ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep096

Training Data Size Matters: A Multi-Metric Framework Revealing How Data Availability Affects Forecasting Method Selection

Tharakesvulu Vangalapat[1]; Priyank Raj Sharma[2]; Somnath Banerjee[3]

[1]Sr. Principal Data Scientist, AI Architect Broadridge Capital One Austin, TX, USA
[2]Senior Manager AMFAM USA
[3]Senior Data Engineer, AI USA

Publication Date: 2026/09/12

Abstract:

- ***Background:***
  Practitioners receive contradictory guidance on forecasting method selection because existing comparisons use fixed train-test splits without examining how training data size affects rankings.

- ***Methods:***
  We evaluate 11 methods across 11 datasets (2,515–48,204 observations) using two protocols (60/20/20 and 70/15/15 splits) with five metrics (MAPE, SMAPE, RMSE, MAE, MASE).

- ***Results:***
  With 60% training data, Exponential Smoothing ranks best (4.80). With 70% training data, Prophet dominates (4.47), followed by LightGBM (4.75) and ETS (5.02). ETS improves 1.75 positions and LightGBM improves 1.56 positions, whereas Exponential Smoothing declines 1.24 positions. ElasticNet ranks last under both protocols (7.69 and 8.31).

- ***Conclusions:***
  A critical threshold at 65–70% training data governs when sophisticated methods surpass simpler alternatives, providing the first selection guidelines based on data availability for practitioners.

Keywords: *Time Series Forecasting, Method Selection, Training Data Requirements, Empirical Evaluation, Comparative Study.*

How to Cite: Tharakesvulu Vangalapat; Priyank Raj Sharma; Somnath Banerjee (2026) Training Data Size Matters: A Multi-Metric Framework Revealing How Data Availability Affects Forecasting Method Selection. *International Journal of Innovative Science and Research Technology*, 11(9), 76-84. https://doi.org/10.38124/ijisrt/26sep096

I. INTRODUCTION

Time series forecasting drives critical decisions across finance [1], healthcare [2], energy management [3], and transportation [4]. However, practitioners face contradictory guidance when selecting methods. Some studies claim that simple statistical methods remain competitive with complex approaches [5], [20], whereas others demonstrate superior machine learning performance [6].

These contradictions leave practitioners uncertain about which methods to deploy. A financial analyst forecasting quarterly revenue must choose between Prophet, ARIMA, and XGBoost without clear selection criteria. An energy manager predicting electricity demand faces similar uncertainty. Current guidance typically recommends trying multiple methods and selecting the best performer, but this approach provides no principled framework for understanding when each method type excels.

We hypothesize that these contradictions stem from an overlooked moderating variable: training data availability. Most comparative studies employ fixed train-test splits (e.g., 70/30 or 80/20) without investigating whether split ratios systematically affect method rankings. If rankings change with data availability, then existing claims about best methods lack critical contextual information. A method that excels with 80% training data might underperform with only 60% training data, and vice versa.

The underlying mechanisms support this hypothesis. Simple methods such as exponential smoothing estimate few

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep096

parameters describing trend and seasonal components, thereby avoiding overfitting with limited data and providing robust predictions. In contrast, sophisticated ensemble methods such as XGBoost construct thousands of decision trees capable of modeling complex nonlinear patterns but require sufficient data to distinguish genuine patterns from noise. The optimal method therefore depends on whether the available data can support complex models or requires simpler approaches.

➢ *Research Questions*

This study addresses three questions. RQ1: Do method rankings remain stable when training data increases from 60% to 70%, or do they change systematically? RQ2: Which methods benefit most from additional training data, and which perform better with limited data? RQ3: Can we identify training data thresholds that determine optimal method selection?

➢ *Contributions*

This research makes six contributions: C1: the first systematic comparison of 60/20/20 versus 70/15/15 splits, isolating training data size effects; C2: identification of the 65–70% range as a critical threshold; C3: empirical evidence that ElasticNet consistently ranks last; C4: a consistency-aware ranking framework aggregating five metrics; C5: evaluation across 11 modern datasets spanning 2000–2026; and C6: actionable deployment guidelines for common data availability scenarios.

➢ *Related Work*

Autoregressive Integrated Moving Average (ARIMA) models [7] and exponential smoothing [8] represent classical approaches that estimate parametric models of trend, seasonality, and error. Prophet [9] modernized these approaches with automatic changepoint detection and multiple seasonality modeling, and NeuralProphet [10] subsequently extended this framework with neural components. Tree-based methods, including Random Forest [13], XGBoost [14], LightGBM [15], and CatBoost [16], transform forecasting into supervised learning using lagged features. ElasticNet [11], [12] applies combined L1/L2 regularization for linear regression.

Recent neural architectures include N-BEATS [17], Informer [6], and FEDformer [18]. However, Nie et al. [19] found that transformers often underperform classical methods on datasets with limited observations, motivating our focus on classical and machine learning methods.

The M-competitions [5], [20], [21] established gold standard benchmarking, and Cerqueira et al. [22] provided evaluation guidelines emphasizing proper splitting and multiple metrics. Despite this extensive body of work, no existing study has systematically investigated how training data size affects method rankings. All major comparisons use fixed splits, leaving practitioners uncertain whether rankings generalize across data availability scenarios.

II. METHODS

➢ *Datasets*

We evaluate methods on 11 datasets spanning 2000–2026: five financial markets (S&P 500 with 2,515 observations, Bitcoin with 4,149, Ethereum with 3,000, Gold with 5,328, and NASDAQ with 6,554), one commodity (Crude Oil with 6,383), one real estate index (Real Estate Exchange-Traded Fund (ETF) with 5,364), one transportation dataset (Traffic with 48,204), and two utilities (Energy with 7,267 and Electricity with 17,520). This collection spans stable utility demand through volatile cryptocurrency markets, ensuring that findings generalize across domains.

