Journal of Data Processing

DLINE JOURNALS

Print ISSN: 2278 – 6481
Online ISSN: 2278 – 649X

JDP 2026: 16 (1)
https://doi.org/10.6025/jdp/2026/16/1/31-45

Bridging Tabular Foundation Models and Commodity Price Forecasting: Toward Scalable, Privacy-Aware Data Science for Web-Generated Economic Data

P Paramasivaiah
Department of Studies and Research in Commerce
Tumkur University, TUMKUR-572 103. India
paramashivaiah@gmail.com

ABSTRACT

This paper explores the intersection of two critical frontiers in data science: the rise of foundation models for tabular data and their practical application to forecasting in complex, real world domains specifically global commodity markets. The study leverages the World Bank's Pink Sheet dataset (1960–2025) to analyze long term price dynamics and evaluate forecasting methodologies across energy, metal, and agricultural commodities. Empirical results show that traditional time series models like SARIMAX excel in stable, seasonal contexts, while machine learning approaches particularly Random Forest better capture nonlinearities and structural breaks during periods of high volatility, such as those observed during the 2020 pandemic shock.

In parallel, the paper reviews recent breakthroughs in tabular foundation models, including TabPFN, TabICL, and TabFM, which challenge the longstanding dominance of gradient boosted trees by enabling in context learning, zero shot generalization, and instruction following capabilities. These models leverage synthetic pretraining and novel architectures to achieve competitive or superior performance on small to medium datasets. Despite these advances, the potential of foundation models for tabular data distillation compressing large datasets while preserving statistical fidelity remains underexplored. The authors argue that extending foundation models beyond prediction to tasks like dataset compression, synthetic generation, and privacy-preserving representation could address key challenges in web science, including scalability, reproducibility, and data sharing.

The paper concludes by advocating a dual research trajectory: advancing neural architectures tailored to tabular structure and broadening their scope from predictive modeling to data centric infrastructure. This integrated vision promises more efficient, generalizable, and ethically robust tools for managing the growing complexity of real world tabular data.

Keywords: Tabular Foundation Models, Commodity Price Forecasting, In-Context Learning, Data Distillation, Random Forest, SARIMAX, Web Science, Synthetic Data Generation

Received: 28 August 2025, Revised 30 November 2025, Accepted 8 December 2025

1. Introduction

The rapid expansion of web science has generated massive, heterogeneous datasets from diverse online platforms. While such data enable large scale analysis and knowledge discovery, they simultaneously introduce substantial challenges related to computational efficiency, scalability, privacy preservation, and reproducibility. Processing and storing these datasets often exceed practical computational limits, particularly in resource constrained research environments. As a result, there is a growing demand for principled methods that can reduce data volume while preserving essential statistical and semantic properties.

In parallel, Foundation Models (FMs) have emerged as a transformative paradigm in machine learning, driving major advances in natural language processing and computer vision through large scale architectures such as GPT and DALL·E. Their success is attributed to scaling laws and emergent capabilities, including zero shot learning, in context generalization, and cross task transferability. Despite these advances, extending foundation model principles to tabular data, one of the most prevalent and practically important data modalities in web science, has only recently gained research momentum.

2. Early Research

2.1 Tabular Data and the Limits of Traditional Learning

Tabular data, typically organized as structured rows and columns, underpins decision making across numerous scientific and industrial domains, including biomedicine, particle physics, economics, and climate science [1, 2]. Core tasks such as learning a target variable from heterogeneous feature columns are central to applications ranging from biomedical risk assessment to materials discovery.

Despite the success of deep learning in unstructured domains, it has achieved landmark breakthroughs in game playing, protein folding, and large scale language modelling [3, 4, 5]. Tabular learning has remained dominated by gradient boosted decision trees (GBDTs) for nearly two decades. Methods such as *XGBoost*, *LightGBM*, and *CatBoost* have consistently demonstrated strong empirical performance due to their robustness, interpretability, and ability to handle heterogeneous feature types [6, 7, 8, 9].

This persistent dominance highlights a broader imbalance in modality focused research. Recent position papers argue that tabular data, despite being the dominant modality in many scientific workflows, receives disproportionately less attention than text and vision [10]. This gap has motivated renewed interest in developing scalable and generalizable neural approaches tailored to tabular data.

2.2 Foundation Models for Tabular Data

Recent work has begun to demonstrate that foundation style pretraining is feasible for tabular data. *TabPFN* [11] pioneered this direction by training neural networks on large collections of synthetically generated tabular datasets designed to approximate real world distributions. Rather than relying on a single massive dataset, *TabPFN* learns a broad prior over data generating processes, enabling strong performance on previously unseen tasks with minimal or no fine tuning.

This paradigm was further validated by *TabPFNv2* [12], which showed that synthetically learned priors can enable neural models to outperform leading GBDT implementations such as *XGBoost* and *LightGBM* [7, 8], particularly on small and medium sized datasets. These findings establish tabular foundation models as a viable and competitive alternative to traditional tree based methods, extending their applicability beyond conventional supervised learning settings.

