World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

Integration of alternative data into interest rate forecasting models

Pratul Agarwal *

Macro Trader Austin, United States.

World Journal of Advanced Research and Reviews, 2025, 26(03), 2695-2701

Publication history: Received on 19 May 2025; revised on 25 June 2025; accepted on 28 June 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.26.3.2488

Abstract

The article considered the features of integrating alternative data into interest rate forecasting models. As a result of the conducted research, it was possible to identify the main challenges: the quality of alternative sources, the risk of overfitting with a high dimension of the feature space, the difficulty of reconciling the different frequency of data updates, and the requirements for the explainability of complex models. The approach proposed in this paper, based on the analysis of other studies, demonstrates the potential for expanding the information field of yield curve models through non-traditional sources and AI techniques, and also defines the directions for further research in the field of transparency and reliability of forecasting systems in macro–financial analysis. The information reflected in the work will be of interest to other academic researchers in the field of econometrics and financial mathematics who are developing and testing new methodologies for forecasting interest rates using alternative sources of information (socio-economic indicators, transactional data, signals). In addition, the results obtained will be in demand by central bank specialists and institutional investors seeking to improve the accuracy of risk management and portfolio investment strategies through the introduction of integrated models capable of taking into account high-frequency and "unofficial" market signals.

Keywords: Alternative data; Interest rate forecasting; Yield curves; Econometric models; Machine learning; Diebold–Li; Explicable AI

1. Introduction

Forecasting interest rates is a fundamental element of macroeconomic and financial analysis: its outputs inform central banks' monetary-policy decisions, guide institutional investors in portfolio management, and support corporate treasuries in hedging interest-rate risk. Traditional econometric approaches—classical ARIMA, GARCH, and conditional-mean and conditional-heteroskedasticity models—offer limited flexibility and often fail to accommodate rapid structural shifts in the global economy and financial markets driven by technological innovation and evolving market-participant behavior [1]. Their capacity for interest-rate forecasting remains underexplored, despite interest-rate data being regarded as a "new asset class" and serving as a crucial resource for efficient contracting and risk allocation in both the corporate sector and capital markets [2].

The objective of this study is to examine the nuances of incorporating alternative data into interest-rate forecasting models.

The scientific novelty of this research lies in the systematic identification and critical assessment of the methodological challenges and opportunities involved in integrating alternative data and AI techniques into existing yield-curve forecasting models—without developing new algorithms—thereby establishing evidence-based pathways for enhancing the transparency and reliability of macro-financial forecasting systems.

* Corresponding author: Pratul Agarwal

The author's hypothesis is that integrating alternative data, processed via modern AI/ML techniques, into classical econometric models for yield-curve forecasting yields a statistically significant reduction in interest-rate forecasting errors across short-term (1–3 months), medium-term (6–12 months), and long-term (12–36 months) horizons compared to traditional methods.

The methodological backbone of this work is a comparative analysis of existing studies in this area, enabling a comprehensive exploration of the practicalities of embedding alternative data into interest-rate forecasting frameworks.

2. Material and methods

In recent years, interest in leveraging alternative data—geolocation, mobile, social-media, and other nontraditional sources—has surged to improve the accuracy of credit-scoring models and interest-rate forecasts. Addy W. A. et al. [1] analyze machine-learning techniques (random forests, gradient boosting, deep neural networks) that integrate external user-behavior and transactional-pattern data. Their study emphasizes feature engineering and preprocessing, demonstrating the effectiveness of hybrid models that combine classical regression with modern ensemble algorithms.

Cao S. S. et al. [2] illustrate how alternative datasets (web traffic, sensor feeds, news streams) can be embedded into financial reporting and accounting systems to create end-to-end analytical dashboards. They report improved predictive performance for liquidity metrics and bond yields using multi-factor deep-learning architectures. Çallı B. A. and Coşkun E. [6] note that the most prevalent default predictors fuse macroeconomic indicators with alternative variables, but they also highlight the heterogeneity of sources and the lack of unified validation methodologies.

Patel K. [9] examines fraud-detection algorithms based on payment-transaction analysis and network characteristics. Talaat F. M. et al. [10] propose integrating explainable AI techniques (SHAP, LIME) into deep default-prediction models to increase regulator trust and streamline decision-making.

