Faithful and Interpretable Explanations for Complex Ensemble Time Series Forecasts using Surrogate Models and Forecastability Analysis

Yikai Zhao
Amazon Web Services
Austin, Texas, USA
yikai@amazon.com

Jiekai Ma
Amazon Web Services
Seattle, Washington, USA
jiekal@amazon.com

ABSTRACT

Modern time series forecasting increasingly relies on complex ensemble models generated by AutoML systems like AutoGluon, delivering superior accuracy but with significant costs to transparency and interpretability. This paper introduces a comprehensive, dual-approach framework that addresses both the explainability and forecastability challenges in complex time series ensembles. First, we develop a surrogate-based explanation methodology that bridges the accuracy-interpretability gap by training a LightGBM model to faithfully mimic AutoGluon's time series forecasts, enabling stable SHAP-based feature attributions. We rigorously validated this approach through feature injection experiments, demonstrating remarkably high faithfulness between extracted SHAP values and known ground truth effects. Second, we integrated spectral predictability analysis to quantify each series' inherent forecastability. By comparing each time series' spectral predictability to its pure noise benchmarks, we established an objective mechanism to gauge confidence in forecasts and their explanations. Our empirical evaluation on the M5 dataset found that higher spectral predictability strongly correlates not only with improved forecast accuracy but also with higher fidelity between the surrogate and the original forecasting model. These forecastability metrics serve as effective filtering mechanisms and confidence scores, enabling users to calibrate their trust in both the forecasts and their explanations. We further demonstrated that per-item normalization is essential for generating meaningful SHAP explanations across heterogeneous time series with varying scales. The resulting framework delivers interpretable, instance-level explanations for state-of-the-art ensemble forecasts, while equipping users with forecastability metrics that serve as reliability indicators for both predictions and their explanations.

ACM Reference Format:
Yikai Zhao and Jiekai Ma. 2025. Faithful and Interpretable Explanations for Complex Ensemble Time Series Forecasts using Surrogate Models and Forecastability Analysis. In *KDD '25: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 03–04, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

KDD '25, August 03–04, 2025, Toronto, ON, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Time series forecasting is a cornerstone of decision-making across numerous sectors, including demand planning, financial prediction, resource management, and climate modeling. The pursuit of higher predictive accuracy has driven the development of sophisticated machine learning (ML) models, particularly complex ensemble models often generated by Automated Machine Learning (AutoML) frameworks. These ensembles, which combine the strengths of diverse forecasting techniques, have demonstrated significant improvements in predictive performance [6].

However, the very complexity that fuels this enhanced performance often renders these powerful time series ensemble models opaque "black boxes" [8]. In applications where forecasts inform critical decisions, this lack of transparency can erode trust and hinder the adoption of these advanced models [8]. This paper addresses the specific challenge of explaining forecasts generated by such state-of-the-art time series ensemble models.

As a prominent example of such systems, we consider AutoGluon-TimeSeries (AutoGluon-TS). AutoGluon-TS has demonstrated strong empirical performance across 29 benchmark datasets, outperforming a variety of traditional forecasting models in both point and quantile prediction. In many cases, it even exceeds the performance of the best-in-hindsight ensemble of previous methods [6]. The high accuracy of AutoGluon-TS, and indeed many cutting-edge time series ensemble systems, is achieved by creating complex combinations of diverse models. These can include statistical methods (like ARIMA, ETS), deep learning models (like DeepAR), and foundational models (like Chronos). This inherent heterogeneity and multi-layered structure make it exceptionally difficult to apply standard Explainable Artificial Intelligence (XAI) techniques directly and reliably across the entire ensemble [1].

A particular challenge with these complex time series ensembles, as exemplified by AutoGluon-TS, is that their constituent models often require fundamentally different approaches to explanation. This makes it impossible to apply a single explanation method, such as Shapley Additive exPlanations (SHAP), uniformly and meaningfully across all member models. For instance, our preliminary investigations using permutation feature importance directly on an AutoGluon ensemble yielded unstable and inconsistent results in faithfulness tests. This observation aligns with documented limitations of permutation-based methods, which can disrupt the learned relationships between features and the target variable [10].

