
RESEARCH ARTICLE

Machine Learning for Business Process Risk Prediction and Operational Performance Optimization

Rahat Hasan¹✉, Md Eamam Hasan², Tanvir Ahammad³, Rahmatur Rab Mohammed Noman⁴, Md Nazmul Hasan Hasib⁵, Md. Evrahim Kholil Ullah⁶ and Mahfujul Karim Sejan⁷

¹Information Technology and Project Management, Westcliff University, CA, USA; <https://orcid.org/0009-0003-2783-1832>

²Department of Business Administration, International American University, CA, USA; <https://orcid.org/0009-0002-6310-5910>

³Department of Business Administration, University of the Potomac, Washington, DC, USA; <https://orcid.org/0009-0004-1489-9552>

⁴Department of Business Administration, University of the Potomac, Washington, DC, USA; <https://orcid.org/0009-0006-1668-2984>

⁵Department of Business Administration, University of the Potomac, Washington, DC, USA; <https://orcid.org/0009-0004-4676-9576>

⁶Department of Business Administration, University of the Potomac, Washington, DC, USA; <https://orcid.org/0009-0006-5774-1709>

⁷Department of Business Administration, Washington University of Science and Technology, Virginia, USA; <https://orcid.org/0009-0002-2785-2601>

Corresponding Author: Rahat Hasan, **E-mail:** r.hasan.7505@westcliff.edu

ABSTRACT

This study develops and evaluates a machine learning framework for predicting business process delays and supporting risk-based operational prioritization under limited review capacity. Using the BPI Challenge 2017 event log, comprising 31,509 loan-application cases and 1,202,267 events, the study applies chronological, case-disjoint validation to assess predictive performance. Following data screening, 16,481 training, 4,025 validation, and 6,275 testing cases were selected for event-prefix analysis. Prolonged processing was defined using the 75th-percentile duration threshold (30.9 days) derived exclusively from training data. Logistic regression, random forest, and histogram gradient boosting models were evaluated against an elapsed-time baseline. The validation-selected logistic regression model achieved an average precision of 0.394 (95% CI: 0.371–0.414) and ROC-AUC of 0.686 on the independent test set. Under a simulated 10% review-capacity constraint, risk-based prioritization identified 17.3% of prolonged cases, compared with 16.4% using elapsed-time ranking and 10.1% through random selection. Although the improvement over the elapsed-time baseline was statistically inconclusive, the findings demonstrate the feasibility of predictive analytics for retrospective business process risk assessment. The study provides a reproducible framework for operational decision support while highlighting the need for prospective validation before claims of improved process efficiency can be established.

KEYWORDS

predictive process monitoring; business process management; interpretable machine learning; chronological validation; loan applications; review prioritization.

ARTICLE INFORMATION

ACCEPTED: 01 December 2025

PUBLISHED: 20 December 2025

DOI: 10.32996/jbms.2025.7.10.7

1. Introduction

Digital enterprise systems record the timing and sequence of operational work. Process mining uses these traces to explain execution patterns [1], while predictive process monitoring estimates outcomes before cases terminate [3,4]. For financial-service workflows, delays can constrain review capacity and degrade responsiveness, but a model with good discrimination does not

necessarily prescribe useful operational action. Predictive accuracy, review capacity and intervention effectiveness are different estimands [9,10].

This paper tests whether prefix-only information from a public loan-application log predicts unusually prolonged processing and whether a fixed review budget captures more prolonged cases than simpler ranking heuristics. The contribution is a documented, reproducible comparison joining temporal holdouts, calibrated-score assessment, supplementary remaining-time forecasting and noncausal capacity triage; neither algorithmic novelty nor causal process optimization is asserted.

2. Literature Review

Tax et al. [3] introduced neural sequence modeling for event-log predictions; Teinemaa et al. [4] benchmarked outcome-oriented process monitoring, and Verenich et al. [5] studied remaining-time forecasting. Predictive stability along case trajectories was emphasized by Teinemaa et al. [6]. Ceravolo et al. [7] discuss deployability and ongoing methodological challenges; explanation usefulness depends on human context and model choice [8]. Alarm-based intervention strategies [9] and causally informed, capacity-limited allocation [10] demonstrate why good risk ranking does not identify a beneficial treatment policy. Accordingly, this study restricts optimization claims to retrospective prioritization rather than realized case-time reductions.

