ACCURIA

28 February 2025

Theodoros Gavriilidis
Data Scientist Accuria Ltd.
theodore.gavriilidis@accuria.com

Re-Examining Recovery Rate Forecasting with Machine Learning

Executive Summary

Two years ago, our research on recovery rate prediction demonstrated that integrating machine learning techniques—such as neural networks and ensemble methods—with rule-based models such as Cubist, significantly outperformed traditional regression models (Gavriilidis and Heppe 2023). Moreover, while servicer-provided recovery expectations (Business Plans, BP) proved less reliable as stand-alone predictors, their inclusion as model features substantially enhanced predictive performance. These findings highlighted the benefits of advanced modelling techniques, the value of combining industry expertise with data-driven approaches, and emphasised the importance of routinely updating existing BPs.

Building on this foundation, our latest research—conducted in collaboration with Job Reijns, a master's student at Erasmus University Rotterdam—expands our dataset, refines our methodologies, and introduces a new suite of models, including Random Forests and XGBoost. By incorporating engineered features, we observe improved predictive accuracy in short-term forecasts, although long-term forecasts remain more challenging. These results not only confirm our previous performance gains with machine learning models but also illuminate the trade-offs between enhanced accuracy, increased model complexity, and reduced interpretability.

The improved short-term predictive accuracy offers direct business benefits by enabling financial institutions to:

- Enhance loan pricing accuracy and reduce the risk of misvaluation,
- Optimise debt collection strategies through more precise recovery forecasts, and
- Strengthen investor confidence in NPL securitisations with transparent, data-driven insights.

This article is based on the results presented in Reijns (2024) master thesis; we encourage readers to consult the complete paper for a more comprehensive analysis.

Table of Contents

Methodology

Expanding upon our prior research, we evaluate the performance of previously untested models—Random Forest and XGBoost—while reassessing Cubist (Gavriilidis and Heppe 2023). Concurrently, we introduce feature engineering by incorporating autoregressive (AR) and dynamic variables such as average annual recovery rate since default and standard deviation of recovery rate since default. All models are trained and evaluated within a standardised forecasting framework for both one-year and three-year recovery predictions and are benchmarked against traditional regression models namely Ordinary Least Squares (OLS) and industry-standard BPs.

i. Data Preparation

This year's research leverages an expanded dataset comprising over 860,000 loans—substantially larger than our previous sample. The dataset includes both static attributes (such as loan characteristics, collateral details, and original and updated BPs) and engineered features derived from post-default behaviours. The loans span multiple vintages, with defaults dating back over twenty years.

Consistent with our previous work, the data were aggregated at the borrower level to reflect the structure of recorded collections. Following best practices from the literature (Anthony Bellotti et al. 2021), we excluded observations with implied recovery rates outside the plausible range of [-25, 125] percent, calculated as a percentage of Gross Book Value (GBV), due to the unavailability of Exposure at Default (EAD) for most claims. Moreover, the dataset was organised as panel data, incorporating realised recoveries, claims, and BP projections as time-series variables.

ii. Data Exploration

The final dataset comprises 816,000 unique counterparties with a total of 860,000 loans—56% of which are corporate and 35% are secured—with an average vintage of 10.5 years. The observed mean recovery rate is 12%, substantially lower than the projected lifetime recovery rates of 28.83% and 33.45% outlined in the original and updated business plans, respectively. Notably, approximately 83% of the loans remain open, indicating that recoveries are still in progress and subject to right-censoring. Additionally, many loans are affected by left-censoring, as their recovery trajectories are only partially observed, particularly for older vintages where the complete post-default period is not recorded.

The distribution of borrower-level recovery rates is bi-modal, with distinct concentrations near 0 and 1. Further stratification by case status and vintage underscores significant differences between open and closed cases. As illustrated in Figure 1, open cases—especially those from recent vintages—tend to exhibit lower observed recovery rates, likely due to the limited time available for full resolution. In contrast, Figure 2 shows that closed cases consistently achieve higher recovery rates, with the majority either reaching full recovery or settling at intermediary levels, as evidenced by a secondary mode around 50%. Collectively, these findings highlight the need to integrate both the temporal dynamics of recovery—accounting for censoring—and the resolution status of cases into any modelling framework for recovery rates.

iii. Feature Splitting

This research evaluates model performance based on three groups of features: static, autoregressive, and dynamic. Static features capture loan, borrower, and contract-specific characteristics recorded at the most recent cut-off date. These attributes remain constant over time and serve as the foundation for all models. In contrast, autoregressive features incorporate lagged recovery rates, linking past performance to current forecasts. Dynamic features, meanwhile, evolve over time to capture trends in loan performance.