➢ *Forecasting Methods*

We evaluate 11 methods in three categories. *Baselines* (3): Moving Average computes rolling window means, Exponential Smoothing applies exponentially decreasing weights, and Seasonal Naive uses same season values from the previous cycle. *Classical statistical* (3): ARIMA [7] with parameter selection based on autocorrelation, Error, Trend, Seasonality (ETS) [8] with automatic model selection, and Prophet [9] combining additive decomposition with changepoint detection. *Machine learning* (5): ElasticNet [11] with L1/L2 regularization, Random Forest [13], XGBoost [14], LightGBM [15], and CatBoost [16]. All machine learning methods use features derived from lagged values (1–7 steps), rolling statistics (seven-day mean and standard deviation), and temporal encodings (day of week, month, and quarter).

➢ *Dual Evaluation Protocol*

Our core innovation compares identical test sets under two training allocations. Protocol 1 (60/20/20): 60% training, 20% validation, and 20% test, representing limited data scenarios arising from collection costs, privacy constraints, or phenomenon recency. Protocol 2 (70/15/15): 70% training, 15% validation, and 15% test, following recommendations from Cerqueira et al. [22] and Hyndman and Athanasopoulos [23]. The 10 percentage point difference provides approximately 17% more training samples. Both protocols evaluate on identical test periods, ensuring that only training data quantity differs.

➢ *Evaluation Metrics and Ranking Framework*

We use five complementary metrics to ensure robust assessment. Mean Absolute Percentage Error (MAPE) provides scale-independent percentage error suitable for cross-dataset comparison, although it penalizes underforecasts asymmetrically. Symmetric Mean Absolute Percentage Error (SMAPE) addresses this limitation with a symmetric denominator. Root Mean Square Error (RMSE) penalizes large errors more heavily through squaring, whereas Mean Absolute Error (MAE) provides interpretable absolute error robust to outliers. Mean Absolute Scaled Error (MASE) scales errors relative to a naive seasonal baseline, enabling meaningful cross-dataset comparison.

For each dataset d and metric k, methods receive ranks from 1 (best) to 11 (worst). Average ranks are computed as:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep096

$$\text{Average}_m = \frac{1}{|M| \times |D|} \sum_{d \in D} \sum_{k \in M} \text{Rankrank}_{m,d,k}$$

Where $|M| = 5$ metrics and $|D| = 11$ datasets, yielding 55 individual rankings per method. This framework reduces sensitivity to any single metric or dataset.

➢ *Plain Language Summary*

The amount of training data available determines which forecasting method works best. When historical data is scarce (below 65% of the available series), simple approaches such as exponential smoothing avoid overfitting and produce more reliable predictions. When sufficient data is available (70% or more), sophisticated methods such as Prophet and LightGBM estimate complex temporal patterns more reliably and outperform simpler alternatives. This study systematically demonstrates this relationship, offering practical guidance that resolves long-standing disagreements in the forecasting literature.

III. RESULTS

➢ *Overall Method Rankings*

Table I presents our central finding: method rankings change substantially between limited data (60%) and adequate data (70%) conditions.

Under Protocol 1, Exponential Smoothing achieves the best average rank (4.80), demonstrating that simple methods excel when data is limited. Under Protocol 2, Prophet dominates (4.47), followed by LightGBM (4.75) and ETS (5.02), confirming that sophisticated methods outperform baselines when adequate data enables reliable parameter estimation.

Table 1 Method Rankings Under Two Training Protocols

Method	60/20/20	70/15/15	Change
Prophet	5.45	4.47	+0.98
LightGBM	6.31	4.75	+1.56
ETS	6.76	5.02	+1.75
ARIMA	5.87	5.80	+0.07
Moving Avg	5.44	5.84	−0.40
CatBoost	5.04	5.84	−0.80
Exp Smooth	4.80	6.04	−1.24
XGBoost	6.27	6.18	+0.09
Random Forest	6.02	6.71	−0.69
Seasonal Naive	6.35	7.05	−0.71
ElasticNet	7.69	8.31	−0.62