To overcome these challenges, we propose a surrogate-based explanation methodology. The core idea is to decouple the complex prediction task from the explanation task. We first leverage the power of AutoGluon-TS to train a high-performance forecasting ensemble, treating it as a black box f_{AG}. We then train a simpler, inherently more interpretable model – specifically, LightGBM – to act as a surrogate, aiming to accurately mimic the point forecasts generated by the AutoGluon-TS. Once this surrogate model, f_{LGBM}, achieves high fidelity in replicating f_{AG}'s predictions, we can apply well-established and efficient explanation techniques to it. Additionally, our approach incorporates complementary techniques, such as forecastability analysis, to provide a more robust framework for interpreting complex forecasts and assessing their reliability.

A critical consideration often overlooked in forecast explainability is the inherent forecastability of the time series being analyzed. Explanations derived from models attempting to predict inherently unpredictable or chaotic data may themselves be misleading or unreliable, regardless of the sophistication of the explanation technique. This creates a fundamental challenge: without assessing the intrinsic predictability of the underlying data, users may place unwarranted confidence in explanations of forecasts that are essentially unpredictable. Our framework addresses this challenge by integrating forecastability analysis with model explanations, providing users with critical context about when to trust both forecasts and their explanations. This integration is particularly valuable in business environments where heterogeneous time series with varying levels of predictability must be processed and interpreted efficiently.

The main contributions of this paper are:

- Validation of a surrogate model approach (LightGBM+TreeSHAP) for explaining complex AutoML time series forecasts (e.g., AutoGluon-TS).
- Quantitative evaluation of surrogate explanation faithfulness using feature injection, showing high correlation with known ground truth effects.
- Integration of spectral predictability and a filter mechanism by comparing it to its white noise benchmarks
- Highlighting the necessity and providing a method for per-item normalization enabling comparable SHAP explanations on heterogeneous series.

This work aims to offer a practical, validated approach for presenting trustworthy insights into the forecasts of sophisticated ensemble time series models.

2 RELATED WORK

This section reviews existing work relevant to explaining time series forecasts, focusing on techniques applicable to surrogate modeling, per-item normalization, and forecastability analysis.

2.1 Surrogate Models for XAI

Surrogate modeling is a practical technique for model-agnostic explainability where an interpretable model (e.g., LightGBM) is trained to mimic a complex black-box system's input-output behavior [4, 8]. This widely applicable technique continues to be adapted for specific domains, including time series forecasting [20], offering valuable simplified global insights into the black box and enhancing comprehension [4]. Harnessing these benefits critically hinges on ensuring high *fidelity*: the surrogate must accurately replicate the original model's predictions, as explanations lack meaning otherwise [21]. Fortunately, fidelity is a quantifiable and verifiable metric, making trustworthy explanation via surrogates a manageable goal rather than an insurmountable barrier. While inherently interpretable models remain ideal [21], surrogates provide an essential bridge when state-of-the-art accuracy necessitates a black-box approach but practical deployment requires understanding its behavior. Consequently, when implemented with rigorous fidelity validation, surrogate modeling offers a robust and highly valuable methodology for practical XAI.

This study adopted LightGBM as the surrogate model, a decision underpinned by two key factors: its consistently demonstrated high performance in forecasting accuracy (e.g., [13]) and its inherent tree-based architecture. This structural characteristic is particularly advantageous as it integrates seamlessly with the efficient TreeSHAP algorithm for the calculation of SHAP values [16].

2.2 Per-Item Normalization for Time Series Explainability

Per-item normalization is crucial for meaningful time series explainability, particularly with heterogeneous scales common in business forecasting [18]. Real-world applications often involve series with vastly different magnitudes, and without normalization, methods like SHAP generate feature attributions biased by absolute scale [17].

This scale-dependency fundamentally distorts local explanations across items when applied to unnormalized heterogeneous series. The core issue stems from the SHAP base value (ϕ_0), which typically represents the average model prediction across the mixed-scale background data. This global ϕ_0 thus becomes unrepresentative for any specific item whose scale deviates significantly from this overall average. Consequently, the SHAP additivity property ($\hat{y} \approx \phi_0 + \sum \phi_j$) forces the feature contributions ($\sum \phi_j$) to absorb this baseline mismatch. In both low- and high-volume cases, individual SHAP values (ϕ_j) are disproportionately scaled primarily to compensate for the inappropriate global baseline, rather than accurately reflecting the true, context-specific marginal impact of features relative to that item's own scale. This significantly obscures genuine local insights.