Related cross-domain research offers complementary, but not directly comparable, evidence. Rahman et al. [15] applied stacking ensembles and SHAP explanations to employee-attrition prediction, illustrating an enterprise-risk application of interpretable machine learning. Gomes et al. [16] reviewed generative-AI administrative workflows and emphasized the distinction between reported time savings and demonstrated financial or operational returns. Javed et al. [17] examined temporal testing, calibration and uncertainty in population-health risk prediction. These studies motivate transparent evaluation and decision-oriented interpretation; none establishes effectiveness of intervention policies for the loan-process event log studied here.

3. Materials and Methods

3.1 Source and cohort construction

The original compressed XES log [2] was parsed trace-by-trace (SHA-256: 183c5e5189282779...). Event activity and timestamps were checked; events were stably ordered in UTC and apparent repeated event tuples retained. The source contained 31,509 traces, 1,202,267 events and 26 distinct activities. An eligible case required the final application-status milestone to be A_Pending, A_Denied or A_Cancelled; its timestamp was used as the terminal point. Nonterminal cases were excluded, not counted as favorable outcomes. Chronological cutoffs were 2016-08-15 and 2016-10-18; training and validation cases had to finish before their respective next-period cutoff.

Audit Measure	Value	Cohort/condition	Count
Cases (raw)	31,509	Terminal-eligible	31,411
Events (raw)	1,202,267	Train ten-event snapshots	16,481
Distinct activities	26	Validation snapshots	4,025
Long-duration threshold	30.93 days	Test snapshots	6,275

Table 1. Raw-log audit and post-filter ten-event eligible cohorts.

3.2 Outcome, features and leakage control

At event 10, eligible cases must remain active and must not already exceed the long-duration limit. Duration is the observed interval from first event to final application-status milestone. The binary outcome equals one when this duration exceeds the 75th percentile of training durations. Prefix features comprise elapsed time, last-event gap, distinct activities, application/offer/work counts, rework count, hour and weekday, events in the prior 24 hours and the last two activity labels. No eventual case status, future event or realized total duration enters a feature. The evaluation nevertheless selects cases known retrospectively to terminate and can overstate deployment performance for live censored cohorts.

3.3 Learning, validation and supplementary task

Candidate models are logistic regression with elapsed time alone, full-feature logistic regression, random forest [11] and histogram gradient boosting (a tree boosting family conceptually related to [12]). Hyperparameters are chosen by validation average precision (AP), with final discrimination and Brier score evaluated once on the holdout. The validation set also chooses classification thresholds for F1. Case-bootstrap intervals use 200 samples; at prefix stages 5 and 20 the best-selected model family is retrained with the locked hyperparameters. For remaining-time forecasting, histogram boosting regression is compared against a training-median baseline using MAE and RMSE. Held-out permutation importance measures model reliance, not causal impact [13]. The cross-domain temporal evaluation in [17] provides methodological context, but no health-data observations are used in this study.

3.4 Retrospective capacity analysis

Cases are ranked by model risk, elapsed processing time, earliest start time and random selection at budgets of 5%, 10% and 20%. Capture is the fraction of eventually prolonged holdout cases selected within each budget. These rankings pool decisions made at

different timestamps and therefore are offline diagnostics, not an executable daily staffing schedule. Any hypothetical net-value calculation uses fixed assumed intervention efficacy and review cost rather than effect estimates from observed interventions [10].

4. Results and Analysis

4.1 Cohort composition and discrimination

After filtering, 31,411 cases had a usable terminal application milestone. The ten-event snapshot cohorts comprised 26,781 cases. The delayed-case rate was 25.0% in training, 18.5% in validation and 25.1% in the final holdout. The changed prevalence reinforces the need to interpret AP relative to the underlying class share. Table 2 summarizes the untouched holdout.