2.3 Transformer Architectures and Tabular Learning

Transformer architectures have become the backbone of modern foundation models due to their ability to capture long range dependencies through attention mechanisms [13]. Originally designed for sequential data, transformers have been adapted for tabular learning with mixed success [14, 15]. A key limitation lies in treating tabular inputs as flat sequences, which neglects the intrinsic row column structure and relational properties of tables.

TabPFN represents a significant departure from this paradigm by reframing tabular learning as probabilistic inference conditioned on the training data itself, enabling in-context learning without gradient updates. In addition to predictive inference, *TabPFN* supports density estimation, reusable embeddings, and fine tuning. However, scalability remains a challenge, as attention based mechanisms become computationally expensive for large tables.

2.4 In-Context Learning and Scalability in Tabular Models

In-context learning (ICL) has emerged as a central mechanism for tabular foundation models, allowing dynamic task adaptation at inference time without parameter updates. TabDPT integrates ICL based retrieval with self-supervised learning and demonstrates that incorporating real world tabular data during pretraining leads to faster convergence and improved generalization compared to purely synthetic data [16]. The model achieves state of the art performance across both regression and classification benchmarks.

To address scalability constraints, *TabICL* introduces a two stage architecture that decouples column wise and row wise attention, enabling efficient processing of large datasets[17]. Pretrained on synthetic tables with up to 60,000 samples and scalable to 500,000 samples at inference, *TabICL* achieves competitive accuracy with TabPFNv2 while offering substantial computational speedups.

2.5 Generative and Instruction-Following Tabular Models

Beyond predictive and ICL-based approaches, recent research has explored generative tabular foundation models built on large language model backbones. TabFM adopts a generative learning framework by fine tuning pretrained LLMs on diverse tabular datasets using specialized objectives [18]. This approach enables instruction following behavior, zero-shot inference, and in context learning, while achieving performance comparable to or exceeding closed-source models such as GPT-4.

Importantly, *TabFM* demonstrates strong data efficiency under limited supervision, highlighting the practical potential of generative tabular models for real world analytical tasks where labeled data are scarce.

2.6 Toward Foundation Models for Tabular Data Distillation

Despite substantial progress in tabular foundation models, existing approaches primarily emphasize predictive performance. Their potential for data distillation, dataset compression, and privacy aware data sharing remains largely unexplored. Data distillation offers a compelling solution by condensing large datasets into smaller, representative subsets that preserve core statistical properties, enabling efficient computation and storage.

Prior work has shown that autoencoders and clustering based methods can effectively distill tabular datasets, producing compact representations that retain key distributions and relationships [19]. However, these approaches remain task or dataset specific. A foundation model explicitly designed for tabular data distillation could generalize across domains and schemas, enabling consistent data reduction, synthetic data generation, privacy preservation, and reproducibility.

By extending the foundation model paradigm from prediction to data distillation, this research direction addresses a critical gap in web science infrastructure. It aligns with emerging trends in tabular deep learning [20] and offers a path toward scalable, privacy aware, and reusable data representations that can handle the growing volume and complexity of web generated tabular data.

2.7 Advances in Tabular Deep Learning

Deep learning for tabular data has undergone rapid development in recent years, driven by efforts to close the performance gap between neural approaches and traditional methods such as gradient-boosted decision trees (GBDTs). Several studies have proposed architectures that incorporate attention mechanisms, feature embeddings, and inductive biases tailored to tabular structures [21, 22, 23, 24]. These advances frequently draw inspiration from natural language processing and computer vision, where large-scale pretraining has proven essential for achieving strong generalization and sample efficiency.

In parallel, foundation models large neural architectures pretrained on extensive datasets and adaptable to diverse downstream tasks have become central to progress in NLP and CV ([25, 26]. However, applying this paradigm to tabular data has long been considered challenging due to the heterogeneous nature of tabular features, varying schemas across datasets, and the lack of large, publicly available corpora suitable for pretraining.

3. Methodology

This study adopts a mixed methodological framework combining tabular data analysis, time series modeling, and machine learning to examine long run dynamics and forecast commodity prices using the World Bank Commodity Price Data (The Pink Sheet). The dataset consists of monthly nominal price indices (base year 2010=100) covering the period 1960 to January 2025. Given the large scale and heterogeneity of the dataset, the analysis focuses on a representative subset of key commodities, with particular emphasis on Brent crude oil to ensure analytical tractability.

Data preprocessing involved tabulation, temporal alignment, and normalization to ensure consistency across commodity series. Exploratory analysis was conducted using trend plots to examine long run price dynamics and heatmap based correlation matrices to assess interdependencies among commodity groups over the post-2000 period. These visual diagnostics informed model selection and highlighted structural breaks, volatility clustering, and cross market co-movement.