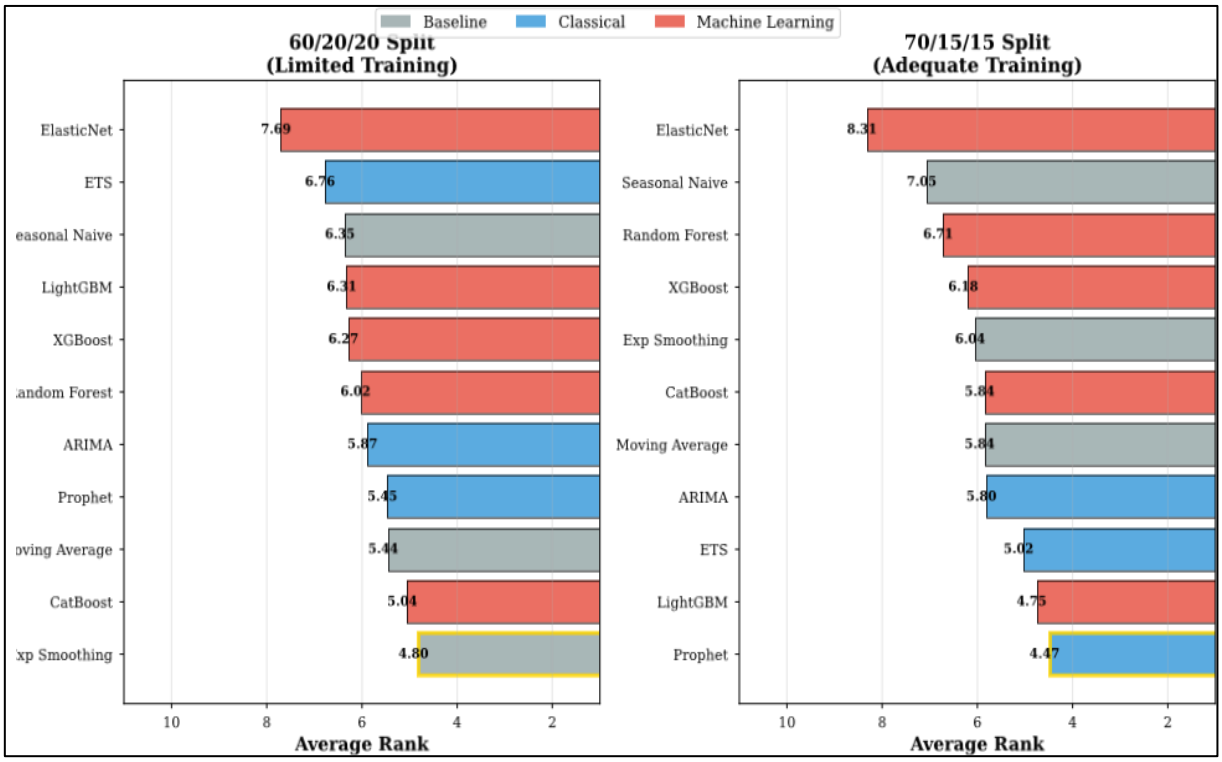

Fig 1 Rankings Under 60/20/20 (Left) and 70/15/15 (Right) Protocols.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep096

Among individual methods, LightGBM improves by 1.56 positions (from 6.31 to 4.75), indicating that gradient boosting effectively leverages additional data to model complex patterns. ETS shows the largest improvement at 1.75 positions (from 6.76 to 5.02), because longer training histories enable better seasonal parameter estimation. Conversely, Exponential Smoothing drops by 1.24 positions (from 4.80 to 6.04), Moving Average declines by 0.40 positions, and Seasonal Naive falls by 0.71 positions. These simple methods cannot exploit additional data for improved modeling complexity. XGBoost (+0.09) and ARIMA (+0.07) remain notably stable across protocols. ElasticNet ranks last under both protocols (7.69 and 8.31), confirming that linear models cannot capture nonlinear temporal patterns regardless of data quantity.

Fig. 1 visualizes these changes, illustrating the transition from simple method dominance to sophisticated method superiority.

➢ *Method Movement Patterns*

Fig. 2 reveals three distinct movement patterns: *improvers* (Prophet at +0.98, LightGBM at +1.56, and ETS at +1.75) whose rankings improve with additional data; *decliners* (Exponential Smoothing at −1.24, CatBoost at −0.80, Seasonal Naive at −0.71, Random Forest at −0.69, ElasticNet at −0.62, and Moving Average at −0.40) whose rankings worsen; and *stable methods* (XGBoost at +0.09 and ARIMA at +0.07) with minimal change.