Applying per-item normalization (e.g., Z-scoring) addresses this by reframing the task to explain relative deviations from an item-specific baseline. This yields equitable, comparable explanations aligned with business focus on relative impacts [9], aiding workflows like exception handling. Conversely, ignoring normalization risks misleading interpretations and suboptimal decisions, undermining trust [18], making per-item normalization a fundamental step for reliable explanations supporting business processes.

2.3 Time Series Forecastability Analysis

Real-world business datasets contain heterogeneous time series; some exhibit clear patterns, while others are erratic and resist reliable prediction regardless of model sophistication [19]. Attempting advanced modeling on inherently unpredictable series sets unrealistic expectations for stakeholders and risks misguided decisions

based on a false sense of precision. Identifying those time series with a low forecastability score *before* modeling is crucial in large-scale environments like supply chains with numerous item-location combinations, including sporadic "long-tail" items, enabling tiered forecasting strategies [11].

Explanations derived from models attempting to predict inherently chaotic or noisy data may themselves be unreliable [18]. Combining forecastability analysis with model explanations yields a more complete picture: understanding not only feature contributions but also whether the underlying data supports reliable prediction. Recent work has focused on quantitative measures to assess time series forecastability prior to modeling:

Spectral Predictability: The spectral predictability score [7] is computed as the entropy of the power spectral density after trend removal. This metric provides a model-agnostic assessment of a time series' intrinsic forecastability.

Benchmarking approaches: Comparing a time series' forecastability metrics against benchmarks derived from noise patterns with equivalent characteristics (length, sparsity) provides a reference point for assessing whether a series contains meaningful signal or is predominantly noise [12]. Time series that score below these benchmarks are likely to be inherently unpredictable.

Our work builds on these approaches by integrating spectral predictability analysis with surrogate-based model explanations, creating a comprehensive framework that considers both model behavior and data characteristics when assessing forecast explainability.

3 METHODOLOGY

The end-to-end process involves training the complex AutoGluon TS model, engineering relevant and explainable features, training a LightGBM surrogate to mimic the normalized AutoGluon predictions, applying TreeSHAP to the surrogate, evaluate faithfulness of the SHAP values derived from the surrogate model, optionally caliberate the SHAP values to be additive to the AutoGluon's raw forecasts[1] and finally provide explanation based on SHAP values. This process could be augmented by forecastability as a filter or confidence score mechanism for the forecasts and explanations.

3.1 Feature Engineering for Explainability

We engineered a distinct feature set (X_{eng}) as interpretable input for the LightGBM surrogate model. This set differs from features typically used by complex forecasters like AutoGluon by deliberately limiting extensive lag features (to avoid diluting SHAP explanations) while augmenting with other signals. This comprehensive set comprised: time-based features (week, month, day of week, etc.), key target lags, rolling/expanding window statistics (mean, std dev, skewness, etc.), percentage changes, signal decomposition features (trend, seasonality), and history/age features. These features were chosen to provide interpretable drivers for the explanation model.

3.2 Per-Item Normalization Strategy

Given the wide variation in demand scales in the M5 dataset, applying a consistent normalization scheme per-item is crucial before training the surrogate and calculating SHAP values.

[1]Details on the calibration method are provided in Appendix.

We employ per-item standardization (z-score normalization). For each item i, we calculate the mean μ_i and standard deviation σ_i of its historical demand y_i. The target variable for the surrogate model becomes the normalized AutoGluon forecast: $z_{AG,k} = (\hat{y}_{AG,k} - \mu_i)/\sigma_i$. The LightGBM model f_{LGBM} is trained to predict $z_{AG,k}$. SHAP values ϕ_k are then computed to explain $f_{LGBM}(x_{eng,k}) \approx z_{AG,k}$. For final interpretation in the original demand units, the SHAP values can be denormalized by multiplying by σ_i.

