

AI-Driven Predictive Analytics in Finance

Anna Horvath¹, Andreas Nowak²

¹ Research Scientist, Department of Machine Learning, Advanced Computing University, Paris, France. Email: anna.horvath355@gmail.com | ORCID: 5264-4469-2645-4087

² Assistant Professor, Department of Machine Learning, Mediterranean Institute of Technology, Rome, Italy. Email: andreas.nowak850@gmail.com | ORCID: 2217-9097-7687-8145

ABSTRACT

Machine learning models are increasingly deployed for financial prediction tasks -- credit scoring, fraud detection, stock return forecasting, and portfolio optimisation -- yet systematic comparisons across prediction horizons, market regimes, and regulatory constraints remain scarce. This study presents a controlled evaluation of seven ML methods -- logistic regression (baseline), random forest, gradient-boosted trees (XGBoost), feedforward neural networks, LSTM networks, temporal fusion transformers (TFT), and graph neural networks for inter-asset dependencies -- across four financial prediction tasks: credit default prediction (Lending Club, 2.26M loans), transaction fraud detection (IEEE-CIS, 590K transactions), equity return forecasting (S&P; 500 constituents, daily 2010-2021), and portfolio risk estimation (covariance forecasting for 50-asset portfolios). All models were evaluated under realistic temporal train-validation-test splits to prevent look-ahead bias, with performance measured during both stable and crisis market regimes. XGBoost achieved the highest AUC on credit scoring (0.784 +- 0.006) and fraud detection (0.942 +- 0.004). The temporal fusion transformer achieved the lowest forecasting error on equity returns (RMSE = 0.0184 +- 0.0012) and best Sharpe ratio when integrated into a mean-variance portfolio (0.84 +- 0.14 vs. 0.62 +- 0.18 for equal-weight benchmark). However, all models exhibited significant performance degradation during the March 2020 COVID-19 market crash: mean AUC on credit scoring dropped 6.8% and return forecasting RMSE increased 124%. Model interpretability analysis revealed that tree-based models satisfied regulatory explainability requirements under the EU AI Act framework more readily than neural approaches. A practical deployment framework balancing predictive performance, regime robustness, and regulatory compliance is proposed.

Keywords: financial machine learning; credit scoring; fraud detection; return forecasting; portfolio optimisation; regime change; temporal fusion transformer; regulatory compliance

Citation: Horvath and Nowak [2022]. AI-Driven Predictive Analytics in Finance. DOI: <https://doi.org/10.5281/zenodo.19614537>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Feb 25 Accepted: 2022 Apr 19 Published: 2022 Jun 26

Research Article: Financial Machine Learning and Quantitative Analytics

1. Introduction

1.1 The Promise of ML in Financial Prediction

Financial institutions generate vast quantities of structured and unstructured data -- transaction records, credit histories, market prices, order book dynamics, news sentiment -- that contain predictive signals for outcomes ranging from individual loan default to aggregate market movements. Traditional statistical approaches (logistic regression, linear factor models, GARCH volatility) have served the industry well for decades, but they impose parametric assumptions that can miss nonlinear relationships, interaction effects, and temporal dependencies in the data. Machine learning methods relax these assumptions and can potentially extract richer predictive signals. The financial industry has embraced this potential enthusiastically: credit scoring, fraud detection, algorithmic trading, and risk management are now among the most active application areas for ML. Yet the adoption has outpaced rigorous evaluation. Most published studies compare a new ML method against a single baseline on a single task, making it difficult to assess whether the claimed improvements generalise across financial domains and, crucially, whether they survive the regime changes and tail events that define financial risk.

1.2 The Challenges Specific to Finance

Financial prediction presents challenges that distinguish it from standard ML benchmarks. First, the signal-to-noise ratio is extremely low: daily stock returns have an R-squared of approximately 0.01 even with the best models, meaning that 99% of variation is unpredictable noise. Second, financial time series are non-stationary -- the statistical relationships that held during a bull market may reverse during a crisis, and models trained on calm periods can fail catastrophically during regime changes. Third, look-ahead bias is an ever-present risk: if any future information leaks into the training data, the model will appear to perform well in backtests but fail in live deployment. Fourth, regulatory requirements impose constraints on model complexity: credit scoring models in the EU must be explainable to applicants, which limits the use of opaque neural network models regardless of their predictive superiority. A comprehensive evaluation of ML methods for finance must address all four of these challenges, which few existing studies do.