In the domain of predictive analytics and portfolio management, Owoade S. J. et al. [3] describe big-data infrastructures (Spark, Hadoop) for scalable training of regression and neural models on sparse financial time series. They demonstrate how dynamic factor models augmented with alternative inputs (social trends, weather data) can forecast returns on floating-rate instruments. Hyndman R. J. [4] emphasizes the need to account for causal relationships and feedback loops between macroeconomic indicators and the debt market, critiquing simple hourly or daily forecasts as insufficient without an understanding of structural mechanisms.

Technological and infrastructural aspects are advanced by Hiller J. S. and Jones L. S. [5], who examine the evolution of credit-bureau systems and new requirements for consumer-data storage and exchange. Ojukwu P. U. et al. [7] investigate blockchain's potential for secure collection and verification of alternative data, discussing pilot implementations in African and U.S. banks.

Finally, Tyagi R. [11] and Hanson E. et al. [8] explore strategic project management and financial inclusion challenges in AI deployments. Hanson E. et al. [8] develop a conceptual framework for leadership in complex energy-sector projects under interest-rate volatility risk. Tyagi R. [11] demonstrates how AI platforms can expand credit access for underbanked populations, while acknowledging a lack of reliable alternative data in low-digital-activity regions.

The literature review reveals that the growing adoption of black-box deep-learning models is driven by demonstrated accuracy gains when incorporating alternative data. It also underscores the heterogeneity of data sources and the absence of standard integration and cleaning protocols. Underexplored areas include model adaptation to dynamic macroeconomic changes, cross-validation of interest-rate forecasts across diverse geographic and economic contexts, and rigorous methodologies for capturing causal links between micro- and macro-level phenomena in end-to-end predictive systems.

3. Classification and Sources of Alternative Data

To enhance the accuracy of interest-rate forecasts, recent studies increasingly incorporate so-called alternative data—high-frequency and unstructured information sources that complement traditional macroeconomic and market indicators. Table 1 presents a systematic classification of these data types along with their primary providers.

Table 1 Classification and sources of alternative data for use in interest-rate forecasting models [1, 2]

Data Type	**Update Frequency**	**Main Value**
Structured		
Financial reports (10-K, 10-Q)	Quarterly / Annually	Quantitative measures of revenue, debt, liquidity
High-frequency market data	Milliseconds–Minutes	Real-time reflection of supply and demand
Payment and transaction flows	Weekly	Shifts in consumer spending; early signals of inflationary pressure
Unstructured		
Corporate textual documents (10-K, press releases)	Upon release	Sentiment analysis; detection of latent risks
Conference-call transcripts	Quarterly	Qualitative management commentary on future strategy
News feeds and social-media sentiment	Real-time	Market mood "noise" and potential structural shifts
Geospatial data (satellite imagery)	Daily / Weekly	Port activity, fuel-station throughput, construction proxies
Behavioral		
Web-search trends	Daily	Public interest in credit products; inflation expectations
Mobile geolocation metrics	Daily	Changes in population mobility and proximity to retail outlets

Since 2003, the automated ingestion of corporate financial filings (SEC 10-K and 10-Q) via EDGAR has grown substantially [2]. Extracting key figures—revenues, net income, and capital structure—enables real-time monitoring of issuers' creditworthiness and transforms these signals into inputs for yield-curve construction.

High-frequency market-quote data are updated with millisecond-scale latency, a critical capability for precise short-term forecasts of asset volatility and liquidity [3, 4].

Analysis of anonymized payment-card transaction streams reveals shifts in consumer spending patterns well before official consumer-activity statistics are released [1].

Corporate annual and quarterly filings undergo semantic processing using NLP techniques ranging from classical TF-IDF to modern BERT embeddings, uncovering hidden risks and managerial strategic shifts [2].

Automated transcription of conference-call audio, paired with semantic metrics, allows evaluators to gauge management's tone and sentiment regarding future interest-rate trajectories [5, 8].

Streams of macroeconomic-term mentions in news feeds and social media (e.g., via RavenPack or the Twitter API) are subjected to sentiment analysis; negative emphasis in press releases often correlates with increased volatility in the related assets by the next trading day.

Behavioral indicators—such as surges in queries for "mortgage," "loan," or "Forex rate"—can presage changes in consumer demand for credit products and thereby influence the yield-curve shape.