Model	AP	ROC-AUC	Brier	F1
Prefix LR	0.394	0.686	0.174	0.449
Elapsed-only LR	0.380	0.675	0.180	0.494
Histogram boosting	0.391	0.682	0.176	0.456
Random forest	0.394	0.688	0.175	0.481

Table 2. Held-out prediction performance after validation-only model selection.

The validation-selected Prefix LR achieved AP 0.394 and ROC-AUC 0.686, with Brier score 0.174; the elapsed-only baseline achieved AP 0.380. The AP interval was 0.371–0.414. These bootstrap intervals describe case sampling, not institutional or temporal generalizability. Eight equal-count calibration bins produced an expected calibration error of 0.015; calibration plots complement Brier-score reporting [14].

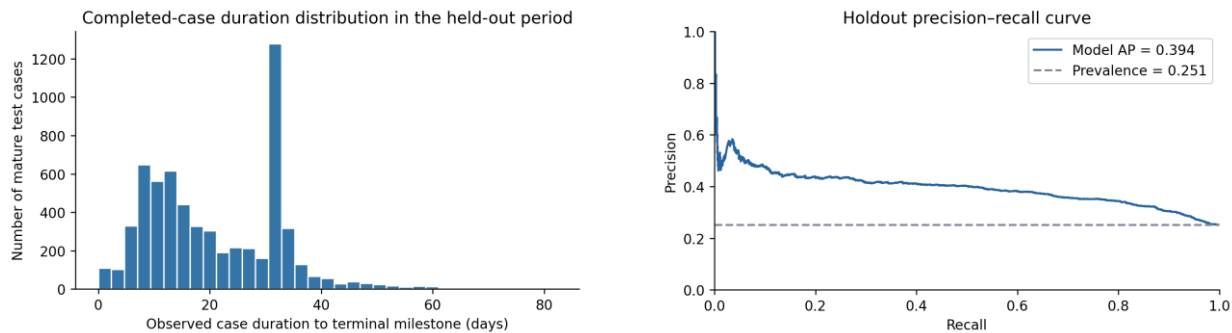


Figure 1. Holdout case-duration distribution (left) and precision-recall curve for the validation-selected model (right). Source: original event log.

4.2 Remaining time, interpretation and prefix checks

For observed-terminal snapshots, histogram boosting remaining-time regression obtained test MAE 9.77 days and RMSE 11.59 days; the training-median baseline returned MAE 9.96 days. The regression model was not superior on validation MAE; its small holdout MAE advantage should not be over-interpreted. The test-set permutation analysis ranked *last_activity*, *elapsed_hours*, *previous_activity* as the top features by AP decrease. These associations are predictive and may reflect progress through the process rather than modifiable causes. Sensitivity across the 5-, 10- and 20-event stages appears in Table 3.

Prefix stage	Cases	Delayed share	AP	ROC-AUC
5 events	6,283	25.1%	0.337	0.640
10 events	6,275	25.1%	0.394	0.686
20 events	5,279	22.4%	0.436	0.731

Table 3. Prefix-stage robustness for the selected estimator family.

4.3 Offline capacity-based prioritization

At a 10% review budget, model-ranked selection captured 17.3% of ultimately prolonged cases, while random selection captured 10.1% and elapsed-time ranking captured 16.4%. The predictive policy's precision at capacity was 43.5%. The paired case-bootstrap difference from elapsed-time ranking at 10% capacity was 1.04 percentage points (95% CI -0.70 to 2.61); these resampling intervals do not establish operational efficacy. Figure 2 shows review-capacity sensitivity. The observed advantages or disadvantages are retrospective case-selection statistics, not measured reductions in process duration.

