.0196	0.0253	0.0264

The benefit of equal pooling decreases as dependence rises, while Dempster's Brier excess increases. This experiment illustrates the mechanism in controlled synthetic data. It does not estimate the effect of dependence in the insurer panel.

S6 Rolling-origin and continuous-outcome checks

The expanding-origin evaluation fits models through $t-2$, uses $t-1$ for validation and tests on year t. Table S6 includes the current-ratio rank. Figure S4 shows the full classifier comparison, and Figure S5 adds the univariate rank.

Table S6: Expanding-window rolling-origin AUC.

Test origin t	Current SR	Logit	Forest	Dempster
2019	0.752	0.707	0.735	0.733
2020	0.729	0.728	0.707	0.718
2021	0.662	0.664	0.686	0.674
2022	0.725	0.726	0.733	0.734
2023	0.724	0.722	0.719	0.724
Mean	0.718	0.709	0.716	0.717

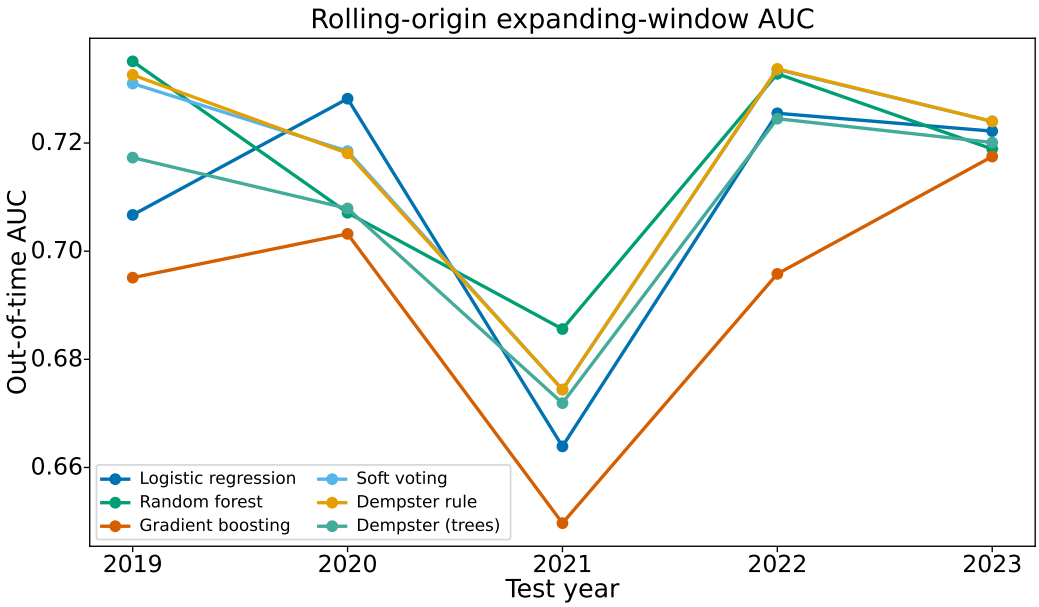


Figure S4: Expanding-window AUC for the fitted classifiers and combination rules.