3.3 Faithfulness Evaluation

Faithfulness ensures explanations accurately reflect a model's reasoning or true feature influence, a critical check beyond predictive fidelity [18, 21, 22]. Evaluating faithfulness rigorously is challenging due to the typical absence of ground truth explanations for real-world models [5]. A key strategy to overcome this, recognized in recent evaluation benchmarks [15], is the use of synthetic or known ground truths; this principle has been practically applied, for instance, to evaluate anomaly explanations against generated ground truths [3]. Our work employs a specific form of this strategy: a feature injection experiment, conceptually illustrated in the Appendix (Table 3). This approach also serves as a practical debugging test and sanity check for explanation reliability [2, 14]. The method involves introducing a synthetic feature with a known, predefined impact on the target variable, retraining the surrogate model on this modified data, and extracting the explanation (e.g., SHAP values, $\phi_{injected}$) for the synthetic feature. Faithfulness is then quantified by comparing the extracted explanation $\phi_{injected}$ against the known ground truth effect; high correlation indicates the method accurately captures the injected influence [18].

This technique provides a valuable quantitative faithfulness measure for at least one feature's effect and serves as a crucial debugging tool [5, 14]. However, its primary limitation is that it mainly validates the explanation for the injected feature's main effect, potentially not fully capturing faithfulness regarding complex interactions among the original features [14, 18, 22]. Despite this, positive results build confidence and provide an important sanity check, aligning with established principles for evaluating XAI against known or synthetic truths [3, 15].

3.4 Forecastability Analysis Implementation

To provide context for forecast and explanation reliability, we quantified the intrinsic predictability of each time series using the Spectral Predictability (SP) approach, assessing both its average level and its stability over time.

1. Preprocessing: Each series was preprocessed by removing leading zeros, yielding processed series y'_i with effective length L_i. The effective length L_i will be at least 4 times of prediction length based for the forecasting purpose.

2. Average SP & Benchmark Comparison: The overall SP score, SP_i, was computed for each processed series y'_i. To account for SP's length dependency, this score was compared against a length-specific pure noise benchmark, $SP_{noise}(L_i)$. Series with SP_i near or below this benchmark were flagged, suggesting predictability close to random noise.

4 RESULTS

This section presents and discusses the results from our experiments to evaluate the proposed framework. The experiments were developed using the M5 Forecasting Competition dataset[2], a real-world hierarchical collection featuring daily sales data from Walmart stores. This dataset spans 1,941 days and encompasses multiple organizational levels, including 10 stores, 7 departments, and 3 categories, tracking a total of 3,049 unique items. To determine the most appropriate forecasting level for our experiments, we conducted the spectral predictability analysis across different hierarchical levels, summarized in the Appendix (Table 1). This analysis indicated that the store-department level (70 time series) provided a strong balance of aggregation and inherent forecastability (approximately 94% of series deemed forecastable), leading us to select this granularity. Our study focuses on this subset of 70 store-department daily demand time series. An initial exploration revealed significant heterogeneity in demand scales across different store-department combinations, with mean daily demand varying considerably. This observation strongly motivated the need for a per-item (per-series) normalization strategy before modeling and explanation. For our experimental setup, we utilized the first 1,913 days of data as the training set and performed forecasting for the subsequent 28-day horizon. The actual data from days 1,914 to 1,941 (inclusive) served as the test set for forecast accuracy validation and explanation analysis.

4.1 Surrogate Model Fidelity

We first assessed the fidelity of a LightGBM surrogate in approximating the normalized point forecasts produced by the high-performing AutoGluon ensemble. Fidelity was quantified by comparing predictions from both models on the test set using per-series metrics, including mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared error (RMSE). Figure 1 illustrates a good alignment, with predicted values tightly clustered around the identity line. This high fidelity enables the use of the complex AutoGluon model for accurate forecasting, while facilitating post hoc interpretability via TreeSHAP applied to the computationally efficient LightGBM surrogate. This fidelity check is a first sanity check for such XAI system.