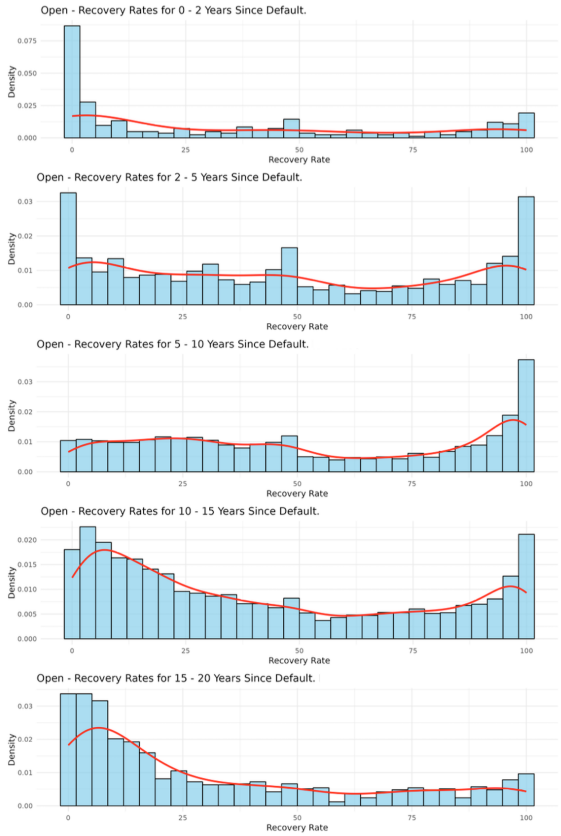

Figure 1: Recovery rate distribution in open cases segmented based on Years Since Default (Reijns 2024)

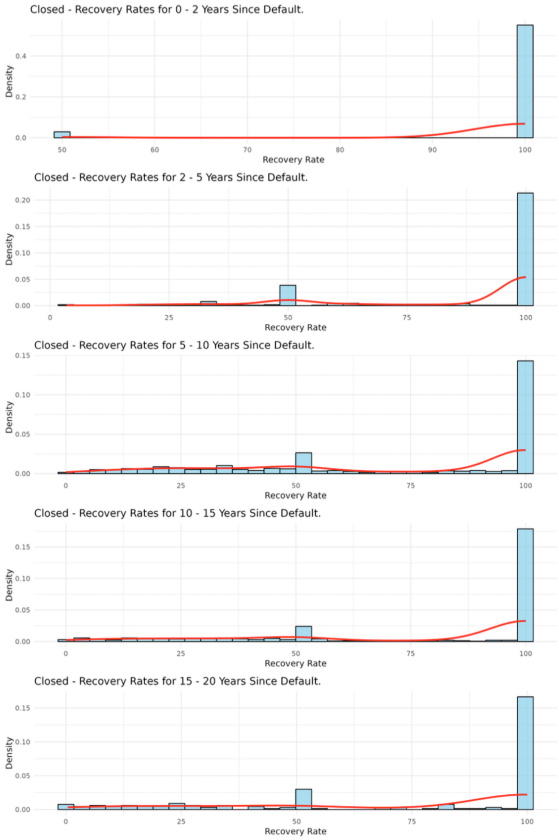

Figure 2: Recovery rate distribution in closed cases segmented based on Years Since Default (Reijns 2024)

Static features include categorical and numerical variables that describe fundamental loan attributes, such as:

- Loan characteristics: Gross Book Value (GBV) bucket, secured status, corporate borrower indicator.
- Property information: Region, property type, loan-to-value (LTV) ratio.
- Loan history: Default period, lifetime recovery rate from the original BP.
- Portfolio structure: Count of associated properties, guarantees, or loans.

Dynamic features track the evolution of recovery rates and loan performance over time, allowing models to differentiate between different observation periods for the same loan. These include:

- Recovery trends: Lagged, mean, and standard deviation of recovery rates since default.
- Time-related metrics: Years since the last recovery.
- Current status: Accumulative recovery rate and BP for the current period.