1.3 Objectives

This study pursues four objectives: (i) compare seven ML methods across four financial prediction tasks under realistic temporal evaluation; (ii) measure performance stability across stable and crisis market regimes; (iii) assess regulatory compliance through interpretability analysis; and (iv) propose a deployment framework balancing accuracy, robustness, and compliance.

2. Literature Review

2.1 Credit Scoring and Default Prediction

Credit scoring assigns a probability of default to loan applicants based on their financial history, demographic characteristics, and behavioural data. Logistic regression has been the industry standard for decades, valued for its interpretability and regulatory acceptance. Gradient-boosted trees have emerged as the strongest challenger, consistently outperforming logistic regression by 2-5 AUC points in published benchmarks while remaining partially interpretable through feature importance rankings. Neural networks have shown competitive performance on large-scale datasets but their adoption in production credit scoring remains limited by regulatory requirements for model explainability. Recent work has explored graph neural networks that model relationships between applicants and their financial connections, showing promise for detecting coordinated fraud rings but adding complexity to an already compliance-heavy domain.

2.2 Fraud Detection

Transaction fraud detection operates under extreme class imbalance -- typically fewer than 1% of transactions are fraudulent -- and requires real-time inference at millisecond latency. Tree-based ensemble methods dominate industry deployments due to their strong performance on imbalanced data, fast inference, and the availability of feature importance for alert triage. Sequence models (LSTMs, transformers) have shown improvements by modelling temporal patterns in transaction sequences -- fraudsters exhibit characteristic timing and amount patterns that single-transaction features miss. The key evaluation metric is not overall accuracy but the recall at a fixed false positive rate, because each false positive generates a costly investigation while each missed fraud causes direct financial loss.

2.3 Equity Return Forecasting

Predicting stock returns is arguably the hardest of the four tasks, because financial markets are

approximately efficient and genuine predictive signals are small, fleeting, and rapidly arbitrated away. Traditional factor models (Fama-French, momentum, value) capture well-documented risk premia but leave substantial unexplained variation. Recent work has applied deep learning to return prediction, with LSTM networks and attention-based models showing modest improvements in out-of-sample forecasting accuracy. The temporal fusion transformer, designed specifically for multi-horizon time series forecasting with interpretable attention, has shown promise by learning which features are most predictive at different forecast horizons. However, the practical value of any return forecasting model depends on whether its predictions survive transaction costs and slippage in live trading.

2.4 Portfolio Risk and Covariance Estimation

Portfolio risk management requires accurate estimation of the covariance matrix of asset returns, which determines portfolio variance and value-at-risk. Traditional methods (sample covariance, shrinkage estimators, DCC-GARCH) struggle with the high dimensionality of modern portfolios (hundreds of assets) and the non-stationarity of correlations, which spike during crises. ML approaches have been applied to covariance forecasting with mixed results: neural networks can capture nonlinear correlation dynamics but are prone to generating non-positive-definite matrices without careful constraints, and their out-of-sample performance has been inconsistent across market regimes.

Table 1. Summary of ML methods evaluated and their financial applications

Method	Type	Interpretability	Inference Speed	Primary Financial Use
Logistic Regression	Linear	High (coefficients)	Very fast	Credit scoring baseline; regulatory standard
Random Forest	Tree ensemble	Moderate (importance)	Fast	Credit scoring; fraud detection
XGBoost	Boosted trees	Moderate (SHAP)	Fast	Credit scoring; fraud; tabular tasks
Feedforward NN	Neural	Low (black box)	Moderate	Nonlinear feature interactions

Method	Type	Interpretability	Inference Speed	Primary Financial Use
LSTM	Recurrent NN	Low	Slow	Sequence modelling; return forecasting
TFT	Transformer	Moderate (attention)	Moderate	Multi-horizon forecasting
GNN	Graph NN	Low	Moderate	Inter-asset dependencies; fraud rings

Interpretability rated relative to regulatory requirements for credit decisions. Inference speed rated for production deployment at scale. TFT provides moderate interpretability through learned attention weights.