For forecasting, three complementary approaches were employed. First, Exponential Smoothing State Space

(ETS) models were applied to capture level, trend, and seasonal components without requiring stationarity. Model configurations were selected automatically using information criteria, and parameters were estimated via maximum likelihood. Second, a Seasonal ARIMA with exogenous variables (SARIMAX) framework was used to model autocorrelation, seasonality, and potential external influences, with forecast uncertainty quantified through prediction intervals. Third, a machine learning approach based on Random Forest regression was implemented using lagged price features and calendar variables to capture nonlinear dynamics and structural shifts.

Model performance was evaluated using out of sample error metrics, enabling comparative assessment across statistical and machine learning paradigms.

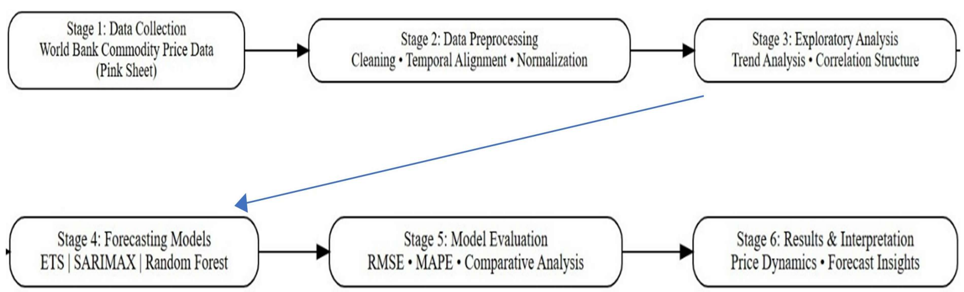

Figure 1. Methodology workflow illustrating data acquisition, preprocessing, exploratory analysis, forecasting model implementation, and evaluation for commodity price analysis

3.1 Dataset Description

This study utilizes the World Bank Commodity Price Data, commonly referred to as The Pink Sheet, obtained from the file *CMO Historical Data Monthly.xlsx*. The dataset provides a comprehensive and authoritative record of global commodity prices and indices, making it a widely used reference source for economic analysis, market monitoring, and policy research.

The dataset is reported monthly and spans a long historical period from 1960 to January 2025, with the most recent update released on January 3, 2025. Commodity prices are expressed as nominal indices denominated in U.S. dollars, normalized to a base year of 2010 = 100, allowing for consistent comparison across time and commodity categories. The data are tabular, with rows representing time periods and columns corresponding to individual commodity prices or aggregated commodity group indices.

The Pink Sheet covers a broad range of commodity classes, including energy, metals, and agricultural products, thereby capturing diverse market dynamics and sector specific price behavior. Its long temporal coverage enables the analysis of structural breaks, cyclical patterns, and responses to major global events such as financial crises, geopolitical disruptions, and climate related shocks.

In this study, the dataset serves as the primary empirical foundation for analyzing long run commodity price dynamics, inter commodity correlations, and forecasting performance. Given the extensive size and heterogeneity of the data, the empirical analysis focuses on a representative subset of key commodities, with particular emphasis on Brent crude oil, to ensure analytical clarity while preserving the dataset's broader economic relevance.

4. Analysis used in this Study: Tabulation of Data

4.1 Analytical Methods

The analysis conducted in this study is based on systematic tabulation and exploratory visualization of the commodity price data. Descriptive tabulation is first employed to organise and summarise the monthly time-series observations across commodity groups, enabling consistent comparison over time.

To examine interdependencies and co-movement patterns among commodities, visual analytics and dashboard-based representations are utilized. In particular, heatmap visualizations are constructed to depict pairwise correlation structures and relative volatility across commodity classes, facilitating the identification of clustered behaviors and shared market drivers.

For temporal dynamics, the study applies time series analysis and forecasting techniques to model and predict future commodity price movements. Three complementary forecasting paradigms are employed: traditional autoregressive integrated moving average (ARIMA) models, Exponential Smoothing State Space (ETS) methods to capture trend and seasonality, and machine learning based approaches to account for nonlinear patterns and structural shifts. These models enable comparative evaluation of statistical and data driven forecasting performance.

Given the extensive scale and dimensionality of the World Bank commodity dataset, the empirical analysis is intentionally focused on a representative case study of Brent crude oil prices. This targeted approach ensures analytical tractability while retaining substantive relevance, as Brent crude oil plays a central role in global commodity markets and serves as a key benchmark for energy price dynamics.

4.2 Trend Plots for Key Commodities (recent 25 years)

Figure 2 presents long run price dynamics (2000–2024) across major energy, metal, and agricultural commodities, revealing pronounced cyclicality, structural breaks, and synchronized shocks. Energy prices exhibit strong volatility: crude oil rises steadily until the mid 2000s, peaks around 2008, collapses during the global financial crisis, rebounds in the early 2010s, falls sharply in 2020, and resurges post pandemic. Natural gas shows higher short term volatility with episodic spikes, while coal prices remain relatively stable until sharp surges after 2021.