To assess the incremental value of temporal information, the study defines three feature sets.

1. **Static Feature Set** – Consists solely of static features.
2. **AR(1) Feature Set** – Expands the static set by including a lagged recovery rate.
3. **Dynamic Feature Set** – The most comprehensive set, including all static and dynamic features.

Comparing model performance across these feature sets enables us to clearly evaluate the benefits of integrating dynamic, time-evolving information into forecasting models.

Performance Evaluation

The evaluation of model performance primarily relies on the out-of-sample root mean square error (RMSE) and the out-of-sample R-squared (R^2_{OOS}), which together measure the accuracy of the forecasts relative to a benchmark, in this case, the true mean of the test observations (Anthony Bellotti et al. 2021; Hawinkel, Waegeman, and Maere 2023). RMSE quantifies predictive accuracy, with lower values indicating better performance, while R^2_{OOS} measures the proportion of variance explained by the model, with higher values signifying greater explanatory power.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad R^2_{\text{OOS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

i. One-Year-Ahead

We begin by evaluating the one-year-ahead forecasting performance of our models, aggregating forecasts across all years rather than distinguishing among specific years.

Figure 3 displays the out-of-sample RMSE values for all tested models and feature sets. Random Forest (RF) and XGBoost (XGB) models achieve the lowest RMSE values, indicating superior predictive accuracy. However, as detailed in Reijns (2024), when investigating further RF exhibits clear signs of overfitting, rendering XGB the more reliable option.

Figure 4 further supports these findings as XGB consistently attains a positive R^2_{OOS}, demonstrating its effectiveness in capturing underlying patterns. In contrast, the OLS model struggles to capture the linear relationships inherent in the data, resulting in near-zero R^2_{OOS}. These results are in line with previous studies (Gavriilidis and Heppe 2023; A. Bellotti, Gambetti, and Vrins 2019; Baranauskaite, Gambetti, and Vrins 2020).

Overall, both Figure 3 and 4 provide compelling evidence for the benefits of incorporating autoregressive and dynamic features. While all models show improvements with these features, the enhancements are most pronounced for Cubist, whereas OLS exhibits the least improvement.

ii. Three-Years-Ahead

Building upon our one-year-ahead results, we extend our evaluation to generate two- and three-year-ahead predictions using a recursive forecasting approach. In this method, the forecast for period $t+1$ is used as an input to predict period $t+2$, and so on, ensuring consistency across multi-year forecasts.

Our analysis covers a three-year period from 2021 to 2023. Due to data constraints, extending the forecasting horizon further would substantially reduce the number of observations, potentially compromising the robustness of the results.

Table 1 summarise the performance for all three feature sets. Interestingly, our analysis reveals that multi-year forecasts perform best when relying solely on the static feature set, regardless of the model used. This is evidenced by a lower RMSE and improved R^2_{OOS} compared to forecasts generated using autoregressive features. We attribute this outcome to the accumulation of errors introduced by recursive forecasting when autoregressive features are included, as observed in both our AR(1) and Dynamic feature sets.

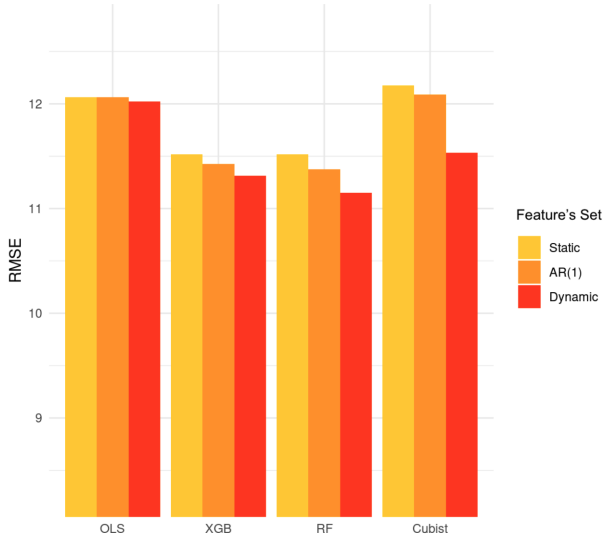

Figure 3: Comparison of out-of-sample RMSE for one year ahead predictions across OLS, XGB, RF, and Cubist models (Reijns 2024)