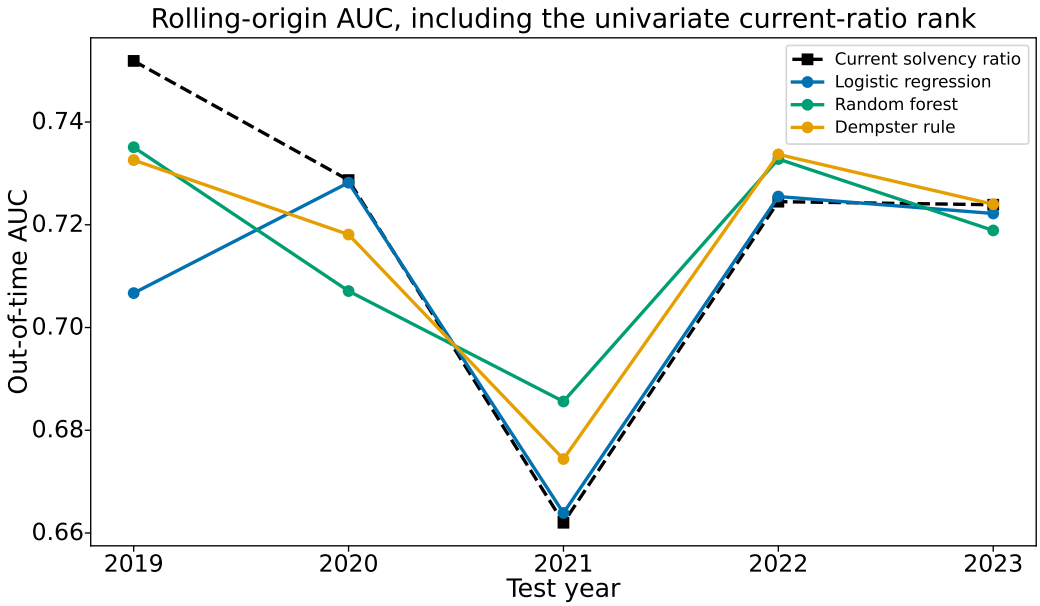


Figure S5: Expanding-window AUC including the univariate current-ratio rank.

Forecasting the continuous next-year solvency-ratio change gives RMSE 85.2 for the full ridge model and 90.0 for the forest, compared with 87.4 for a zero-change forecast and 87.5 for training-mean climatology. A univariate ridge on the current solvency ratio has RMSE 84.1 and MAE 37.8, the best values in this comparison. Ranking firms by its predicted change gives AUC 0.724. The full ridge and forest give 0.644 and 0.617.

S7 Alternative labels and winsorisation

Table S7 reports alternative event definitions on Window 1. A 20 percentage-point drop that also ends below 200% or 150% is a different event from the primary unrestricted deterioration label.

Table S7: Label sensitivity on Window 1. Δ is the best fusion rule minus the best single model.

Event	Test prevalence	Logit AUC	Best single AUC	Δ fusion
Absolute ≤ -10 pp	34.9	0.661	0.661	-0.003
Absolute ≤ -20 pp (primary)	23.4	0.727	0.727	-0.009
Absolute ≤ -30 pp	15.9	0.775	0.775	-0.002
Relative $\leq -10\%$	19.0	0.624	0.646	-0.002
Relative $\leq -20\%$	7.0	0.648	0.688	-0.012
20 pp and $R_{t+1} < 200$	8.1	0.593	0.675	0.004
20 pp and $R_{t+1} < 150$	2.3	0.602	0.679	-0.015

Winsorising each input at the 1st and 99th percentiles estimated on model-fit observations gives Window 1 AUC 0.727 for logistic regression, 0.709 for the forest and 0.714 for Dempster fusion. The main conclusion is unchanged.

S8 Calibration, decision utility and the worked score

Figure S6 shows the Window 1 reliability curves. Figure S7 reports decision-curve net benefit. These are descriptive comparisons at the observed test origins. Table S8 gives selected net-benefit values.