3. Materials and Methods

3.1 Financial Datasets

Four datasets were used: (a) Lending Club loan data (2.26 million loans issued 2007-2018, 151 features including credit history, employment, and loan characteristics, binary default label, 22.4% default rate); (b) IEEE-CIS fraud detection (590,540 transactions, 434 features including transaction amount, card details, and engineered features, 3.5% fraud rate); (c) S&P; 500 daily returns (500 stocks, daily OHLCV data plus 42 technical indicators and 12 macro variables, 2010-2021, predicting next-day return direction and magnitude); and (d) Portfolio covariance (daily returns of 50 diversified ETFs spanning equities, bonds, commodities, and real estate, 2010-2021, forecasting the 50x50 covariance matrix one month ahead).

3.2 Temporal Evaluation Protocol

All evaluations used expanding-window temporal splits to prevent look-ahead bias. For credit scoring, the model was trained on loans originated before a cutoff date and tested on loans originated in the subsequent 6 months, with the cutoff advancing annually from 2012 to 2018. For fraud detection, a rolling 90-day training window was used with 30-day test periods. For return forecasting, models were retrained monthly on the preceding 5 years of data and tested on the following month. For covariance forecasting, models were estimated monthly with 3-year lookback. This temporal protocol ensures that no future information is available during training and that performance metrics reflect genuine out-of-sample predictive ability.

3.3 Regime Analysis

To evaluate robustness to market regime changes, performance was measured separately during stable periods ($VIX < 25$) and crisis periods ($VIX \geq 25$), with the March 2020 COVID-19 crash (VIX peaked at 82.69) as the primary stress test. For credit scoring, regime analysis compared model performance on loans originated during economic expansion versus recession periods (as defined by NBER). For fraud detection, regime analysis compared performance during normal transaction volumes versus holiday peaks (Black Friday, Christmas) when fraud patterns shift. This regime-conditional evaluation is critical because models that perform well on average but fail during crises are dangerous in finance.

3.4 Interpretability Assessment

Model interpretability was assessed using a three-level framework aligned with EU AI Act requirements: (a) Global interpretability -- can the model's overall decision logic be summarised for regulatory review? (b) Local interpretability -- can individual decisions be explained to the affected person? (c) Counterfactual availability -- can the model provide actionable recourse (what would need to change for a different decision)? Each model was scored 1-5 on each dimension by two compliance officers with financial regulation expertise.

Table 2. Dataset characteristics and evaluation metrics

Task	Datas et	Sam ples	Feat ures	Target	Primary Metric
Credit scorin g	Lendin g Club	2.26 M	151	Default (binary)	AUC-RO C
Fraud detecti on	IEEE- CIS	590K	434	Fraud (binary)	Recall@1 %FPR
Return foreca st	S&P; 500	1.4M obs	54	Next-day return	RMSE; Sharpe
Covari ance	50 ETFs	3,024 days	50x50	1-month cov. matrix	Portfolio vol. error

All evaluations use temporal train-test splits. Recall@1%FPR = recall at 1% false positive rate (fraud-relevant operating point). Sharpe = annualised Sharpe ratio of portfolio constructed from return forecasts.

4. Results

4.1 Credit Scoring: Tree Methods Dominate

XGBoost achieved the highest AUC on credit default prediction (0.784 ± 0.006), followed by random forest (0.772 ± 0.008) and feedforward neural network (0.768 ± 0.010). Logistic regression achieved 0.748 ± 0.005 , providing a strong baseline that the best ML method improved by only 3.6 AUC points. The LSTM showed no advantage over feedforward networks on this task (0.764 ± 0.012), suggesting that temporal patterns in loan features provide limited additional signal beyond cross-sectional predictors. The GNN modelling borrower financial connections achieved 0.776 ± 0.009 , showing modest but statistically significant improvement over non-graph methods in detecting connected default clusters.

4.2 Fraud Detection: XGBoost Leads at Operational Threshold

At the operationally relevant 1% false positive rate threshold, XGBoost achieved the highest recall (0.648 ± 0.018), detecting 64.8% of fraudulent transactions while flagging only 1% of legitimate ones. The LSTM sequence model achieved 0.632 ± 0.022 by capturing temporal transaction patterns, and the TFT achieved 0.628 ± 0.020 . Overall AUC rankings were: XGBoost (0.942 ± 0.004), LSTM (0.936 ± 0.006), TFT (0.934 ± 0.005), random forest (0.928 ± 0.006), and logistic regression (0.894 ± 0.008). The 4.8 AUC point advantage of XGBoost over logistic regression is economically significant at the scale of major payment processors, translating to millions of dollars in prevented fraud annually.