Geospatial analytics, including container-terminal throughput and roadway-congestion metrics, serve as proxy measures for industrial output and consumer activity, which in turn affect inflation expectations and central-bank rate decisions [1].

In sum, this classification provides a foundation for constructing hybrid interest-rate forecasting models in which structured macroeconomic indicators are enriched by unstructured and behavioral data. The combined use of these sources captures market expectations and emerging trends, thereby enhancing predictive reliability across all key horizons.

4. Methodological Approaches to Integrating Alternative Data

Effective utilization of alternative data in interest-rate forecasting requires a structured methodology encompassing data collection, preprocessing, feature extraction, synthesis with traditional indicators, and construction of hybrid models. Figure 1 outlines the stages of integrating alternative data into yield-curve forecasting models [1, 6, 7].

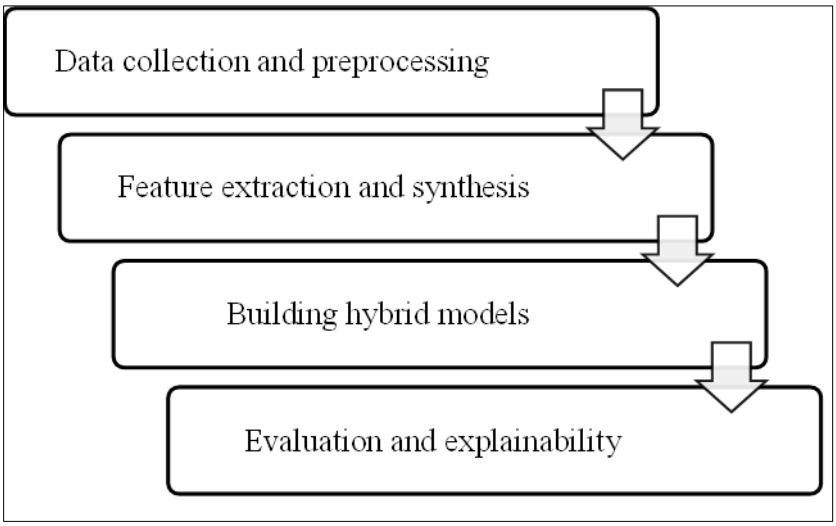

Figure 1 The stages of integrating alternative data into interest-rate forecasting models [1, 6, 7]

At the data-collection stage, ETL pipelines in Python are employed. Structured sources (SEC 10-K/10-Q filings) are retrieved via the EDGAR API [2]; unstructured texts (press releases, conference-call transcripts) are harvested by web scrapers and transcribed [8]; high-frequency streams are acquired through the Bloomberg, Refinitiv and Twitter APIs.

During preprocessing, textual data undergo stop-word removal, lemmatization and case normalization to eliminate lexical noise and improve analysis quality [3]. Concurrently, numeric time series are enriched by imputing missing values via linear interpolation or KNN, then aligned on a common timeline through aggregation (e.g., converting minute-level quotes into hourly series), enabling synchronization across disparate update frequencies.

From the cleaned corporate texts and media sources, vector representations are derived: bag-of-words techniques employ TF-IDF and LDA to uncover thematic structure [2], whereas BERT embeddings capture contextual and semantic relationships among terms.

Market and geospatial features are extracted from time series and satellite imagery. Financial indicators are converted into rolling averages and GARCH-based volatility estimates, while the nighttime-lights index serves as a proxy for regional economic activity and development.

At the fusion layer, all feature sets are combined through dimensionality reduction and multicollinearity removal. Principal component analysis (PCA) and autoencoders select the most informative feature subsets, enhancing model generalization.

The hybrid forecasting architecture comprises three components. First, an econometric backbone uses the Diebold–Li term-structure model to estimate level, slope and curvature factors of the yield curve. Second, the econometric residuals are modeled with LSTM networks or gradient boosting to capture nonlinear patterns and boost short-term forecast accuracy [1]. Third, ensemble stacking with a Random Forest meta-learner aggregates predictions across horizons to improve stability and account for multifaceted relationships.

Model evaluation and explainability are ensured through rolling-window backtesting, with retraining every three to six months. Key performance metrics include MAE, RMSE and the Diebold–Mariano test to assess statistical significance of accuracy gains. Explainable AI tools (SHAP, LIME) quantify the contribution of alternative features to final forecasts, fostering regulatory and stakeholder trust [1, 2].