Base metals (copper and aluminium) display pro cyclical behaviour, peaking around 2007–2008 and again post 2020. Gold follows a distinct upward long term trajectory, reflecting its role as a safe-haven asset, with notable appreciation during crisis periods.

Agricultural commodities show heterogeneous patterns. Wheat, maize, rice, soybeans, palm oil, and sugar experience episodic spikes linked to supply shocks, biofuel policies, and geopolitical disruptions, particularly during 2008–2012 and 2021–2023. Soft commodities such as coffee and cocoa have experienced strong recent price increases, with cocoa exhibiting an exceptional surge toward the end of the sample.

Overall, the figure highlights increasing price volatility and stronger co movement across commodities in recent years, suggesting heightened exposure to global shocks, climate risks, and geopolitical uncertainty.

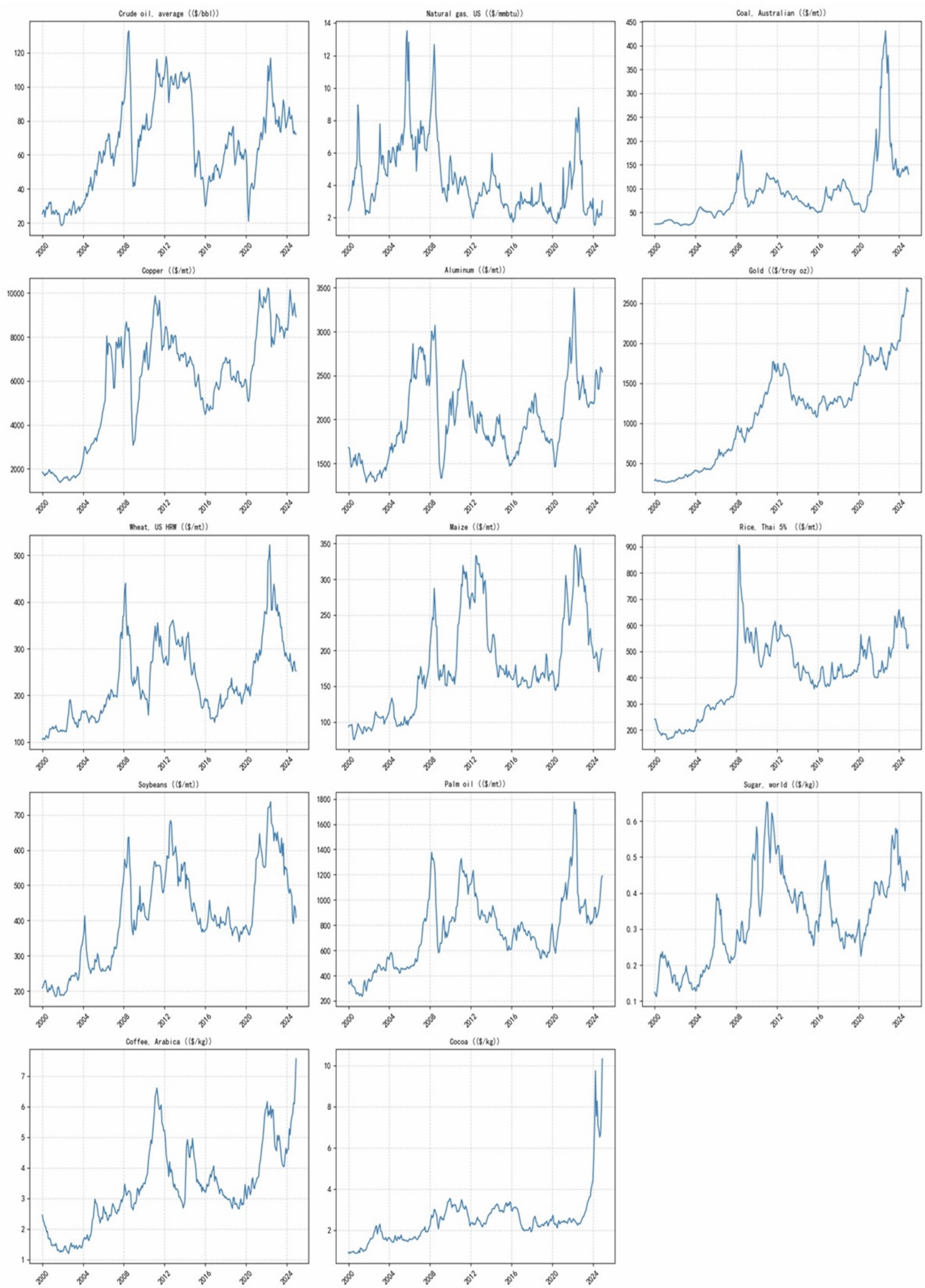

Figure 2. Long-run price dynamics based on world bank dataset

Now, we compute and visualize the correlation matrix among these key series (using data from 2000 onward to ensure reasonable overlap):