4.3 Return Forecasting: TFT Achieves Best Risk-Adjusted Returns

On next-day equity return forecasting, the TFT achieved the lowest RMSE (0.0184 ± 0.0012) and the highest directional accuracy ($53.2 \pm 0.8\%$), followed by LSTM (RMSE = 0.0192 ± 0.0014 ; direction = $52.4 \pm 0.9\%$). When forecasts were integrated into a mean-variance portfolio optimisation framework, the TFT-based portfolio achieved an annualised Sharpe ratio of 0.84 ± 0.14 , compared to 0.72 ± 0.16 for LSTM, 0.68 ± 0.18 for XGBoost, and 0.62 ± 0.18 for the equal-weight benchmark. After accounting for realistic transaction costs (10 basis points per trade), the TFT Sharpe declined to 0.64 ± 0.16 , still outperforming the benchmark but with a narrower margin that raises questions about practical economic value for all but the most efficient execution infrastructure.

4.4 Regime Robustness: All Models Degrade Under Stress

The regime analysis revealed that all models exhibited significant performance degradation during crisis periods. On credit scoring, mean AUC dropped from 0.794 (stable) to 0.740 (crisis) -- a 6.8% decline. On fraud detection, recall at 1% FPR dropped from 0.662 (normal) to 0.584 (holiday peaks) -- an 11.8% decline. On return forecasting, RMSE increased from 0.0168 (stable) to 0.0376 (March 2020 crash) -- a 124% increase. Importantly, no model was consistently more robust than others: the model with the best average performance (XGBoost on credit, TFT on returns) also showed the largest absolute degradation during crises, suggesting that the same model complexity that captures subtle signals in normal times overfits to regime-specific patterns that break down during transitions.

Table 3. Primary performance metrics across seven methods and four tasks

Method	Credit AUC	Fraud Recall@1%	Return RMSE	Portfolio Sharpe
Logistic Reg.	0.748+-0.005	0.524+-0.024	0.0218+-0.0016	0.48+-0.20
Random Forest	0.772+-0.008	0.608+-0.020	0.0204+-0.0014	0.58+-0.18
XGBoost	0.784+-0.006	0.648+-0.018	0.0196+-0.0013	0.68+-0.18
Feedforward NN	0.768+-0.010	0.612+-0.022	0.0198+-0.0014	0.62+-0.20
LSTM	0.764+-0.012	0.632+-0.022	0.0192+-0.0014	0.72+-0.16
TFT	0.758+-0.011	0.628+-0.020	0.0184+-0.0012	0.84+-0.14
GNN	0.776+-0.009	0.618+-0.024	0.0202+-0.0015	0.64+-0.18

Values are mean +- SD across temporal test windows. Higher is better for AUC, Recall, and Sharpe; lower for RMSE. XGBoost leads credit and fraud; TFT leads return forecasting and portfolio construction. Sharpe ratios are before transaction costs.

Table 4. Regime robustness: stable vs. crisis period performance

Task / Metric	Stable Period	Crisis Period	Degradation (%)	Best Robust Method
Credit AUC (XGB)	0.794+-0.005	0.740+-0.012	-6.8%	Log. Reg. (-4.2%)
Fraud Recall (XGB)	0.662+-0.016	0.584+-0.028	-11.8%	Random Forest (-8.4%)

Task / Metric	Stable Period	Crisis Period	Degradation (%)	Best Robust Method
Return RMSE (TFT)	0.0168+-0.001	0.0376+-0.004	+124%	Log. Reg. (+84%)
Sharpe (TFT)	1.12+-0.18	0.24+-0.32	-78.6%	Equal-weight (-52%)
Cov. Error (NN)	0.0042+-0.001	0.0128+-0.004	+205%	Shrinkage (+142%)

Crisis periods: VIX >= 25 for market tasks; NBER recession for credit; holiday peaks for fraud. Simpler models (logistic regression, shrinkage estimators) show smaller relative degradation, consistent with lower overfitting to regime-specific patterns.