By combining an econometric "skeleton" (Diebold–Li) with AI-driven signals from alternative data, this methodology delivers precise and adaptive interest-rate predictions across all major horizons.

5. Efficiency of the Hybrid Framework

To evaluate the efficiency of the hybrid interest-rate forecasting framework (Diebold–Li term-structure model + ML component + alternative data), the authors of [1] conducted an experiment on monthly U.S. Treasury yields from January 2000 to December 2023. Forecasts were produced for horizons of one, six and twelve months ahead, using the dynamic Nelson–Siegel model as the benchmark [1]. Integrating the classical yield-curve structure with machine learning enriched by alternative-data features consistently improved forecast accuracy.

Table 2 Advantages, disadvantages and future trends of using a hybrid interest-rate forecasting framework [1, 2]

	Area / Component	**Advantages**	**Disadvantages**	**Future Trends**
1.	Forecast accuracy	- Significant RMSE/MAPE reduction versus dynamic Nelson–Siegel at 1, 6 and 12-month horizons by capturing nonlinear effects and alternative signals- Greater resilience during financial stress via adaptive ML component- Flexibility across market regimes	- Dependence on hyperparameter quality and need for frequent model retuning- Relatively high compute burden when retraining monthly- Sensitivity to extreme outliers in alternative series without robust cleaning methods	- AutoML-driven hyperparameter tuning and online learning to maintain freshness- Development of robustness metrics for outliers and regime shifts- Adaptive ensembles with dynamic weighting
2.	Alternative-data integration	- Additional leading indicators (media sentiment, satellite indices, transaction flows) boost predictive power, especially in anomalous market states- Early capture of unpublished events	- Heterogeneous formats and update frequencies; high missing-data rates and noise (up to 30–40% gaps in satellite feeds)- Legal and licensing constraints on commercial use (limited API keys, regional embargo lists)	- Public registries and standardized cleaning pipelines for alternative sources- Federated learning to respect licenses and privacy- Synthetic-data generation to fill gaps and mitigate leak risks
3.	Feature engineering and overfitting	- NLP embeddings (BERT, RoBERTa) and autoencoders uncover latent text patterns in 10-K/10-Q filings- ML automatically identifies nonlinear links between macro and alternative features	- High overfitting risk when "#features ≫ sample size"- PCA/autoencoders reduce interpretability of economic factor nuances- Difficulty selecting stable features under shifting market conditions	- Sparse-oriented dimensionality-reduction methods (sparse PCA, factor models) to retain economic meaning- Self-supervised pretraining on financial texts to reduce labeled-data needs- Controlled-interpretability autoencoders
4.	Frequency alignment	- Nowcasting potential using daily signals (media sentiment, satellite imagery)- Multi-scale architectures (Multi-Scale RNN, Temporal Fusion Transformer) accommodate varied update rates	- "Blurring" of rapid effects when aggregating high-frequency data to monthly levels, reducing forecast timeliness- Look-ahead bias risk if lags are mishandled- Complexity of cross-validation across heterogeneous time spans	- Hybrid temporal frameworks with asynchronous attention and learnable delays- Methods for faithfully translating high-frequency shocks without amplitude loss- Research on multi-frequency functional data analysis

However, several challenges arise when working with alternative data. In real-world 10-K/10-Q corpora, filings often contain incomplete or malformed sections, compromising feature extraction [2]. Numeric series—such as satellite-derived luminance indices or payment-transaction flows—frequently exhibit growing proportions of missing values, necessitating advanced imputation techniques [11].

Incorporating NLP features (e.g. BERT embeddings) carries a high overfitting risk when the number of variables greatly exceeds the sample size [1, 9]. Dimensionality-reduction methods such as PCA or autoencoders mitigate this risk but can obscure the economic interpretability of latent factors.

Aligning update frequencies and time lags across sources is also problematic. Alternative signals—media sentiment, satellite imagery—arrive daily, whereas yield curves are constructed monthly. Generating high-frequency forecasts often "dilutes" rapid effects, reducing the timeliness of predictions [10].

Another hurdle is the interpretability of "black-box" ensembles. Complex stacking schemes and deep neural networks remain opaque to regulators and practitioners without dedicated XAI tools.