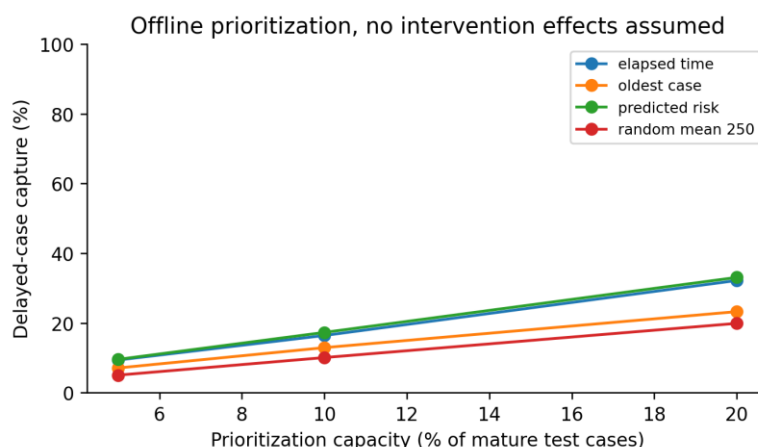


Figure 2. Delayed-case capture under four retrospective allocation policies at varying review capacities; no real interventions were evaluated.

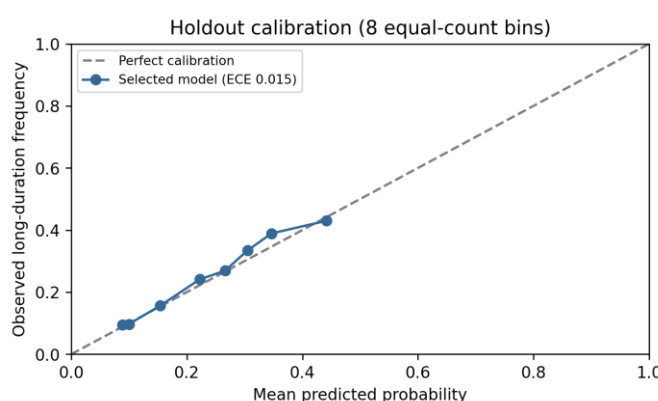


Figure 3. Quantile-bin reliability plot for the validation-selected model on the chronological holdout.

5. Discussion

The evaluation connects early-prefix prediction to the practical question of which cases warrant scarce attention. Whether higher AP converts into improved top-budget capture is an empirical rather than automatic relationship. Chronological splitting and baseline comparisons strengthen interpretability, while a separate regression task tests remaining-time predictability. The method provides transferable code for business-process analytics but does not validate broader gains in IT project management or external financial operations without replication. Consistent with the distinction between workflow indicators and realized financial value in [16], the retrospective capacity analysis should not be treated as measured cost savings.

There are substantial limitations. First, the outcome is a training-relative duration percentile, not a contractual breach or economic loss. Second, restricting to eventually terminated cases and excluding already-long prefixes introduces selection and right-censoring biases. Third, the observed application-status milestone may not capture all subsequent administrative work. Fourth, pooled historic prioritization does not represent one real-time intervention queue. Finally, the log contains no randomized treatment assignment or credible causal effect information, so benefits, losses, return on investment and achieved duration reductions cannot be estimated [9,10]. The feature reliance analysis is not SHAP and must not be interpreted causally [13].

6. Conclusion and Future Research

Using 1,202,267 events from the BPI Challenge 2017 dataset, this study developed and evaluated a reproducible machine learning framework for event-prefix delay prediction, remaining-time estimation, and risk-based operational prioritization. The validation-selected model achieved an average precision of 0.394 on the chronological holdout set. Under a simulated 10% review-capacity constraint, the proposed ranking approach identified 17.3% of prolonged cases, compared with 16.4% using an elapsed-time baseline. Although the improvement was modest and statistically inconclusive, the findings demonstrate the feasibility of integrating predictive analytics with capacity-constrained business process monitoring. Future research should incorporate censoring-aware survival models, rolling-origin validation, multi-institutional datasets, and prospective intervention studies. Overall, this study provides a reproducible foundation for explainable, risk-informed operational decision support without establishing causal improvements in process efficiency.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