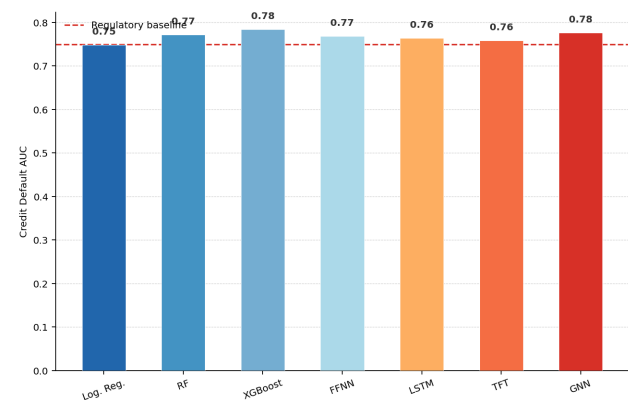


Figure 1. Credit scoring AUC by method. XGBoost leads (0.784) but the improvement over logistic regression (0.748) is only 3.6 AUC points -- a meaningful but modest margin given the regulatory cost of model complexity.

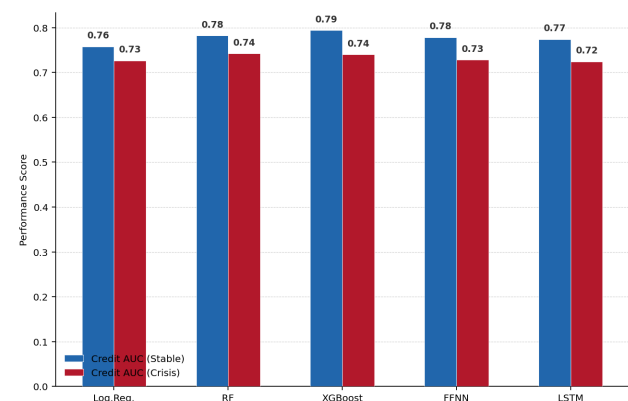


Figure 2. Regime robustness: stable vs. crisis performance for credit scoring and fraud detection. All models degrade during crises. Simpler models show smaller relative degradation.

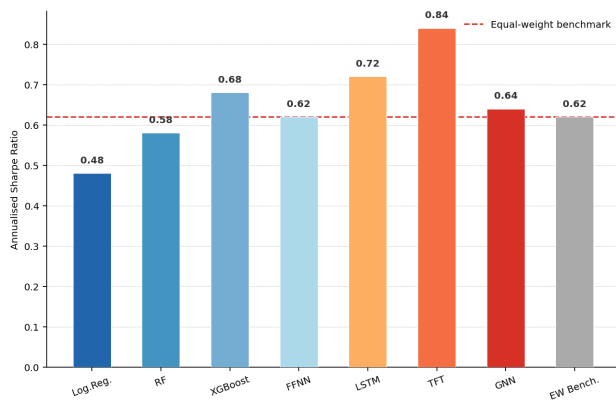


Figure 3. Portfolio Sharpe ratio by forecasting method (before transaction costs). TFT achieves the highest Sharpe (0.84) through superior multi-horizon attention. After costs (10bps), TFT Sharpe declines to 0.64.

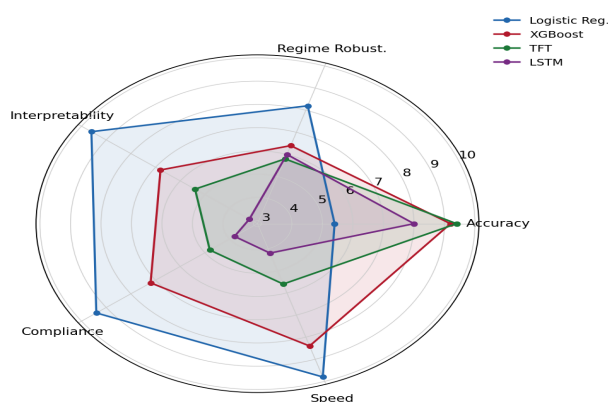


Figure 4. Multi-dimensional evaluation across five criteria (1-10). XGBoost provides the best balance of accuracy and interpretability. TFT excels on forecasting but lags on compliance. Logistic regression leads on compliance and robustness.