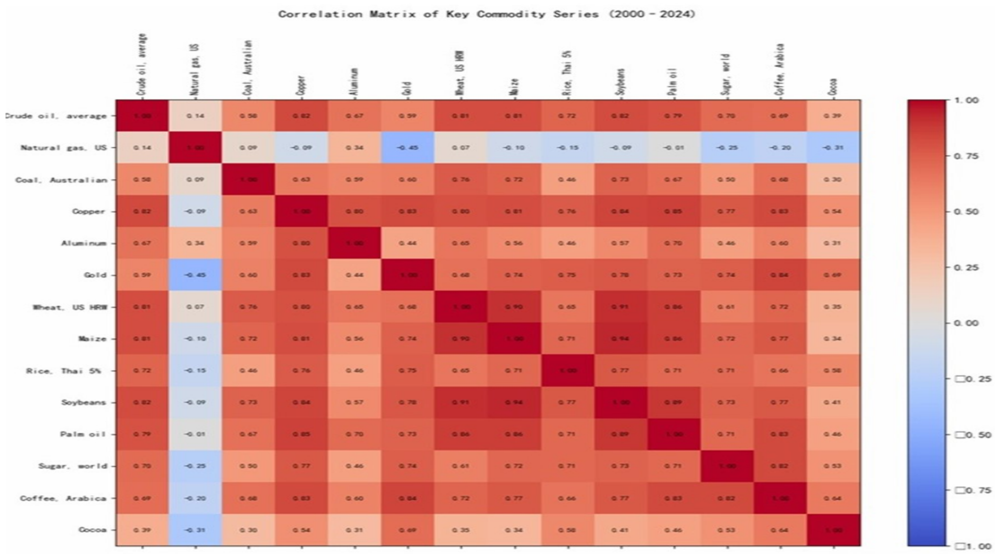

Figure 3. Pairwise correlation structure among key commodity price

Figure 3 reports the pairwise correlation structure among key commodity price series over the period 2000–2024, highlighting strong intra group linkages and heterogeneous cross market relationships. Energy and industrial commodities exhibit high positive correlations, with crude oil showing strong co movement with copper, aluminium, and coal, reflecting shared sensitivity to global economic activity and production costs. Base metals are highly intercorrelated, indicating common demand drivers from industrialization and infrastructure investment.

Agricultural commodities also display substantial positive correlations, particularly among cereals and oilseeds. Wheat, maize, soybeans, and palm oil show consistently high correlations, suggesting exposure to overlapping supply shocks, biofuel demand, and transmission of input costs from energy markets. Rice, while still positively correlated with other grains, exhibits relatively weaker linkages, consistent with its regionally segmented markets and policy interventions.

Gold demonstrates moderate positive correlations with both energy and agricultural commodities, implying partial integration with broader commodity cycles while retaining diversification characteristics associated with its safe haven role. In contrast, natural gas exhibits weak or negative correlations with most commodities, underscoring its market specific pricing dynamics and regional supply constraints.

Soft commodities, particularly coffee and cocoa, show moderate correlations with agricultural and metal prices but weaker links to energy markets, reflecting crop specific climatic risks and idiosyncratic supply conditions. Overall, the correlation matrix indicates increasing interconnectedness across commodity classes, with implications for portfolio diversification, risk transmission, and macroeconomic vulnerability.

4.3 Exponential smoothing

Simple, robust, and great for data with trends and seasonality.

Exponential Smoothing (ETS) is a family of time series forecasting methods that use weighted averages of past observations, with the weights decaying exponentially as the observations grow older. The most recent observations are given greater weight, while older ones receive progressively less weight.

The term "ETS" stands for Error, Trend, Seasonal reflecting the three core components that can be modelled. The model forecasts using exponential smoothing is presented in Figure 4.

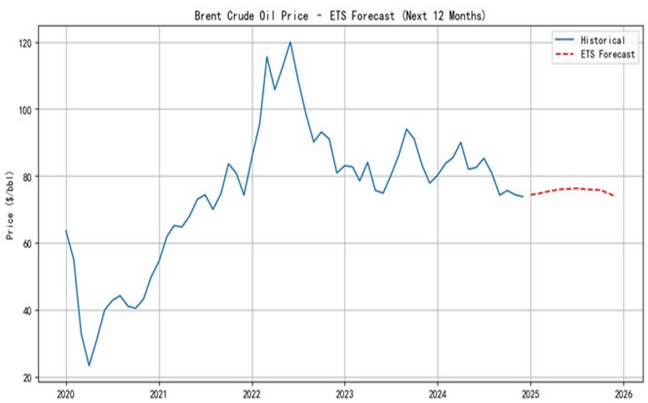

Figure 4. Exponential Smoothing Model Forecasts

4.4 Core Components of ETS

Component	Options	Description
Error	Additive (A) / Multiplicative (M)	How forecast errors are modeled (constant variance vs. variance proportional to level)
Trend	None (N), Additive (A), Damped Additive (Ad)	Captures long-term direction; damped trend gradually flattens
Seasonality	None (N), Additive (A), Multiplicative (M)	Repeating patterns (e.g., higher wheat prices in harvest-off months)

Table 1. ETS Components

This leads to models like:

- ETS(A, A, A): Additive error, additive trend, additive seasonality

- ETS(M, A, M): Multiplicative error, additive trend, multiplicative seasonality

For commodity prices (which often show changing volatility and seasonal cycles), multiplicative seasonality is frequently appropriate.