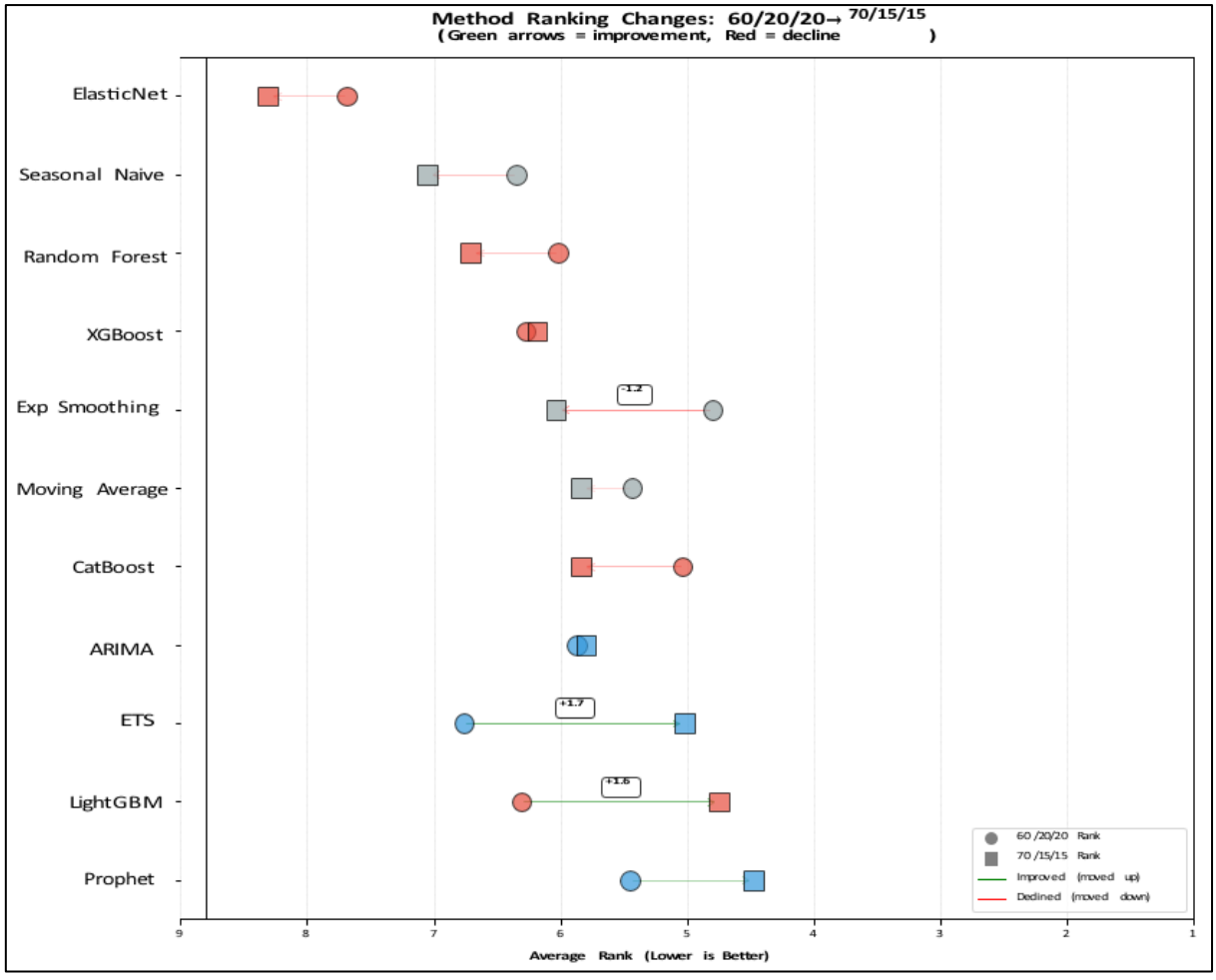

Fig 2 Ranking Changes as Training Data Increases From 60% to 70%.

These patterns offer actionable guidance. Methods that improve with additional data (Prophet, LightGBM, and ETS) should be preferred when practitioners can collect more observations, whereas methods that perform better with limited data (Exponential Smoothing and Moving Average) suit constrained data applications.

➢ *Per-Dataset Performance*

Fig. 3 presents per-dataset heatmaps confirming that ranking changes occur consistently across diverse data characteristics rather than as artifacts of specific datasets. Under Protocol 2, Prophet ranks first on four datasets (Traffic, Energy, Electricity, and Real Estate ETF), demonstrating particular strength on seasonal data with trend changes. LightGBM excels on volatile financial data (Bitcoin and Ethereum), where nonlinear modeling captures complex price dynamics. ETS performs best on stable series (Gold and S&P 500), where state-space models effectively estimate smooth components. Fig. 4 further confirms that ranking shifts occur across the majority of datasets.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep096

➢ *Category Analysis and Validation*

Fig. 5 shows forecasts from three representative datasets, illustrating method-specific strengths across different data patterns. Fig. 6 aggregates results by category: classical methods improve most (by +0.93 positions, from 6.03 to 5.10), baselines decline (by −0.78, from 5.53 to 6.31), and machine learning methods show minimal aggregate change (by −0.09, from 6.27 to 6.36). However, the machine learning stability masks divergent within-category behavior, as LightGBM improves substantially whereas CatBoost, Random Forest, and ElasticNet decline.

Fig. 7 validates these findings across all five metrics: Prophet leads on percentage-based errors (MAPE and

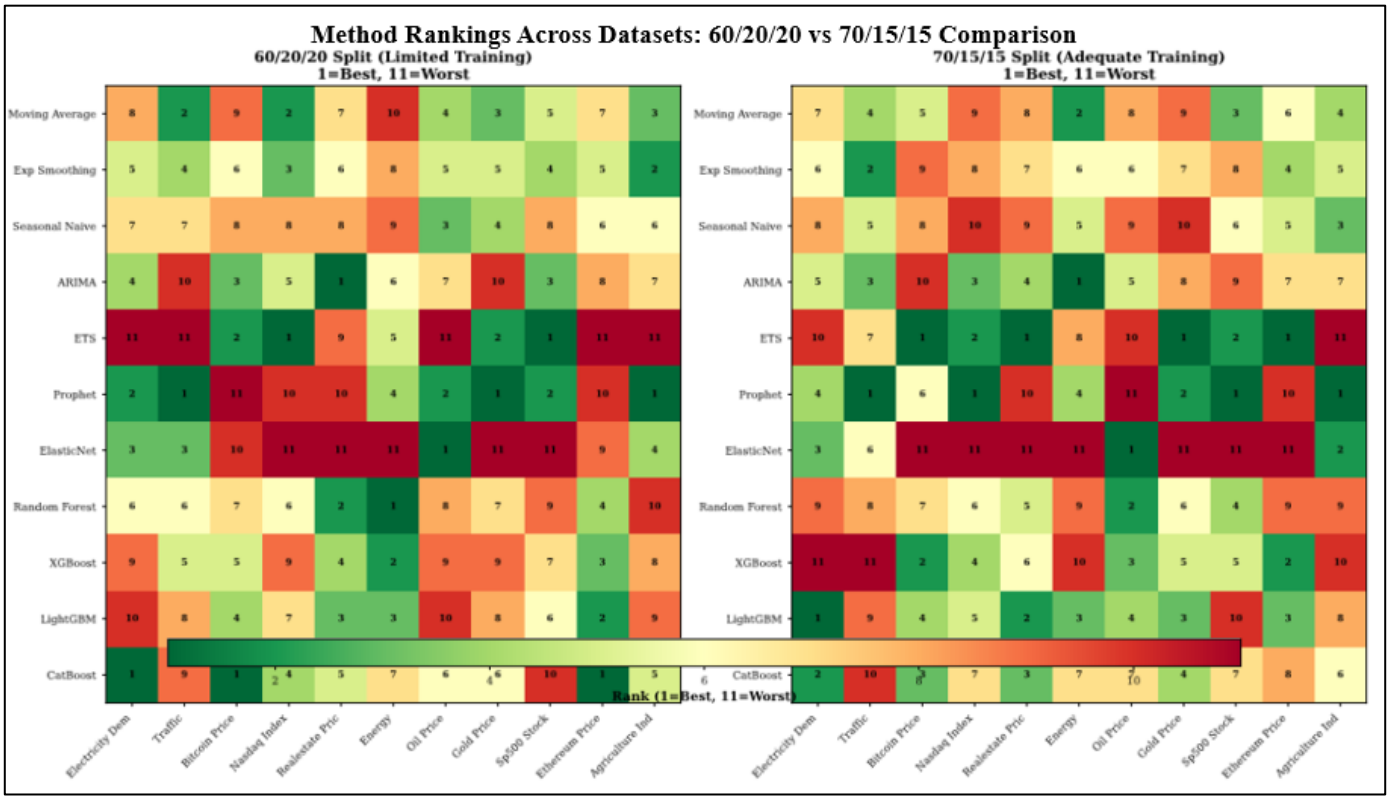

Fig 3 Per-Dataset Rankings for Both Protocols Across All Domains.