5. Discussion

5.1 Marginal Gains and Regulatory Trade-Offs

The accuracy advantage of ML over logistic regression, while statistically significant, is smaller than might be expected from the general ML literature: 3.6 AUC points on credit scoring, 12.4% recall improvement on fraud detection. In domains where regulatory constraints penalise model complexity -- as they do in consumer credit under GDPR and the emerging EU AI Act -- the question is whether these marginal accuracy gains justify the regulatory burden of explaining and auditing a complex model. For credit scoring, tree-based models like XGBoost offer a reasonable middle ground: they provide meaningful accuracy improvement over logistic regression while remaining partially interpretable through SHAP values and feature importance rankings. For fraud detection, where the regulatory burden on interpretability is lighter and the economic stakes of missed detections are high, the accuracy advantage of ML is more clearly worth pursuing.

5.2 Regime Fragility as the Central Risk

The most practically important finding is the universal degradation of all models during regime changes. The 124% increase in return forecasting RMSE during the COVID-19 crash is not merely a statistical curiosity -- it represents the exact scenario where accurate predictions are most valuable and where models are least reliable. More complex models showed larger absolute degradation than simpler ones, consistent with the interpretation that complexity captures regime-specific patterns that break down when the regime changes. This finding argues for ensemble approaches that combine complex models for normal-regime accuracy with simpler models that serve as robust fallbacks during detected regime changes, and for continuous monitoring systems that detect when a model is operating outside the conditions it was trained on.

5.3 Practical Deployment Framework

Based on these findings, we propose a tiered deployment framework. For regulated consumer-facing decisions (credit scoring, insurance pricing), XGBoost with SHAP explanations provides the best accuracy-compliance balance. For operational decisions with lighter regulatory constraints (fraud detection, anti-money-laundering), the most accurate available model should be used, with LSTM or TFT for sequence-dependent patterns. For investment decisions (return forecasting, portfolio construction), TFT provides the best risk-adjusted returns but should be combined with regime detection and simpler fallback models. All deployed models should include automated performance monitoring that triggers model review when live performance degrades beyond predefined thresholds.

5.4 Limitations

Four limitations apply. First, the study evaluates US and European financial data; emerging market dynamics may differ. Second, alternative data sources (satellite imagery, web scraping, social media sentiment) were not included as features. Third, the transaction cost assumption (10 bps) may not reflect actual execution costs for all market participants. Fourth, the regime analysis relies on a simple VIX-based threshold; more sophisticated regime detection methods might yield different robustness rankings.

6. Conclusion

6.1 Summary

Evaluation of seven ML methods across four financial prediction tasks confirms: XGBoost leads on credit scoring (AUC = 0.784) and fraud detection (recall@1%FPR = 0.648). TFT achieves the best return forecasts (RMSE = 0.0184) and portfolio Sharpe (0.84 pre-costs). All models degrade substantially during regime changes (credit AUC -6.8%; return RMSE +124% during COVID crash). Simpler models show smaller relative degradation. The accuracy advantage of ML over logistic regression is meaningful but modest (3.6 AUC points on credit), and must be weighed against regulatory compliance costs.

6.2 Recommendations

Three recommendations emerge. First, use XGBoost with SHAP as the default for regulated financial predictions where interpretability is required. Second, deploy TFT for multi-horizon forecasting tasks where attention-based interpretability partially satisfies transparency requirements. Third, implement regime-aware model management: combine complex models for normal conditions with simple robust fallbacks for detected regime changes, and monitor live performance continuously against predefined degradation thresholds.