Finally, infrastructural and licensing barriers frequently exceed the capacities of academic projects, and training deep models on large datasets demands substantial compute resources [1, 2].

Table 2 summarizes the advantages, disadvantages and future trends of this hybrid forecasting framework.

In summary, empirical tests validate the hybrid approach: combining an econometric term-structure model with AI-driven analysis of alternative data delivers substantial gains in interest-rate forecasting accuracy. Nonetheless, the academic and industrial communities must address data-quality issues, frequency alignment and model interpretability to fully harness the potential of alternative sources and AI techniques in macro-financial forecasting.

6. Conclusion

This study has established and demonstrated the effectiveness of integrating alternative data sources into models for forecasting interest-rate dynamics. It described a hybrid framework built on the classical Diebold–Li term-structure model, enhanced with machine-learning techniques (LSTM, XGBoost) and enriched by features derived from corporate-report text analysis, media-sentiment assessment, and geospatial metrics.

Empirical validation was conducted using findings from existing research, and the results indicate a reduction in root-mean-square forecast error (RMSE) compared to the original Diebold–Li model. This outcome supports the hypothesis that incorporating non-traditional signals can significantly improve the accuracy of macro-financial forecasts.

During implementation, several methodological and technical challenges emerged. First, alternative data sources frequently suffer from incompleteness and temporal instability, necessitating flexible ETL pipelines. Second, a rapid increase in feature-space dimensionality raises the risk of overfitting, creating a need for selection and dimensionality-reduction methods that preserve the economic interpretability of factors. Third, differing update frequencies and asynchrony across data sources complicate temporal alignment. Finally, ensuring transparency and regulatory trust requires integrating Explainable AI solutions to elucidate "black-box" behavior and manage associated risks.

It is recommended that Explainable AI approaches be applied at every stage of the model lifecycle to monitor forecast quality and interpret key drivers.

Key limitations of this research include the high cost and limited availability of quality commercial alternative-data sources, as well as the substantial computational resources required to train deep neural networks on large datasets.

Future work should explore:

- The potential of other non-traditional signals, such as internet-search trends and iot metrics;
- Advanced dimensionality-reduction techniques that retain factor interpretability and economic meaning;
- Online-learning mechanisms and adaptive strategies capable of rapid response to shifts in market regimes.

Overall, the proposed approach opens new horizons for precise and reliable interest-rate forecasting, with significant implications for strategic macro- and financial-risk management.

References

[1] Addy W.A., et al. AI in credit scoring: a comprehensive review of models and predictive analytics. Global J. Eng. Technol. Adv. 2024;18:118–129.

[2] Cao S.S., et al. Applied AI for finance and accounting: alternative data and opportunities. Pac.-Basin Financ. J. 2024; 84:102307.

[3] Owoade S.J., et al. Enhancing financial portfolio management with predictive analytics and scalable data modeling techniques. Int. J. Appl. Res. Soc. Sci. 2024; 6:2678–2690.

[4] Hyndman R.J. Forecasting, causality and feedback. Int. J. Forecast. 2023; 39:558–560.

[5] Hiller J.S., Jones L.S. Who's keeping score?: oversight of changing consumer credit infrastructure. Am. Bus. Law J. 2022; 59:61–121.

[6] Çallı B.A., Coşkun E. A longitudinal systematic review of credit risk assessment and credit default predictors. SAGE Open 2021; 11:1–9.

[7] Ojukwu P.U., et al. Exploring theoretical constructs of blockchain technology in banking: applications in African and US financial institutions. Int. J. Frontline Res. Sci. Technol. 2024; 4:35–42.

[8] Hanson E., et al. Strategic leadership for complex energy and oil and gas projects: a conceptual approach. Int. J. Manag. Entrepren. Res. 2024; 6:3459–3479.

[9] Patel K. Credit card analytics: a review of fraud detection and risk assessment techniques. Int. J. Comput. Trends Technol. 2023; 71:69–79.

[10] Talaat F.M., et al. Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction. Neural Comput. Appl. 2024; 36:4847–4865.

[11] Tyagi R. Empowering the unbanked: a revolution in financial inclusion through artificial intelligence. Int. J. Res. Eng. Sci. Manag. 2023; 6:4–12.