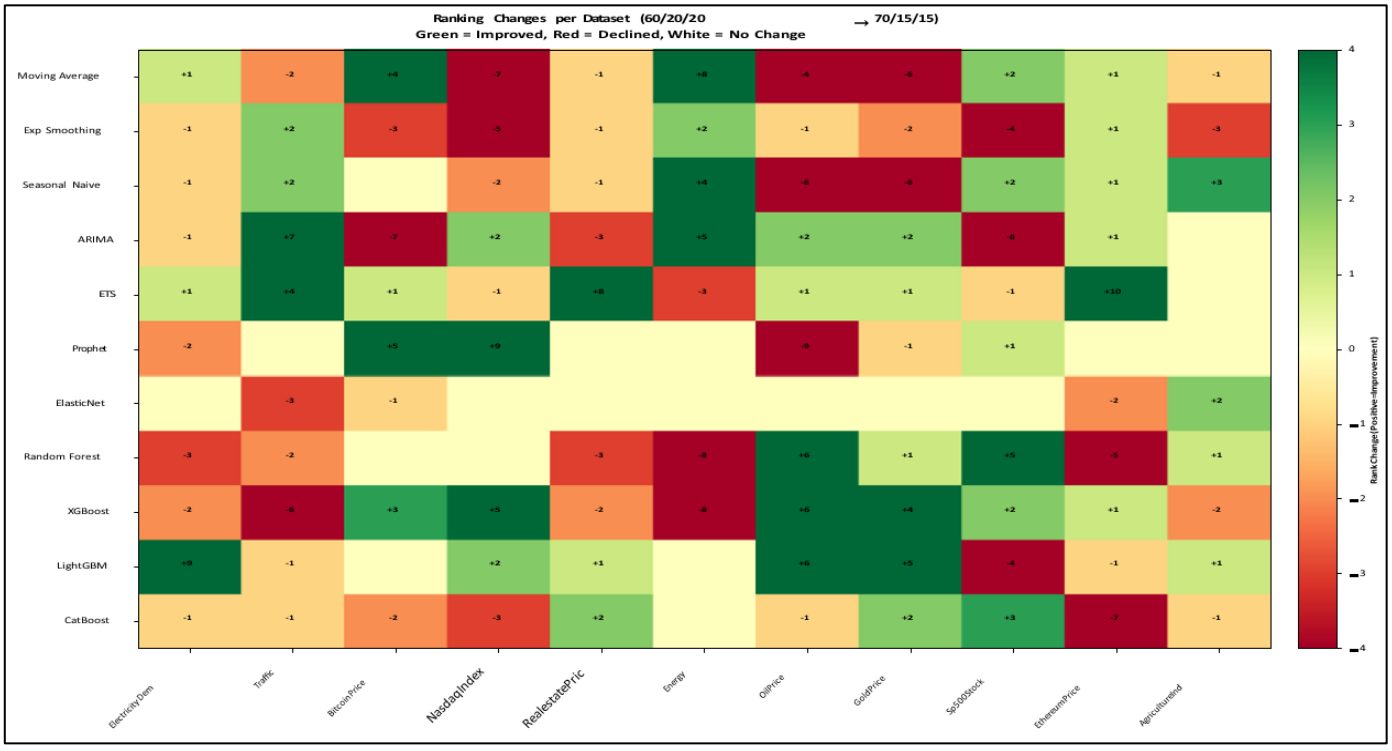

Fig 4 Ranking Changes Per Dataset Confirming Widespread Effects.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep096

SMAPE), whereas LightGBM excels on absolute-error metrics (RMSE and MAE). Fig. 8 reveals computational tradeoffs: gradient boosting requires 30–120 seconds, with training times increasing by 32–38% from Protocol 1 to Protocol 2, whereas Prophet requires 10–30 seconds and ElasticNet trains instantaneously. These costs suggest that practitioners with computational constraints may prefer classical methods even when more data is available.

IV. DISCUSSIONS

➢ *Mechanisms Explaining Ranking Changes*

Three complementary mechanisms explain the observed ranking transitions. First, *parameter estimation requirements*: classical methods such as Prophet, ETS, and ARIMA estimate multiple parameters for trend, seasonality, and changepoints. With 60% training data, insufficient observations exist for robust estimation, particularly for seasonal components that require multiple complete cycles. At 70%, longer histories provide adequate observations for accurate estimation.

Second, *gradient boosting sample requirements*: LightGBM and XGBoost construct thousands of decision trees, and each split requires sufficient samples in both branches. With limited data, insufficient samples at deeper levels force early stopping and underutilize model capacity. Additional data enables deeper, more reliable splits while regularization prevents overfitting.