- [1] van der Aalst, W. M. P. (2022). Process Mining: A 360 Degree Overview. In *Process Mining Handbook* , pp. 3–34. Springer. https://doi.org/10.1007/978-3-031-08848-3_1
- [2] van Dongen, B. F. (2017). BPI Challenge 2017 [Dataset]. 4TU.Centre for Research Data. <https://doi.org/10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b>
- [3] Tax, N., Verenich, I., La Rosa, M., & Dumas, M. (2017). Predictive Business Process Monitoring with LSTM Neural Networks. *CAiSE 2017* , pp. 477–492. https://doi.org/10.1007/978-3-319-59536-8_30
- [4] Teinemaa, I., Dumas, M., La Rosa, M., & Maggi, F. M. (2019). Outcome-Oriented Predictive Process Monitoring: Review and Benchmark. *ACM Transactions on Knowledge Discovery from Data* , 13(2), 17:1–17:57. <https://doi.org/10.1145/3301300>
- [5] Verenich, I., Dumas, M., La Rosa, M., Maggi, F. M., & Teinemaa, I. (2019). Survey and Cross-benchmark Comparison of Remaining Time Prediction Methods in Business Process Monitoring. *ACM Transactions on Intelligent Systems and Technology* , 10(4), Article 34. <https://doi.org/10.1145/3331449>
- [6] Teinemaa, I., Dumas, M., Leontjeva, A., & Maggi, F. M. (2018). Temporal Stability in Predictive Process Monitoring. *Data Mining and Knowledge Discovery* , 32, 1306–1338. <https://doi.org/10.1007/s10618-018-0575-9>
- [7] Ceravolo, P., Comuzzi, M., De Weerd, J., Di Francescomarino, C., & Maggi, F. M. (2024). Predictive process monitoring: concepts, challenges, and future research directions. *Process Science* , 1, Article 2. <https://doi.org/10.1007/s44311-024-00002-4>
- [8] Rizzi, W., Comuzzi, M., Di Francescomarino, C., Ghidini, C., Lee, S., Maggi, F. M., & Nolte, A. (2024). Explainable predictive process monitoring: a user evaluation. *Process Science* , 1, Article 3. <https://doi.org/10.1007/s44311-024-00003-3>
- [9] Fahrenkrog-Petersen, S. A., Tax, N., Teinemaa, I., Dumas, M., de Leoni, M., Maggi, F. M., & Weidlich, M. (2022). Fire now, fire later: alarm-based systems for prescriptive process monitoring. *Knowledge and Information Systems* , 64, 559–587. <https://doi.org/10.1007/s10115-021-01633-w>
- [10] Shoush, M., & Dumas, M. (2022). Prescriptive Process Monitoring Under Resource Constraints: A Causal Inference Approach. In *Process Mining Workshops* , pp. 180–193. https://doi.org/10.1007/978-3-030-98581-3_14
- [11] Breiman, L. (2001). Random Forests. *Machine Learning* , 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- [12] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of ACM SIGKDD* , pp. 785–794. <https://doi.org/10.1145/2939672.2939785>
- [13] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* , 30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
- [14] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *Proceedings of Machine Learning Research* , 70, 1321–1330. <https://proceedings.mlr.press/v70/guo17a.html>
- [15] Rahman, M. A., Devnath, R. K., Niloy, S. B., Mehedi, C. M., Chowdhury, T. H., & Javed, M. I. K. (2025). A Stacking Ensemble Framework for Predicting Employee Turnover: Explainable AI with SHAP. *2025 IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS)*. <https://doi.org/10.1109/COMPAS67506.2025.11381730>
- [16] Gomes, C. A., Javed, M. I. K., Ahammad, T., Mandal, S., Hassan, S., Sejan, M. K., Ponduru, P. S., Bhuiyan, M. M., & Jamali, M. L. (2025). Operational and Financial Evidence for Generative AI in United States Healthcare Administrative Workflows. *Frontiers in Computer Science and Artificial Intelligence*, 4(4), 138–145. <https://al-kindipublisher.com/index.php/fcsai/article/view/13502>
- [17] Javed, M. I. K., Imran, M., Gomes, C. A., & Ponduru, P. S. (2025). Machine Learning-Based Prediction of Poor Self-Rated Health Among U.S. Adults Using Behavioral and Socioeconomic Factors. *World Journal of Advanced Research and Reviews*, 27(1), 2817–2829. <https://doi.org/10.30574/wjarr.2025.27.1.2649>