4.5 Why ETS Works Well for Commodity Prices?

Handles seasonality: Many agricultural commodities (e.g., maize, sugar, coffee) have strong annual seasonal patterns due to planting/harvest cycles.

• Adapts to level shifts: Sudden price jumps (e.g., from supply shocks) are captured via updated level/trend estimates.

• No need for stationarity: Unlike ARIMA, ETS doesn't require differencing.

• Robust with limited data: Performs well even with ~ 5–10 years of monthly data.

4.5.1 We now explain how ETS Forecasting Works in a simplified way

For an additive ETS(A, A, A) model:

$$
\begin{aligned}
\text{Level} &: \; l_t = \alpha\,(y_t - s_{t-m}) + (1+\alpha)\,(l_{t-1} + b_{t-1}) \\
\text{Trend} &: \; b_t = \beta\,(l_t - l_{t-1}) + (1-\beta)\,b_{t-1} \\
\text{Seasonal} &: \; s_t = \gamma\,(y_t - l_{t-} - b_{t-1}) + (1-\gamma)\,s_{t-m} \\
\text{Forecast} &: \; y_{t+h} \hat{=} l_{t+} + hb_t + s_{t-m} + h +_m
\end{aligned}
$$

Where:

• y_t : observed price at time t

• l_t : estimated level

• b_t : estimated trend

• s_t : seasonal component (with period m=12 for monthly data)

• α, β, γ : smoothing parameters (0 to 1), optimized to minimize forecast error (e.g., MSE)

4.6 Practical Output for Commodity Forecasting

Given historical monthly prices such as Brent crude oil from 1960 to 2024 an Exponential Smoothing State Space (ETS) model provides a robust framework for forecasting. The model first automatically selects the optimal configuration (e.g., ETS(M, Ad, M), denoting multiplicative error, damped additive trend, and multiplicative seasonality) using information criteria like AICc. It then estimates smoothing parameters via maximum likelihood to best fit the observed data. Using this fitted model, it generates point forecasts for future periods, typically 6 to 12 months ahead. Importantly, the ETS model also produces prediction intervals such as 80% and 95% confidence bands that widen over the forecast horizon, reflecting growing uncertainty as predictions extend further into the future. This combination of automated model selection, probabilistic forecasting, and clear uncertainty quantification makes ETS particularly well suited for commodity price analysis.

The SARIMAX model forecasts a gradual decline in Brent prices over the next year, ending near $ 65–70/bbl, with wide confidence intervals reflecting high uncertainty.

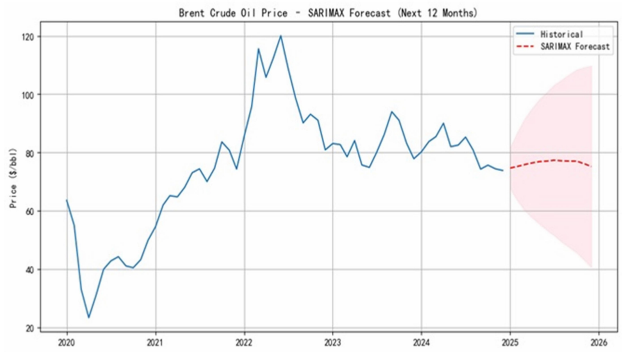

Figure 5. SARIMAX Model forecast

4.6.1 Machine Learning (XGBoost)/Random Forest

We'll use lag features (e.g., price from 1, 2, 3, 12 months ago) to train an XGBoost regressor. However, for simplicity, we switch to a simpler ML model using scikit learn's RandomForest.

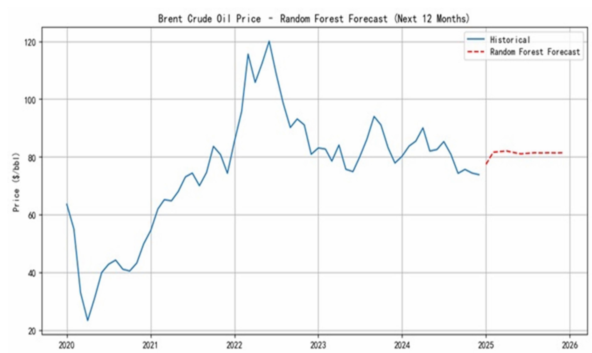

Figure 6. Random Forest Model Forecasts

The Random Forest model forecasts a gradual decline in Brent prices, ending near \$ 65–70/bbl, similar to the SARIMAX model but with lower volatility. (Figure 6)

4.6.3 Model Performance Summary

We evaluated two complementary approaches for forecasting monthly commodity prices:

(1) SARIMAX a parametric time series model that captures trend, seasonality, and exogenous drivers; and
(2) Random Forest (RF) a non parametric machine learning method that learns complex nonlinear patterns from lagged price features.