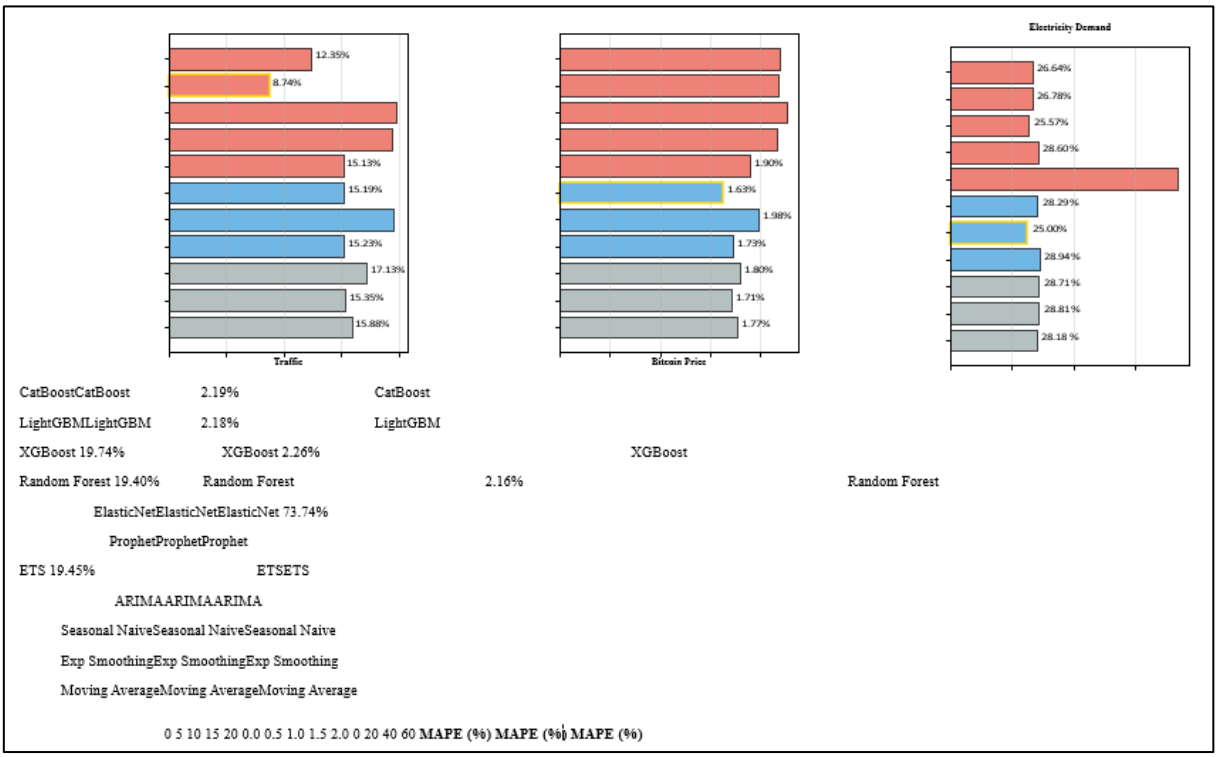

Fig 5 Forecasts on Electricity, Traffic, and Bitcoin Datasets.

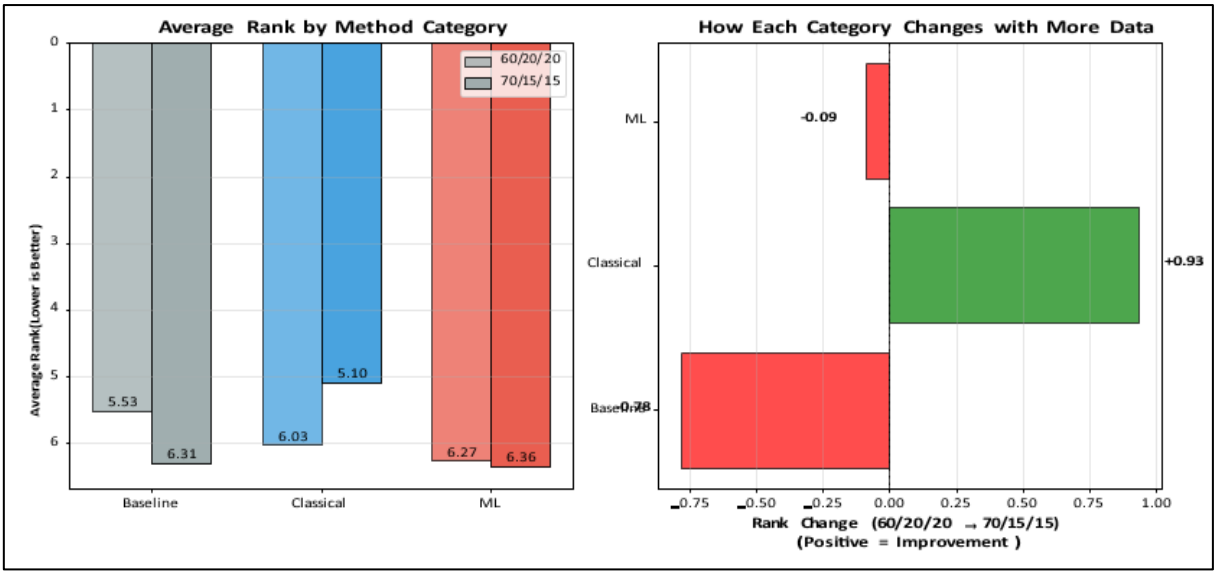

Fig 6 Category-Level Performance Comparison Between Protocols.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep096

Third, *bias–variance dynamics*: simple methods exhibit high bias but low variance. When data is scarce, the advantage of low variance dominates because overfitting risks outweigh approximation errors. With adequate data, sophisticated methods reduce bias through flexible modeling while controlling variance through regularization, thereby shifting the optimal trade-off toward lower-bias models.

➢ *Reconciling Contradictory Literature*

Our findings resolve apparent contradictions in the forecasting literature, where different studies reach opposing conclusions about method superiority.

