



International Congress on Economics, Management and Business Studies

Hosted Online from New York, USA

Date: 23rd June , 2026

Website: <https://econferencia.com>

ARTIFICIAL INTELLIGENCE AND ECONOMETRIC VALIDATION: A THREE-REGIME APPROACH TO FORECASTING RETAIL CHAIN FINANCIAL PERFORMANCE

Khurramov Eldor

Tashkent State University of Economics

Abstract

Artificial intelligence and machine learning models are increasingly used to forecast the financial performance of retail chains, yet their genuine practical effectiveness depends heavily on the choice of econometric validation method. Using a panel dataset of 158,715 store-day observations from 140 outlets of a confectionery-café chain operating in Uzbekistan, this thesis comparatively evaluates XGBoost, CatBoost, and a regularized recurrent (GRU) neural network in forecasting revenue and profitability across three validation regimes of increasing strictness. The results show that model accuracy (R^2) declines sharply as the validation regime becomes stricter, that tree-based ensembles are markedly more stable than the recurrent model, and that Decision Curve Analysis confirms substantial managerial value even where predictive accuracy is modest. The findings demonstrate the critical importance of validation-methodology choice when deploying artificial intelligence in financial practice.

Keywords: Artificial intelligence, machine learning, econometric validation, XGBoost, CatBoost, deep learning, decision curve analysis.

Artificial intelligence algorithms are increasingly applied to forecast financial performance because they can capture complex, non-linear relationships that traditional linear econometric models often miss. In the retail-chain context, such



International Congress on Economics, Management and Business Studies

Hosted Online from New York, USA

Date: 23rd June , 2026

Website: <https://econferencia.com>

models hold promise for forecasting future store revenue or profitability, flagging financially vulnerable outlets, and allocating managerial attention efficiently. However, the high accuracy figures reported for machine learning models frequently misrepresent their real-world effectiveness, because these figures depend critically on the chosen validation method.

The most common error when working with panel-structured data — repeated observations of the same units (stores) over time — is to split training and test sets randomly or purely by time, so that a given store's observations from different days appear in both sets. This produces data leakage: the resulting accuracy score reflects how well the model has memorized a familiar store's future days, not how well it can forecast a genuinely new, previously unseen store. This distinction matters greatly in practice, since the chain's actual forecasting need is to predict the performance of stores it has not yet opened.

At the same time, a low accuracy score does not necessarily imply that a model is managerially useless: even a model that only ranks stores correctly by relative risk can carry substantial decision value. To capture this dimension, this thesis complements conventional accuracy metrics with Decision Curve Analysis (Vickers and Elkin, 2006), a decision-analytic framework originally developed in clinical medicine that compares a classifier's net benefit against “treat-all” and “treat-none” strategies across a range of decision thresholds.

The empirical analysis uses the same panel of 158,715 store-day observations from 140 confectionery-café outlets in Uzbekistan (2022–2026) described above, with an identical 28-variable feature set applied uniformly across models to ensure a fair comparison. Three models were compared: XGBoost and CatBoost gradient-boosted tree ensembles (Chen and Guestrin, 2016; Prokhorenkova et al., 2018), and a regularized, gradient-clipped, five-model-ensemble GRU recurrent network. Three validation regimes of increasing strictness were constructed: (A)



International Congress on Economics, Management and Business Studies

Hosted Online from New York, USA

Date: 23rd June , 2026

Website: <https://econferencia.com>

a lenient time-based split, in which each store's early period is used for training and its later period for testing; (B) a store-level five-fold grouped cross-validation, in which entire stores are held out for testing; and (C) a strict joint store-and-time split, in which training uses only “old” stores observed in an earlier period and testing uses only “new” stores observed in a later period — the most realistic proxy for the chain's actual forecasting task of predicting a newly opened store's future performance. Regime C was repeated across five random seeds to characterize sampling variability.

Model accuracy declined sharply and consistently as the validation regime became stricter. For revenue, XGBoost's R^2 fell from 0.7298 under the lenient regime to 0.4445 under the store-grouped regime and to only 0.1813 under the strict joint regime — a loss of more than 75 percent of explained variance. CatBoost showed an almost identical pattern (0.7344 → 0.4018 → 0.1249). The decline was even sharper for profitability: from 0.6240 to 0.0715 for XGBoost, and from 0.6148 to 0.0839 for CatBoost. The two tree-based models tracked each other closely across all three regimes, indicating that this pattern is not an artifact of a particular algorithm. The GRU model behaved differently: its nominal revenue R^2 under the strict regime (0.2195) was numerically higher than either tree model but carried a standard deviation (± 0.1710) too large to be considered a reliable advantage, while its profitability R^2 under the strict regime was negative (-0.0286) — worse than simply predicting the training-sample mean — even though both tree models retained positive R^2 under the same regime.

This gap persisted despite regularization, gradient clipping, and five-model ensembling applied to the GRU, indicating that it is structural rather than a tuning artifact, consistent with evidence that tree-based ensembles remain highly competitive with deep learning on small-to-medium tabular datasets (Grinsztajn et al., 2022). Despite the pronounced R^2 shrinkage under the strict regime,



International Congress on Economics, Management and Business Studies

Hosted Online from New York, USA

Date: 23rd June , 2026

Website: <https://econferencia.com>

Decision Curve Analysis showed that the machine-learning classifier retained a clear net-benefit advantage over both “treat-all” and “treat-none” strategies across a wide range of decision thresholds when flagging underperforming stores, confirming that the model keeps practical value for identifying which stores warrant managerial attention even where its continuous-outcome accuracy is modest. These findings imply that retail chains and researchers should evaluate forecasting models under store-level and strict joint-split validation regimes rather than relying on lenient time-based splits alone, should favor tree-based ensembles as a strong default choice for retail panel data given limited numbers of independent stores, and should apply decision-analytic tools such as Decision Curve Analysis to assess practical value independently of raw accuracy scores. Future research should extend this validation-sensitivity analysis to additional model architectures and retail sectors.

References

1. Chen, T., Guestrin, C. XGBoost: A scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — 2016. — P. 785–794.
2. Grinsztajn, L., Oyallon, E., Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? // Advances in Neural Information Processing Systems 35 (NeurIPS 2022). — 2022.
3. Hochreiter, S., Schmidhuber, J. Long short-term memory // Neural Computation. — 1997. — Vol. 9, No. 8. — P. 1735–1780.
4. Kanász, R., Gnip, P., Zoričák, M., Drotár, P. Bankruptcy Prediction Using Ensemble of Autoencoders Optimized by Genetic Algorithm // PeerJ Computer Science. — 2023. — Vol. 9. — P. e1257.



International Congress on Economics, Management and Business Studies

Hosted Online from New York, USA

Date: 23rd June , 2026

Website: <https://econferencia.com>

5. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. CatBoost: Unbiased boosting with categorical features // Advances in Neural Information Processing Systems 31 (NeurIPS 2018). — 2018. — P. 6638–6648.
6. Sonkavde, G., Dharrao, D. S., Bongale, A. M., Deokate, S. T., Doreswamy, D., Bhat, S. K. Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models // International Journal of Financial Studies. — 2023. — Vol. 11, No. 3. — P. 94.
7. Vickers, A. J., Elkin, E. B. Decision curve analysis: A novel method for evaluating prediction models // Medical Decision Making. — 2006. — Vol. 26, No. 6. — P. 565–574.