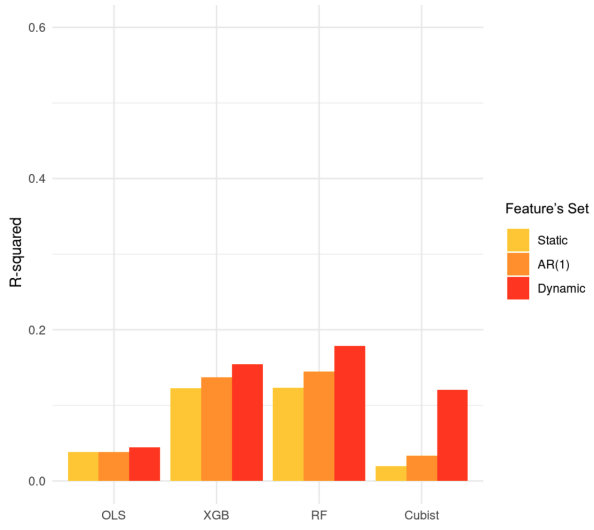

Figure 4: Comparison of out-of-sample R^2_{OOS} for one year ahead predictions across OLS, XGB, RF, and Cubist models (Reijns 2024)

Table 1: Performance Results Across Feature Sets

Feature Set	Metric	OLS	XGBoost	Random Forest	Cubist	BP
Static	RMSE	20.317	17.251	17.232	20.926	24.972
	R^2_{OOS}	0.112	0.360	0.361	0.058	-0.341
Static + Lagged	RMSE	22.200	19.054	21.442	22.058	24.972
	R^2_{OOS}	-0.060	0.219	0.011	-0.046	-0.341
Static + Dynamic	RMSE	20.318	18.740	18.959	20.184	24.972
	R^2_{OOS}	0.112	0.244	0.227	0.123	-0.341

Figure 5 segments the out-of-sample RMSE by default period. The results reveal substantial heterogeneity in predictive accuracy: loans that defaulted prior to 2007 exhibit significantly lower RMSE values compared to those defaulting in more recent periods (e.g., 2016–2017). This discrepancy likely reflects the tendency of older loans to reach a stable or "stale" state, where recoveries remain relatively unchanged, whereas more recent defaults are characterised by greater uncertainty and volatility in recovery rates.

Critical Analysis

i. Performance

This year's research, leveraging an expanded dataset and additional models, demonstrates that XGBoost (XGB) and Random Forest (RF) better capture the variance of recovery rates, as indicated by higher R^2_{OOS} values. However, this improvement comes at the cost of lower accuracy in predicting the mean recovery rate, as reflected by higher out-of-sample RMSE values.

Although direct comparisons are limited due to differences in the datasets between the two studies, comparing