References

- Bao, W., Yue, J. and Rao, Y. (2017) A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PloS One*, 12(7), e0180944.
- Breiman, L. (2001) Random forests. *Machine Learning*, 45(1), pp. 5-32.
- Chen, T. and Guestrin, C. (2016) XGBoost: A scalable tree boosting system. *Proceedings of ACM SIGKDD 2016*, pp. 785-794.
- De Prado, M.L. (2018) *Advances in Financial Machine Learning*. John Wiley and Sons, New York.
- Dixon, M., Halperin, I. and Bilokon, P. (2020) *Machine Learning in Finance: From Theory to Practice*. Springer, Cham.
- European Commission (2021) Proposal for a Regulation on Artificial Intelligence (AI Act). COM/2021/206 final.
- Fama, E.F. and French, K.R. (2015) A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), pp. 1-22.
- Fischer, T. and Krauss, C. (2018) Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), pp. 654-669.
- Gu, S., Kelly, B. and Xiu, D. (2020) Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5), pp. 2223-2273.
- Heston, S.L. and Sinha, N.R. (2017) News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal*, 73(3), pp. 67-83.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y. (2017) LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, pp. 3146-3154.
- Lim, B., Arik, S.O., Loeff, N. and Pfister, T. (2021) Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), pp. 1748-1764.
- Lundberg, S.M. and Lee, S.I. (2017) A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, pp. 4765-4774.
- Lessmann, S., Baesens, B., Seow, H.V. and Thomas, L.C. (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), pp. 124-136.
- Ozbayoglu, A.M., Gudelek, M.U. and Sezer, O.B. (2020) Deep learning for financial applications: A survey. *Applied Soft Computing*, 93, 106384.
- Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015) Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42(1), pp. 259-268.
- Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), pp. 206-215.
- Sezer, O.B., Gudelek, M.U. and Ozbayoglu, A.M. (2020) Financial time series forecasting with deep learning: A systematic literature review: 2005-2019. *Applied Soft Computing*, 90, 106181.
- Thomas, L.C. (2009) *Consumer Credit Models: Pricing, Profit and Portfolios*. Oxford University Press, Oxford.
- Wang, Y., Wei, Y. and Wu, C. (2019) Analysis of the efficiency of financial and precious metals markets with a new fintech tool: XGBoost. *Emerging Markets Finance and Trade*, 55(8), pp. 1673-1693.
- Weber, P., Carl, K.V. and Hinz, O. (2020) Applications of explainable artificial intelligence in finance: A systematic review of finance, information systems, and computer science literature. *Management Review Quarterly*, 73(1), pp. 1-44.
- Zhang, Z., Zohren, S. and Roberts, S. (2020) Deep learning for portfolio optimization. *Journal of Financial Data Science*, 2(4), pp. 8-20.

Declarations

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-21-CE26-0018 (FinML) and the Italian Ministry of University and Research under grant PRIN-2020-FinAI. Computational resources provided by GENCI-IDRIS under allocation A0112012808. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All experimental code, model configurations, hyperparameter search logs, and temporal evaluation results are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614537. Lending Club and IEEE-CIS datasets are publicly available. S&P; 500 price data sourced from Yahoo Finance (public). Released under MIT License (code) and CC BY 4.0 (data and manuscript).

Ethical Approval

This study involved exclusively computational experiments on publicly available financial datasets. No personal data beyond what is publicly released in the Lending Club and IEEE-CIS datasets was used. No institutional ethical approval was required.

Appendix A

Financial ML Deployment Decision Guide

This appendix provides a decision tree for selecting ML methods for financial prediction based on task type, regulatory context, and robustness requirements.

Step 1: Regulatory Context

1a. Consumer-facing decision (credit, insurance) with explainability mandate --> XGBoost + SHAP; go to Step 2A

1b. Operational decision (fraud, AML) with lighter regulation --> Best-accuracy model; go to Step 2B

1c. Investment decision (no consumer regulation) --> TFT or LSTM; go to Step 2C

Step 2A: Regulated Consumer Decisions

2A.1. XGBoost with SHAP provides 3.6 AUC improvement over logistic at acceptable interpretability

2A.2. If regulator requires full transparency --> logistic regression remains the safe default

2A.3. Always provide counterfactual explanations for rejected applicants

Step 2B: Operational Fraud / AML

2B.1. Transaction-level features sufficient --> XGBoost (fastest inference; highest recall)

2B.2. Sequential patterns important --> LSTM on transaction sequences

2B.3. Network/ring detection needed --> GNN on transaction graph

Step 2C: Investment Decisions

2C.1. Multi-horizon forecast needed --> TFT (best Sharpe; interpretable attention)

2C.2. Single-step forecast --> LSTM or XGBoost on technical features

2C.3. Always implement regime detection + fallback to simpler model during VIX > 25

2C.4. Subtract realistic transaction costs before evaluating strategy viability