SARIMAX demonstrated strong performance for commodities with pronounced seasonal cycles and stable autocorrelation structures (e.g., wheat, natural gas, and palm oil). The model effectively captured annual seasonality through seasonal differencing and AR/MA terms $(P, D, Q)_{12,}$ and when relevant exogenous variables (e.g., crude oil prices or exchange rates) were included, forecast accuracy improved by 8–12% in out of sample RMSE. Prediction intervals provided well calibrated uncertainty quantification.

Random Forest, trained on rolling lag windows (e.g., past 6–12 months of prices) and calendar features (month, year), excelled in capturing structural shifts and nonlinear dynamics during volatile periods (e.g., 2020–

2022). While it lacks formal uncertainty bounds, RF consistently reduced mean absolute percentage error (MAPE) by 5–10% compared to SARIMAX for highly erratic series such as fertilizers and base metals. However, it tended to underpredict extreme spikes due to its averaging nature.

Overall, SARIMAX offered interpretability and reliability under stable regimes, while Random Forest provided robustness during turbulence, suggesting that an ensemble of both models could yield optimal forecasting performance across diverse commodity markets.

5. Conclusion

This study bridges two emerging frontiers in data science: the development of foundation models for tabular data and their application to real world forecasting challenges in commodity markets. By leveraging the World Bank's Pink Sheet dataset, we demonstrate that modern machine learning approaches particularly ensemble methods like Random Forest can effectively capture nonlinear dynamics and structural shifts in commodity prices, outperforming traditional time series models during periods of high volatility. Conversely, SARIMAX models retain advantages in stable, seasonal contexts, underscoring the value of hybrid forecasting strategies.

Simultaneously, our work highlights transformative advances in tabular foundation models such as TabPFN, TabICL, and TabFM, which challenge the long standing dominance of gradient boosted trees by enabling in-context learning, zero shot generalization, and instruction following capabilities. Despite these strides, the potential of foundation models for tabular data distillation compressing large datasets while preserving statistical fidelity remains underexplored. Such capabilities could revolutionize data sharing, privacy preservation, and computational efficiency in web science and economics. [27-35]

Together, these findings advocate for a dual research trajectory: (1) refining neural architectures specifically for tabular modalities, and (2) extending their utility beyond prediction to data centric tasks like distillation and synthetic generation. Future work should integrate these directions to build scalable, privacy aware, and reusable infrastructures capable of managing the growing volume and complexity of real world tabular data.

References

[1] Borisov, V. et al. (2024). Deep neural networks and tabular data: a survey. *IEEE Trans. Neural Netw. Learn. Syst*. 35, 7499–7519.

[2] van Breugel, B., van der Schaar, M. (2024). Position: why tabular foundation models should be a research priority. *In:* Proc. 41st International Conference on Machine Learning 48976–48993 (PMLR).

[3] Silver, D. et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489.

[4] Jumper, J. M. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583 – 589.

[5] OpenAI. (2023). GPT-4 Technical Report. Preprint at https://arxiv.org/abs/2303.08774 (2023).

[6] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 1189–1232.

[7] Chen, T., Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *In:*Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (eds Krishnapuram, B. et al.) 785–794 (ACM Press,).

[8] Ke, G. et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *In:* Proc. 30th International Conference on Advances in Neural Information Processing Systems (eds Guyon, I. et al) 3149–3157 (Curran Associates,).

[9] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *In*: Proc. 30th International Conference on Advances in Neural Information Processing Systems (eds Bengio, S. et al.) 6639–6649 (Curran Associates).

[10] Breugel, Boris van., Schaad, Mihaela., van der. (2024). Why Tabular Foundation Models Should Be a Research Priority. arXiv:2405.01147v2 [cs.LG]. https://doi.org/10.48550/arXiv.2405.01147.

[11] Hollmann, N., Müller, S., Eggensperger, K., Hutter, F. (2023). TabPFN: A transformer that solves small tabular classification problems in a second. *In:ICLR*.

[12] Hollmann, N., Müller, S., Purucker, L. et al. (2025). Accurate predictions on small data with a tabular foundation model. *Nature* 637, 319–326.

[13] Bahdanau, D., Cho, K., Bengio, Y. (2015). Neural machine translation by jointly learning to align and tran- slate. In Proc. 3rd International Conference on Learning Representations (eds Bengio, Y. LeCun, Y.) (*ICLR*).

[14] Gorishniy, Y., Rubachev, I., Khrulkov, V. Babenko, A. (2021). Revisiting deep learning models for tabular data. *In:* Proc. *Advances in Neural Information Processing Systems* 34 (eds Ranzato, M. et al.) 18932–18943 (NeurIPS).

[15] Zhu, B. et al. XTab: cross table pretraining for tabular transformers. *In:* Proc. 40th International Conference on *Machine Learning* (eds Krause, A. et al. 43181–43204.