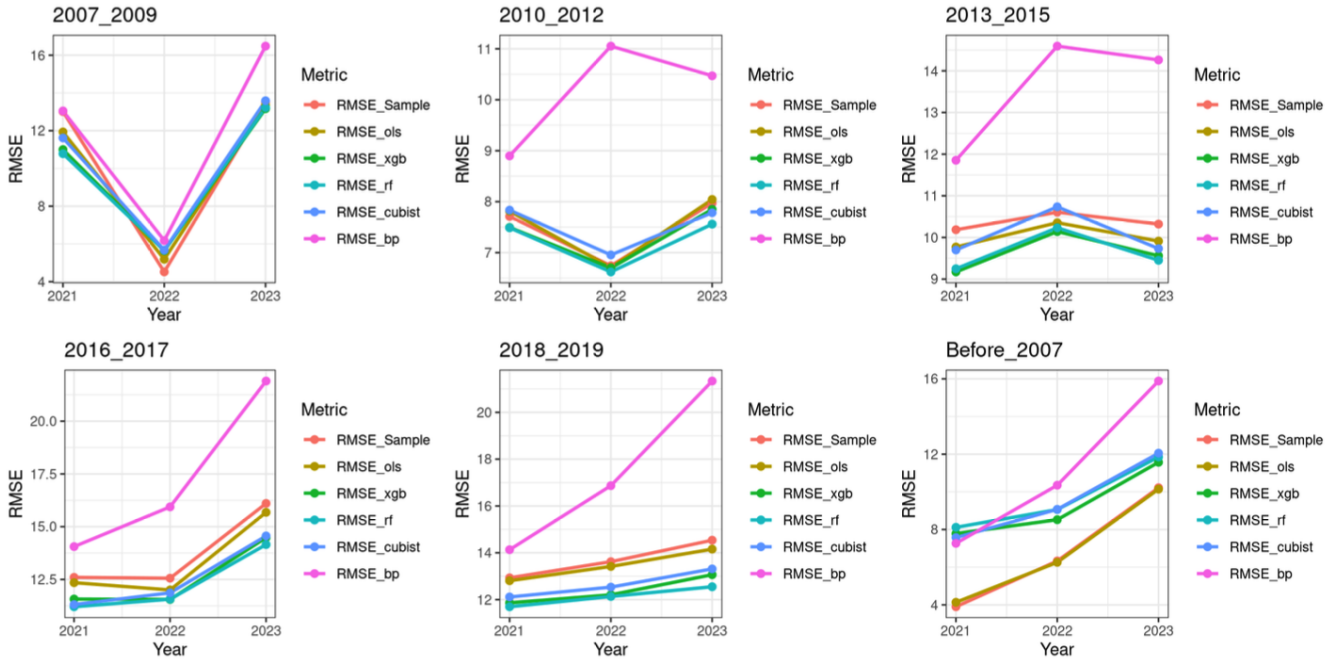

Figure 5: Out-of-sample RMSE comparison across OLS, XGB, RF, and Cubist models, split by different default periods. (Reijns 2024)

the best scores from each provides an indication of the most effective models and approaches tested so far. For example, in three-year-ahead predictions, XGBoost achieved an out-of-sample R^2 of up to 0.36—a significant improvement over our last research's best models, which reached only 0.14. In contrast, the RMSE increased substantially from 7.3 in our previous research to 17.25 in the current study.

Consistent with our previous findings, machine learning models—especially non-linear approaches such as neural networks, ensemble methods, and, in this study, XGBoost and Random Forest—continue to show clear improvements in predictive performance when compared to simple regression models or industry BP. Nonetheless, forecasting multi-year recoveries remains challenging due to the accumulation of errors in recursive predictions, which limits long-term accuracy.

ii. Business Plan as Stand-alone Predictors

Consistent with our previous findings, BP were consistently outperformed by machine learning models at the loan level, reinforcing the importance of regularly updating BP. Despite this, BP remains a critical predictive component, serving as one of the most influential features in our models[1]. These results highlight the value of information captured during the debt collection process, which servicers are able to incorporate into their forecasts to enhance predictive accuracy.

iii. Recovery Rate Segmentation

Building on our previous research, this year we examine first-order effects by analysing how different loan segments influence both realised and expected recovery rates. Our findings reveal significant heterogeneity in re-

[1]Please refer to Reijns (2024), for a more detailed analysis on variable importance.

covery rates across distinct loan cohorts.

Older loans, often inactive for over a decade, exhibit predictable recovery trajectories, with many unlikely to generate future cash flows. In contrast, loans that defaulted more recently show greater variability, reflecting the ongoing recovery process and the inherent uncertainties associated with it. These differences highlight the importance of tailoring prediction models to capture the diverse characteristics of NPL portfolios, ultimately enhancing the accuracy and effectiveness of segment-specific forecasts.

iv. Features Engineering

Expanding our previous research, this year, we explored how engineered features enhance the forecasting capabilities of our models. Our findings confirm that these features improve short-term prediction accuracy. However, their impact diminished in multi-year forecasts due to accumulated prediction errors, which reduced the models' overall performance.

Compared to our last research, our analysis underscores both the benefits of dynamic features in boosting accuracy and the challenges posed by error accumulation in recursive predictions.

Conclusion

Our findings reaffirm the transformative role of machine learning in predicting NPL recovery rates. By leveraging larger datasets and advanced highly non-linear modelling approaches, financial institutions can improve loan pricing accuracy—thereby reducing the risk of miss-valuation—optimise debt collection strategies through more accurate recovery forecasts, and enhance investor confidence in NPL securitisations with transparent, data-driven insights.