Studies claiming that simple methods remain competitive [5], [20] likely used evaluation protocols with limited training data relative to series length. Under such conditions, our results confirm that simple methods do indeed excel. Conversely, research demonstrating sophisticated method superiority [6] likely provided adequate training data, where our findings confirm that complex methods outperform baselines.

Both sets of conclusions are correct within their respective experimental contexts. The contradiction arises from failing to account for training data availability as a moderating variable. Our dual-protocol evaluation demonstrates that both simple and sophisticated methods can be optimal, depending on data availability.

➢ *Practical Guidelines*

Addressing RQ3, our analysis reveals a critical training data threshold between 65% and 70% of the available time series. Below this threshold, simple statistical methods provide superior performance. Above it, sophisticated methods excel. The transition zone of 65–70% exhibits high ranking volatility, making method selection particularly uncertain.

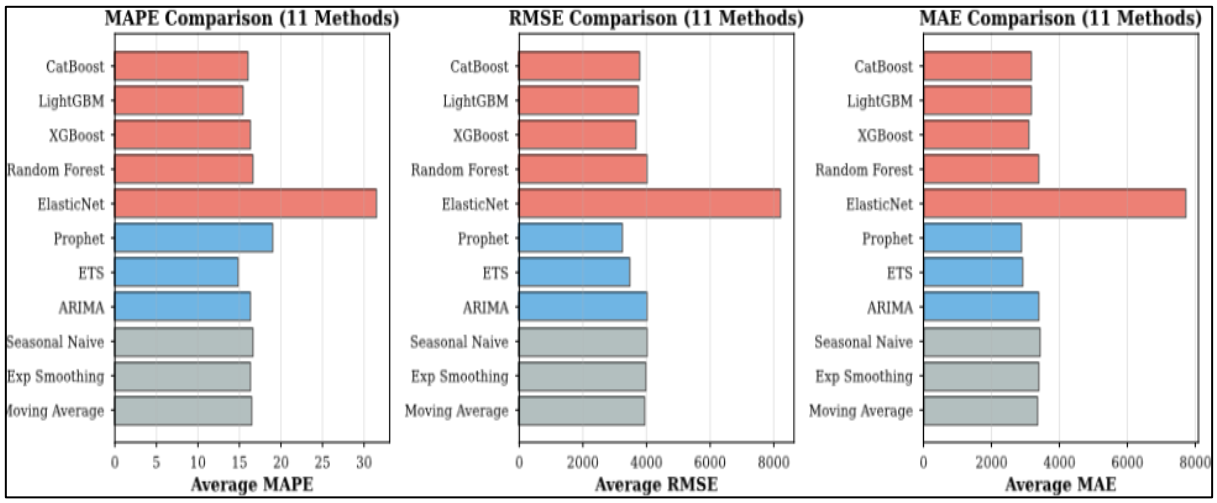

Fig 7 Rankings Across all Five Metrics Confirming Consistency.

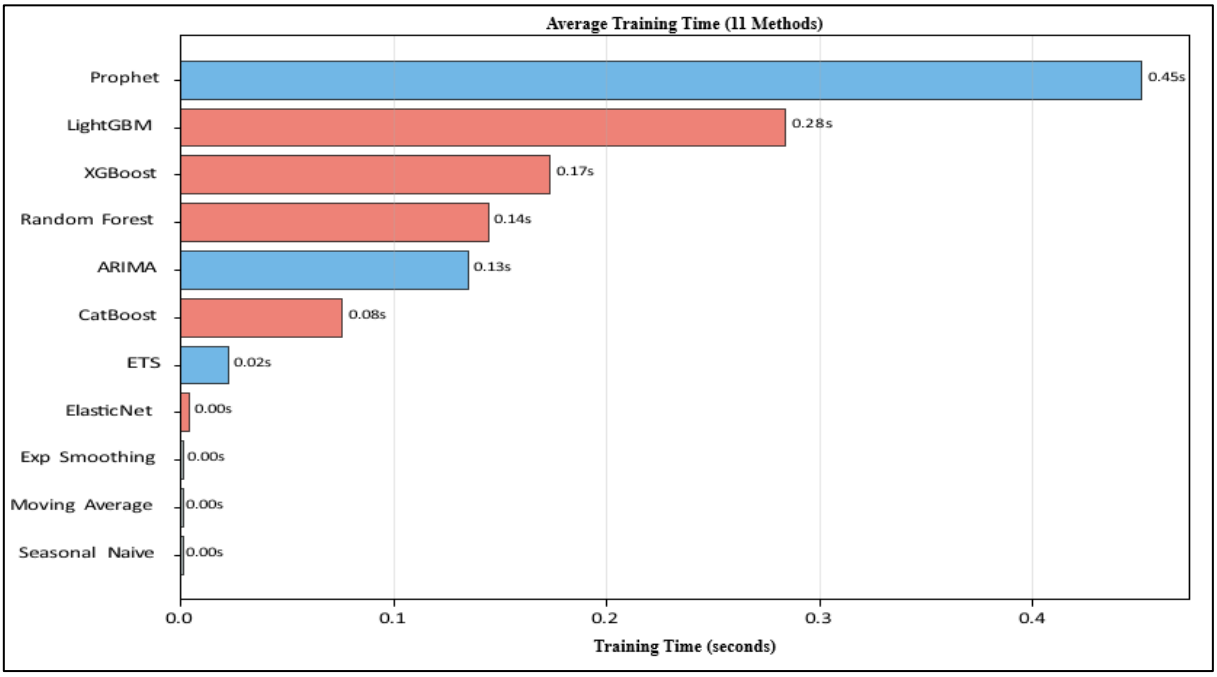

Fig 8 Training Time Comparison Across all Methods.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep096

We propose six deployment scenarios. Scenario 1 (limited data, below 65%): use Exponential Smoothing or Moving Average for new products, new sensors, or privacy-restricted data. Scenario 2 (adequate data with strong seasonality, 70% or above): use Prophet for retail, web traffic, or energy forecasting. Scenario 3 (adequate data with high volatility, 70% or above): use LightGBM for cryptocurrency or high-frequency data. Scenario 4 (adequate data with stable trends, 70% or above): use ETS for gold, utilities, or mature indices. Scenario 5 (uncertain data availability): use ARIMA, which maintains stable performance across both protocols (ranks of 5.87 and 5.80). Scenario 6 (computational constraints): prefer classical methods, as gradient boosting training times increase by 32–38% with additional data.