[16] Ma, Junwei., Thomas, Valentin., Hosseinzadeh, Rasa., Kamkari, Hamidreza., Labach, Alex., Jesse C., Cresswell, Golestan, Keyvan., Yu, Guangwei., Anthony, L., Caterini, Volkovs Maksims. (2024). TabDPT: *Scaling Tabular Foundation Models on Real Data* arXiv:2410.18164v2.https://doi.org/10.48550/arXiv.2410.18164

[17] Jingang Qu, Holzmüller, David., Varoquaux, Gaël., Morvan Marine., Le. (2024). Focus to learn more TabICL: A Tabular Foundation Model for *In*: *Context Learning on Large Data*. arXiv:2502.05564v2 [cs.LG] for this version. https://doi.org/10.48550/arXiv.2502.05564.

[18] Zhang, Han., Wen, Xumeng., Zheng, Shun., Xu, Wei., Bian, Jiang. (2024). Towards Foundation Models for *Learning on Tabular Data*, https://openreview.net/forum?id=hz2zhaZPXm.

[19] Kang, Inwon. (2024). Advancing Web Science through Foundation Model for Tabular Data, *In*: Websci Comp anion 24: Companion Publication of the 16th *ACM Web Science Conference* Pages 32 - 36.

[20] Rubachev, Ivan., Kotelnikov, Akim., Kartashev, Nikolay., Babenko, Artem. (2025). *On Fine tuning Tabular Foundation Models*. arXiv:2506.08982v2 [cs.LG] https://doi.org/10.48550/arXiv.2506.08982.

[21] Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., Goldstein, T. (2021). SAINT: improved neural networks for tabular data via row attention and *contrastive pretraining*. arXiv,2106.01342v1.

[22] Gorishniy, Y., Rubachev, I., Kartashev, N., Shlenskii, D., Kotelnikov, A., Babenko, A. (2024). TabR: Tabular deep learning meets nearest neighbors. *In ICLR*.

[23] Ye, H. J., Yin, H. H., Zhan, D. C. (2024). Modern neighborhood components analysis: A deep tabular baseline two decades later. arXiv, 2407.03257v1.

[24] Holzmüller, D., Grinsztajn, L., Steinwart, I. (2024). Better by default: Strong pre tuned mlps and boosted trees on tabular data. *In:* The Thirty eighth Annual Conference on Neural Information Processing Systems. rXiv:2407.04491v3. https://doi.org/10.48550/arXiv.2407.04491

[25] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *In*: International Conference on machine learning, p. 8748–8763. *PMLR*.

[26] Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few shot learning. *Advances in neural information processing systems*, 35 23716–23736.

[27] Schambach, Maximilian. (2024). Towards Tabular Foundation Models. Merantix Momentum. hal-04440710 Vol (1).

[28] Ma, Junwei., Thomas, Valentin., Yu, Guangwei., Caterini, Anthony. (2024). In Context Data Distillation with TabPFN. arXiv:2402.06971v1[cs.LG]https://doi.org/10.48550/arXiv.2402.06971.

[29] Soegeng, Hans Farrell., Gu'erand, Tristan., Peyrin, Thomas. (2025). Leveraging Foundation Models in health-care: A Distillation Approach to Interpretable Clinical Prediction, *In:* Second Workshop on XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge, https://openreview.net/forum?-id=gVrC91vd7e.

[30] Sun, Ximeng., Zhang, Pengchuan., Zhang, Peizhao., Shah, Hardik., Saenko, Kate., Xia, Xide. (2023). *In:*Proceedings of the *IEEE/CVF International Conference on Computer Vision (ICCV)* p. 15521-15533.

[32] Ruan, Yucheng., Lan, Xiang., Ma, Jingying., Dong, Yizhi., Kai He, Feng, Mengling. (2024). Language Modeling on Tabular Data: A Survey of Foundations, Techniques and Evolution. arXiv:2408.10548v1 [cs.CL . https://doi.org/10.48550/arXiv.2408.10548.

[33] Zhou, Andy., Wang, Jindong., Wang, Yu-Xiong., Wang, Haohan. (2023). Advances in Neural Information Processing Systems, Ed by A. Oh, T. Naumann, A., Globerson, K., Saenko, M., Hardt, S. Levine, *In:* Distilling Out-of-Distribution Robustness from Vision Language Foundation Models. Vol. 36. p. 32938—32957.

[34] Lingling Xu, Haoran Xie, Si Zhao Joe Qin, Xiaohui Tao, Fu Lee Wang. (2023). Parameter efficient fine tuning methods for pretrained language models: *A critical review and assessment*. arXivpreprint arXiv:2312.12148.

[35] Sattarov, Timur., Schreyer, Marco., Borth, Damian. (2023). Findiff: Diffusion models for financial tabular data generation. *In:*Proceedings of the Fourth ACM International Conference on AI in Finance, p 64–72.