Moreover, our results underscore the complementary nature of machine learning and traditional industry expertise. While servicer-provided business plans (BP) remain integral to recovery forecasting, combining them with predictive analytics yields more accurate and actionable insights. Recognising that modelling incorporating BP is a luxury that may not be feasible for all investors, further research on forecasting approaches that do not rely on BP is a priority.

Our advancements in feature engineering and dataset expansion provide evidence of improved short-term predictive accuracy. However, challenges persist in multi-year forecasts due to the accumulation of errors inherent in recursive predictions, which diminishes the effectiveness of dynamic features over longer horizons.

Looking ahead, our research will focus on expanding the dataset to extend the available time series, enabling a more in-depth exploration of the dynamics between open and closed cases, as well as the impact of optimal workout strategy resolution. Moreover, we plan to delve further into variable importance measures and grouping criteria to pinpoint the primary drivers of recovery outcomes. Ultimately, our goal is to empower financial institutions with robust, transparent, and actionable tools for managing and accurately valuing NPL portfolios, thereby fostering a more efficient and resilient market for non-performing loans.

References

Baranauskaite, I., P. Gambetti, and F. Vrins (2020). *Forecasting partial recovery rates for non-performing loans*. URL: https://epublications.vu.lt/object/%20elaba:81591438/.

Bellotti, A., P. Gambetti, and F. Vrins (2019). *Forecasting recovery rates on non-performing loans with machine learning*.

Bellotti, Anthony et al. (2021). "Forecasting recovery rates on non-performing loans with Machine Learning". In: *International Journal of Forecasting*. DOI: 10.1016/j.ijforecast.2020.06.009.

Breiman, L. (2001). *Random forests 45(1), 5–32*. DOI: 10.1023/A:1010933404324. URL: Retrieved%20from%20https://%20doi.org/10.1023/A:1010933404324.

Chen, T. and C. Guestrin (2016). "Xgboost: A scalable tree boosting system." In: *S In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785–794)*. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

Gavriilidis, Theodoros and Burkhard Heppe (2023). "Enhancing recovery rate predictions with ma- chine learning." In: *Accuria.com*. URL: https://accuria.com/enhancing-recovery-rate-predictions-with-machine-learning/.

Hawinkel, Stijn, Willem Waegeman, and Steven Maere (2023). "Out-of-Sample R2: Estimation and Inference". In: *The American Statistician*. DOI: 10.1080/00031305.2023.2216252. URL: https://doi.org/10.1080%2F00031305.2023.2216252.

Quinlan, J Ross (1992). "Learning with continuous classes". In: *Proceedings of the 5th Australian joint conference on artificial intelligence*.

— (1993). "Combining instance-based and model-based learning". In: *Proceedings of the Tenth International Conference on International Conference on Machine Learning*.

Reijns, Job (2024). *Forecasting Recovery Rates of Non-Performing Loans*. URL: https://thesis.eur.nl/.

Appendix

A.1 Random Forest

Random Forest (RF), introduced by Breiman (2001), is an ensemble learning technique that constructs multiple decision trees and aggregates their predictions to improve robustness and reduce overfitting. Each tree is trained on a bootstrap sample of the dataset, and at each split, a random subset of features is selected.

For a regression task, the final prediction $\hat{y}$ for an input x is the average of individual tree predictions:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} h_m(x), \tag{1}$$

where $h_m(x)$ is the prediction of the $m-th$ tree, and M is the total number of trees.

Random Forest reduces variance through averaging. The variance of a single decision tree is Var(h(x)) , and for an ensemble with inter-tree correlation ρ , the variance is:

$$\text{Var}(\hat{y}) = \frac{1}{M}[\text{Var}(h(x)) + (M-1)\rho \cdot \text{Var}(h(x))]. \tag{2}$$

As $M \to \infty$, assuming low correlation ρ , the ensemble variance diminishes, making predictions more stable. This makes Random Forest robust to overfitting while maintaining high predictive accuracy.