➢ *Limitations*

This study has several limitations. We evaluate only two protocols; future work should vary training percentages in finer increments from 50% to 90% to map the complete relationship between data availability and method selection. We exclude deep learning architectures (N-BEATS, Informer, and FEDformer) due to computational costs and inconsistent performance evidence. Our focus on single-step forecasting means that multi-horizon settings may exhibit different data requirements. We employ static splits; longitudinal evaluations tracking performance as training data accumulates would provide complementary insights. Finally, expanding beyond 11 datasets to hundreds across additional industries would strengthen generalizability.

V. CONCLUSION

This study demonstrates that training data size fundamentally determines optimal forecasting method selection. Through dual-protocol evaluation of 11 methods across 11 datasets, we establish that method rankings change substantially as training data increases from 60% to 70%.

Our findings directly address all three research questions. RQ1: Rankings change substantially; Exponential Smoothing ranks best with 60% training, whereas Prophet dominates with 70% training. RQ2: LightGBM improves by 1.56 positions and ETS improves by 1.75 positions, whereas simple baselines decline significantly (Exponential Smoothing by −1.24 and Seasonal Naive by −0.71). RQ3: A critical threshold exists between 65% and 70% training data, above which sophisticated methods surpass baselines.

These findings carry immediate practical implications. Method selection should explicitly consider available training data size, not merely problem domain or computational constraints. With limited data (below 65%), simple methods provide superior performance by avoiding overfitting. With adequate data (70% or above), sophisticated methods excel by reliably estimating complex patterns. This perspective shifts forecasting methodology from seeking universally best methods toward understanding when each method type excels.

Future work should vary training percentages in finer increments, incorporate deep learning methods, examine multi-horizon forecasting, and conduct longitudinal evaluations to refine method selection guidance based on data availability.

AUTHOR CONTRIBUTIONS

T. Vangalapat: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization. P. Sharma: Methodology, Formal Analysis, Investigation, Validation, Writing – Review & Editing, Supervision. S. Banerjee: Methodology, Validation, Writing – Review & Editing, Supervision.

➢ *Conflict of Interest*

The authors declare no competing financial interests or personal relationships that could have influenced this work.

➢ *Data Availability*

All datasets used in this study are publicly available. Financial data were obtained from Yahoo Finance. Traffic data were obtained from the UCI Machine Learning Repository. Energy and Electricity data were obtained from government sources. Code and processed datasets will be made available upon publication.

➢ *Ethics Statement*

This research uses only publicly available datasets and does not involve human subjects. No ethics approval was required.

➢ *Funding Statement*

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

[1]. O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, “Financial time series forecasting with deep learning: A systematic literature review: 2005–2019,” *Appl. Soft Comput.*, vol. 90, p. 106181, 2020.

[2]. V. K. R. Chimmula and L. Zhang, “Time series forecasting of COVID-19 transmission in Canada using LSTM networks,” *Chaos Solitons Fractals*, vol. 135, p. 109864, 2020.

[3]. T. Ahmad, H. Zhang, and B. Yan, “A review on renewable energy and electricity requirement forecasting models for smart grid and buildings,” *Sustain. Cities Soc.*, vol. 55, p. 102052, 2020.

[4]. Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Wang, “Traffic flow prediction with big data: A deep learning approach,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, 2015.

[5]. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and machine learning forecasting methods: Concerns and ways forward,” *PLoS One*, vol. 13, no. 3, p. e0194889, 2018.

[6]. H. Zhou *et al.*, “Informer: Beyond efficient transformer for long sequence time-series forecasting,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 11106–11115.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep096

[7]. G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ: Wiley, 2015.
[8]. R. J. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder, *Forecasting with Exponential Smoothing: The State Space Approach*. Berlin: Springer, 2008.
[9]. S. J. Taylor and B. Letham, "Forecasting at scale," *Amer. Statist.*, vol. 72, no. 1, pp. 37–45, 2018.
[10]. O. Triebe *et al.*, "NeuralProphet: Explainable forecasting at scale," *arXiv preprint arXiv:2111.15397*, 2021.
[11]. H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. B*, vol. 67, no. 2, pp. 301–320, 2005.
[12]. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York: Springer, 2009.
[13]. L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[14]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 785–794.
[15]. G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3146–3154.
[16]. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 6638–6648.
[17]. B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," in *Int. Conf. Learn. Represent.*, 2020.
[18]. T. Zhou *et al.*, "FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *Proc. 39th ICML*, 2022, pp. 27268–27286.
[19]. Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *Int. Conf. Learn. Represent.*, 2023.
[20]. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: 100,000 time series and 61 forecasting methods," *Int. J. Forecast.*, vol. 36, no. 1, pp. 54–74, 2020.
[21]. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," *Int. J. Forecast.*, vol. 38, no. 4, pp. 1346–1364, 2022.
[22]. V. Cerqueira, L. Torgo, and I. Mozetic, "Evaluating time series forecast-ˇ ing models: An empirical study on performance estimation methods," *Mach. Learn.*, vol. 109, no. 11, pp. 1997–2028, 2020.
[23]. R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. Melbourne: OTexts, 2018.