4.2 Surrogate Model Faithfulness

To further validate the reliability of the generated explanations, we evaluated explanation faithfulness through a feature injection experiment. This evaluation compared the denormalized SHAP values derived from the surrogate model for an injected feature (ϕ_{price}) against its known ground truth effect. The results demonstrated high faithfulness, with a Pearson correlation coefficient of 0.961 between the extracted SHAP values and the ground truth impacts. Figure 2 visually reinforces this result, showing a strong positive linear relationship. This high correlation validates faithfulness by confirming that the SHAP values correctly track the direction and relative magnitude of the known influence, even if the absolute scales on the plot differ due to the specific ground truth construction artifact. This evidence substantiates that the surrogate + SHAP framework can accurately recover the influence of individual features when the true effects are known, thereby reinforcing the trustworthiness of the explanations regarding feature effects. Nonetheless, as noted in prior discussions on the limitations of faithfulness evaluation, the reliable attribution of interaction effects remains an unresolved challenge. As such, generalizing the observed faithfulness from the injected feature's main effect to all original features warrants caution.

[2]Dataset available at: https://www.kaggle.com/competitions/m5-forecasting-accuracy/data

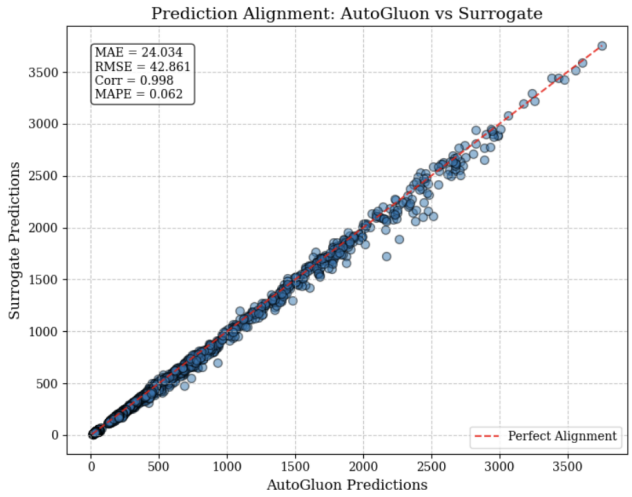

Figure 1: Prediction Alignment between AutoGluon Model and Surrogate Model

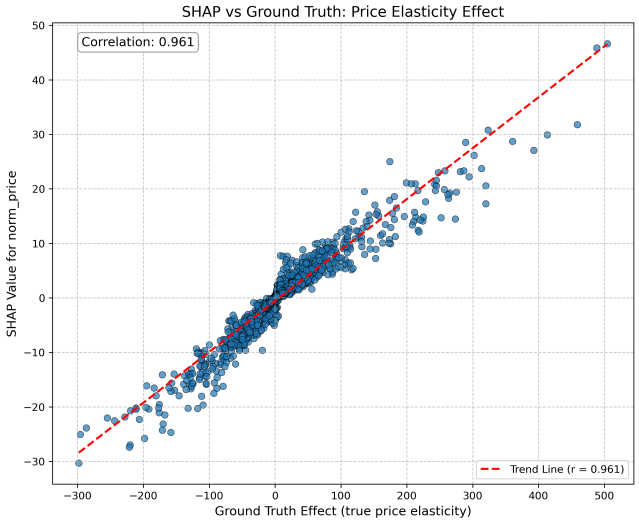

Figure 2: SHAP Faithfullness Validation by Price Effects

4.3 Necessity of Per-item Normalization

Our experiments empirically validated the critical necessity of per-item normalization discussed previously. For example, consider the items whose average demand varies by orders of magnitude, as

shown by example items in the Appendix (Table 2). As visually demonstrated for a low-volume item in the Appendix (Figure 4), applying SHAP without normalization results in distorted explanations. The large global base value forces compensatory SHAP values that obscure the true local feature impacts relative to the item's specific scale. In contrast, the results confirmed that per-item normalization yields explanations reflecting relative feature importance against an appropriate item-specific baseline, enabling meaningful interpretation and comparison across series with differing scales.

4.4 Forecastability Analysis

Beyond evaluating fidelity and faithfulness, we further analyzed relationship between forecastability and forecast accuracy and surrogate model fidelity. These results underscore the critical role of data properties in both generating reliable forecasts and interpreting model explanations.