A.2 XGBoost

XGBoost, short for Extreme Gradient Boosting, is an optimised implementation of the gradient boosting algorithm, designed for efficiency and scalability (Chen and Guestrin 2016). Like traditional gradient boosting, XGBoost builds trees sequentially, where each new tree corrects the errors of the previous ones. However, it incorporates regularisation and second-order optimisation to enhance accuracy while maintaining computational efficiency. At each iteration t , XGBoost minimises a regularised objective function:

$$L^{(t)}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{K} \Omega(f_k), \tag{3}$$

where $l(y_i, \hat{y}_i^{(t)})$ is the loss function (e.g., squared loss for regression, logistic loss for classification), and $\Omega(f_k)$ is a regularisation term penalising tree complexity:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2. \tag{4}$$

Here, T is the number of leaf nodes, λ controls L2 regularisation, and γ penalises excessive tree growth, reducing overfitting. Predictions are updated iteratively using a learning rate η :

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta f_t(x_i). \tag{5}$$

To optimise the loss function efficiently, XGBoost employs a second-order Taylor expansion:

$$L^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t), \tag{6}$$

where g_i and h_i are the first- and second-order gradients of the loss function with respect to the previous predictions. This enables XGBoost to make fast, precise updates while controlling model complexity.

A.3 Cubist

Cubist, introduced by Quinlan (1993) as an extension of the M5 model tree (Quinlan 1992), is a rule-based model that integrates regression trees with linear regression at each node. The model constructs a tree by iteratively splitting the data to minimise variance, where each terminal node T(x) contains a linear model of the form:

$$\hat{y}_{T(x)} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \tag{7}$$

where β_0 and β_j are fitted using ordinary least squares (OLS) on observations satisfying T(x). To prevent overfitting, Cubist applies recursive smoothing, blending the prediction of each node with its parent's prediction:

$$\hat{y}(x) = \frac{n_T \hat{y}_{T(x)} + l \hat{y}_P(x)}{n_T + l} \tag{8}$$

where n_T is the number of observations in $T(x)$, $\hat{y}_P(x)$ is the parent node's prediction, and l is a smoothing constant (typically 15). Additionally, Cubist employs instance-based correction, adjusting predictions using similarity-weighted contributions from k nearest training instances:

$$\hat{y}(x) = \hat{y}T(x) + \sum_{i=1}^{k} w(x_i)(\hat{y}T(x) - \hat{y}T(x_i)) \tag{9}$$

where the weight function $w(x_i)$ assigns higher importance to closer instances based on Euclidean distance. Finally, Cubist supports committees, an ensemble approach where multiple model trees are trained sequentially in a boosting-like framework to reduce residual errors. The number of committees and nearest neighbours is optimised via K-fold cross-validation.

About Accuria

Accuria is a next-generation, asset-agnostic credit technology platform that transforms how financial institutions and investors manage, analyse, and trade credit assets. Evolving from our roots in distressed asset management, our rebranding to Accuria, from NPL Markets, marks a strategic shift to a unified digital ecosystem that supports the entire credit spectrum—from traditional financial instruments and portfolios to debt funds, securitisations, ABCP conduits, and covered bonds.

Our platform seamlessly integrates advanced data management, AI-driven analytics, and a dynamic digital marketplace to convert complex, loan-level data into actionable insights. With solutions that elevate valuations, optimise risk monitoring, and streamline transaction execution, Accuria empowers users to make smarter, faster decisions. Our comprehensive suite of products—including innovative tools like DataHub—enables seamless data integration and delivers comprehensive analytics to support your strategic objectives.

Trusted across 28 countries and managing over €92 billion in assets, Accuria is committed to innovation, accuracy, and a client-first approach. While our identity has evolved, our dedication to providing cutting-edge tools and expert support remains unchanged—ensuring users stay ahead in an increasingly complex financial landscape.

Disclaimer

This paper contains confidential information about Accuria, current at the date hereof. This presentation is not intended to provide the sole basis for evaluating Accuria and should not be considered as a recommendation with respect to it or any other matter. This document and the information contained herein are not an offer of securities for sale in the United States and are not for publication or distribution to persons in the United States (within the meaning of Regulation S under the United States Securities Act of 1933, as amended). This presentation and the information contained herein does not constitute or form part of any (i) offer or invitation or inducement to sell or issue, or any solicitation of any offer to purchase or subscribe for, any securities or (ii) invitation or inducement to engage in investment activity within the meaning of Section 21 of the United Kingdom Financial Services and Markets Act 2000, as amended, nor shall any part of this presentation nor the fact of its distribution form part of or be relied on in connection with any contract or investment decision relating thereto.