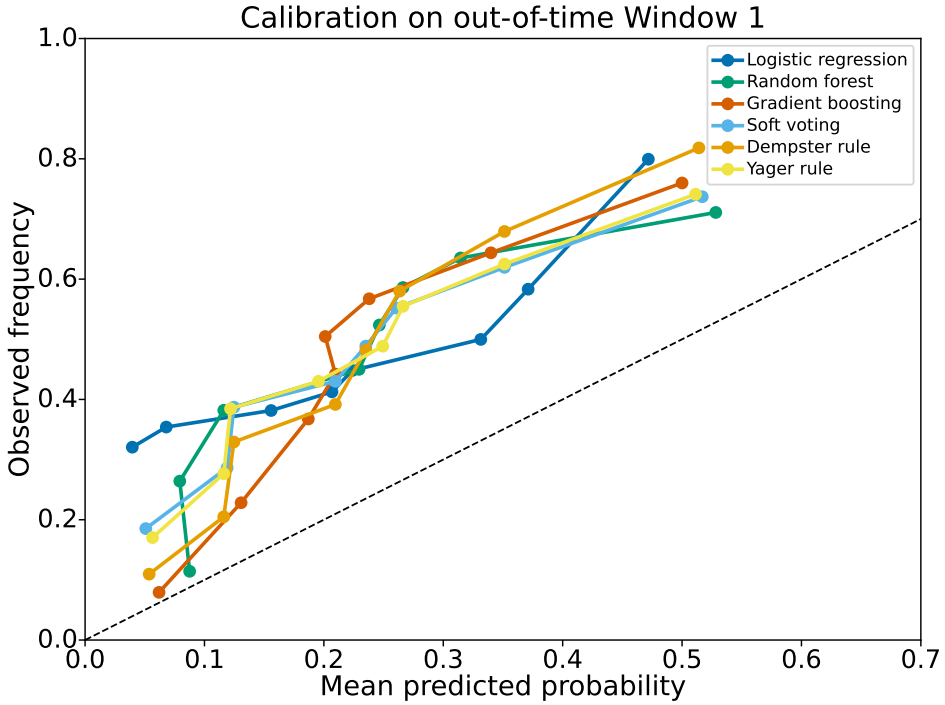


Figure S6: Reliability diagram for Window 1.

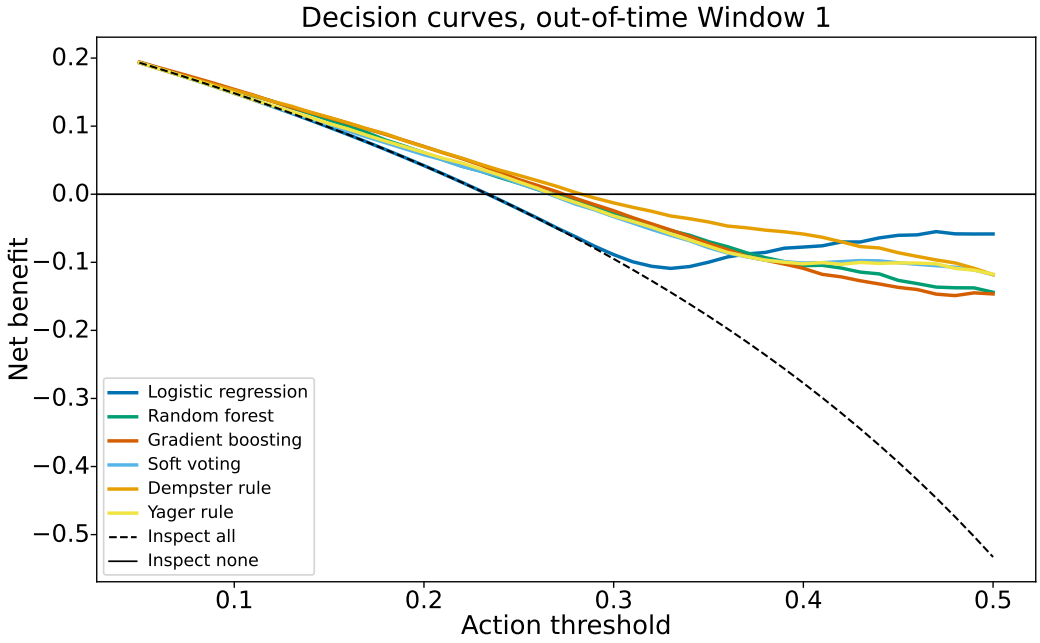


Figure S7: Decision curves for out-of-time Window 1.

Table S8: Selected net-benefit values on Window 1.

Method	$\tau = 0.05$	$\tau = 0.10$	$\tau = 0.20$
Logistic regression	0.193	0.148	0.042
Random forest	0.193	0.152	0.061
Dempster rule	0.193	0.153	0.070

The Window 1 logistic coefficients use standardised selected features and firm-clustered standard errors. Undertaking A in predictor year 2022 (Life) receives predicted probability 0.998. The observed event is a 20 percentage-point solvency-ratio drop. It is not an insolvency finding or an SCR breach. Table S9 reports the coefficients, clustered standard errors and odds ratios.

Table S9: Window 1 logistic coefficients on standardised selected features.

Feature	β	Clustered SE	Odds ratio
Intercept	-0.056	0.029	n/a
Solvency ratio	0.788	0.056	2.200
Sr. ratio change ($t-1$ to t)	0.054	0.036	1.055
Δ inflation	-0.013	0.029	0.988