Our analysis revealed strong positive correlations between average spectral predictability and both AutoGluon-TS forecast accuracy (Figure 3) and surrogate model fidelity (Figure 5 in the Appendix). This demonstrates that time series with more regular, predictable patterns not only yield more accurate forecasts but also enable more faithful mimicry by the surrogate model. The relationship becomes particularly evident at the lower end of the spectrum—both accuracy and fidelity deteriorate significantly for series with spectral predictability approaching or below the pure noise baseline. These findings highlight the practical value of spectral predictability as a diagnostic metric, offering a reliable mechanism to filter potentially problematic series or flag when explanations are generated for forecasts that may be inherently compromised by the unpredictable nature of the underlying data.

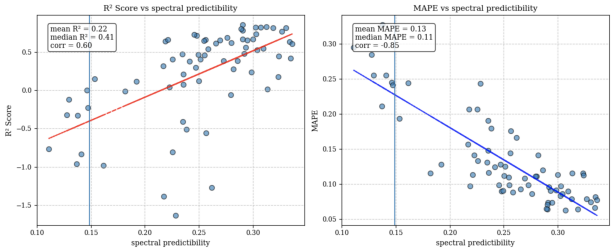

Figure 3: AutoGluon Forecast Accuracy vs. Average Spectral Predictability. Left: using R^2 metric. Right: using MAPE metric.

Note: The vertical line indicates the average spectral predictability score for pure noise, serving as a benchmark for randomness.

4.4.1 Case Examples: High vs. Low Forecastability. To illustrate the practical implications of forecastability, we examine two contrasting case studies, comparing the alignment of AutoGluon forecasts, surrogate model forecasts, and actual demand.

Illustrative case studies, detailed in the Appendix (Section D), contrast a high-forecastability series (`HOUSEHOLD_1` at `CA_3`, Appendix Figure 7) with a low-forecastability one (`HOBBIES_2` at `TX_1`, Appendix Figure 9). The high-forecastability example demonstrates close tracking between actual demand, AutoGluon-TS predictions, and surrogate forecasts, indicating high forecast accuracy and surrogate fidelity. In contrast, the low-forecastability example exhibits poorer accuracy for both models and noticeable divergence between the AutoGluon-TS and surrogate predictions, signifying lower fidelity. These cases visually confirm that higher intrinsic data predictability (context in Appendix Figures 6 and 8) supports both better forecast performance and more faithful surrogate model replication.

These examples visually reinforce the link between a time series' intrinsic forecastability and the performance achievable by both complex forecasting models and the surrogate models intended to explain them. High forecastability correlates positively with both forecast accuracy and surrogate fidelity, whereas low forecastability often leads to reduced performance on both fronts, highlighting the importance of considering data characteristics when interpreting model outputs and their explanations.

These empirical findings strongly advocate for integrating forecastability analysis directly into the explainability workflow. Quantitative forecastability metrics offer significant practical value by serving as an essential filter or confidence score when interpreting explanation results [12, 19]. By assessing metrics against predefined benchmarks (e.g., derived from noise), we can identify series where inherent data characteristics likely undermine model reliability. Low forecastability signals potential issues: the underlying forecast being explained may be inaccurate, the surrogate model might struggle to achieve high fidelity, and the feature relationships captured by the explanation could be unstable or noise-driven. Consequently, explanations associated with low-forecastability scores warrant lower confidence and can be appropriately flagged or filtered in practical applications. Explicitly acknowledging inherent data limitations via forecastability analysis thus provides a crucial layer of understanding. This helps manage user expectations regarding prediction certainty and prevents over-reliance on model outputs or their explanations when the underlying data simply does not support high confidence [18]. Ultimately, this combined framework of explainability and forecastability enables a more nuanced, trustworthy, and responsible interpretation of advanced forecasting systems.

5 CONCLUSION

This paper addressed the critical challenge of explaining complex ensemble time series forecasting models, specifically those generated by AutoML systems like AutoGluon, where direct explanation methods often prove unstable or infeasible. We proposed and validated a surrogate-based methodology, training a LightGBM model to mimic AutoGluon's point forecasts with high fidelity. We generated stable local feature attributions, whose faithfulness was confirmed via a rigorous feature injection experiment. We also demonstrated the crucial role of per-item normalization in enabling meaningful interpretations across heterogeneous time series. This work highlighted the significant benefit achieved by integrating these surrogate-based explanations with forecastability analysis (e.g., using spectral predictability). This combination allows users to calibrate their trust in an explanation based on the inherent

predictability, which we found directly correlates with both improved forecast accuracy and higher surrogate model fidelity, or lack thereof, in the underlying time series. This synergy bridges the gap between understanding model behavior and acknowledging data limitations, preventing over-reliance on potentially misleading explanations when data characteristics inherently limit predictive certainty. Such integrated approaches, balancing model interpretability with an understanding of data forecastability, are vital for the transparent and responsible deployment of advanced forecasting systems.

A PREDICTABILITY ANALYSIS RESULTS

Table 1: M5 Dataset Predictability Analysis by Hierarchical Level

Level	# Time Series	# Forecastable Series	Forecastable (%)
Store	10	10	100.0%
Department	7	7	100.0%
Store-Department	70	66	94.3%
Product	3049	496	16.3%

Note: A time series was deemed 'forecastable' if its mean spectral predictability score was higher than its baseline score. The implementation of the spectral predictability analysis is detailed in the main text.

B PER-ITEM NORMALIZATION ILLUSTRATION

Table 2: Example Items with Different Demand Scales

Item ID	Mean Weekly Demand	Std Dev Weekly Demand
Item 0 (High Vol)	1179.86	315.96
Item 1 (Med Vol)	116.80	26.83
Item 2 (Low Vol)	11.66	2.84

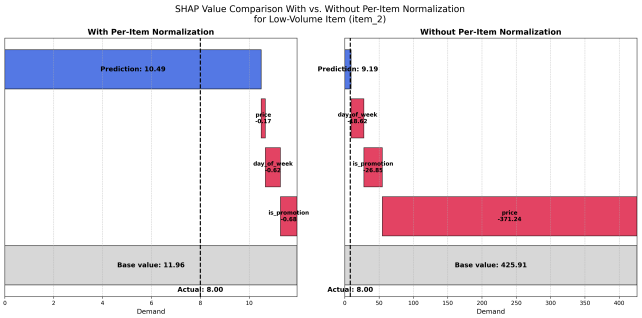

Figure 4: Conceptual Illustration of SHAP Values With and Without Per-Item Normalization for Items of Different Scales.

C CORRELATION BETWEEN SPECTRAL PREDICTABILITY AND FIDELITY

This section places the correlation plot regarding Spectral Predictability and fidelity. The plot has been referenced in the main body of the paper.

D EXAMPLE OF HIGH VS LOW FORECASTABILITY

This section provides the SHAP Explanations of two department-stores' forecasting as the illustrution. The HOUSEHOLD_1 of CA_3

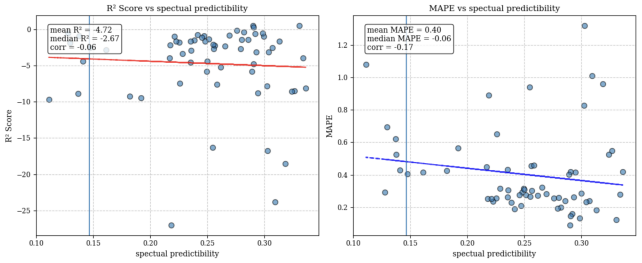

Figure 5: Surrogate Model Fidelity vs. Average Spectral Predictability. Left: using R^2 metric. Right: using MAPE metric.

store is highly forecastable (high spectral predictibility), while HOBBIES_2 at TX_1 store has low spectral predictibility.

First, consider HOUSEHOLD_1 at CA_3, identified as a series with relatively high forecastability, as shown in Figure 6 in the Appendix. Figure 7 in the Appendix displays the forecast results. The plot clearly shows strong, regular patterns in the actual demand (green line). This plot indicates high forecast accuracy for both the AutoGluon-TS ensemble in this high-forecastability scenario. Furthermore, the surrogate's predictions align tightly with AutoGluon-TS's predictions, demonstrating high fidelity.

In contrast, Figure 9 in the Appendix presents results for HOBBIES_2 at TX_1, a series characterized by a relatively low forecastability, as shown in Figure 8 in the Appendix. The actual demand here is much more erratic and lacks the clear seasonality seen previously. Consequently, both the AutoGluon and surrogate forecasts exhibit lower accuracy, struggling to capture the sharp peaks and troughs in the actual demand, although they follow the general level. Critically, there are noticeable divergences between the AutoGluon and surrogate model forecasts at several points across the horizon. This visual gap signifies reduced surrogate fidelity compared to the high-forecastability example, confirming that it is more challenging for the surrogate to mimic the AutoGluon-TS's behavior when the underlying series is inherently less predictable.

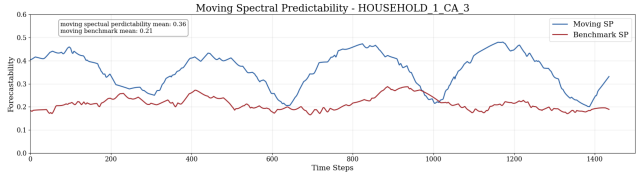

Figure 6: spectral Predictibility vs. Benchmark (random noise) for HOUSEHOLD_1 at CA_3 Store

E EXAMPLE OF FEATURE INJECTION FAITHFULNESS TEST

This section provides a conceptual illustration of the feature injection experiment used to evaluate explanation faithfulness, as discussed in Section 3.3. This method allows for quantitative assessment by creating a known ground truth effect for at least one feature, helping to validate if explanation techniques accurately capture feature influence.

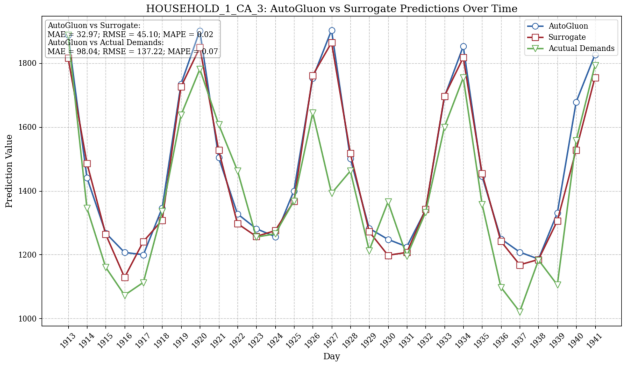

Figure 7: Comparison of Surrogate and AutoGluon Forecasts for HOUSEHOLD_1 at CA_3 Store

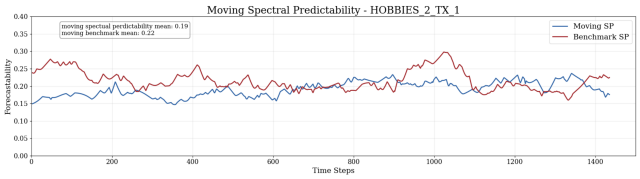

Figure 8: spectral Predictibility vs. Benchmark (random noise) for HOBBIES_2 at TX_1 Store

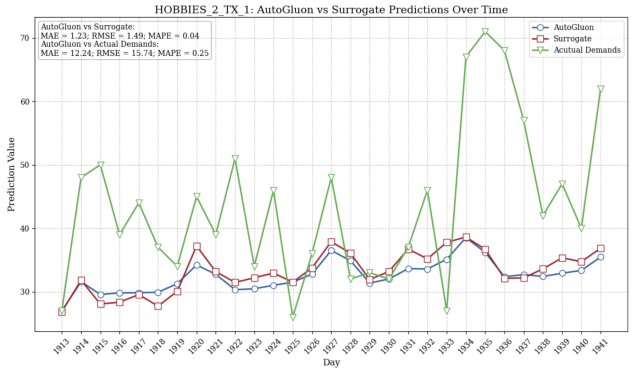

Figure 9: Surrogate vs. Auto Gluon Forecasting for HOBBIES_2 at TX_1 Store forecasting

Table 3: Conceptual Example of Feature Injection: Simulating Price Effect on Demand

ID	Date	Price ($)	Price Effect (On Demand)	Old Demand	New Demand
ITEM_A	2025-05-01	10.00	0	100	100
ITEM_A	2025-05-02	11.00	-10	102	92
ITEM_B	2025-05-01	20.00	0	50	50
ITEM_B	2025-05-02	21.00	-5	51	46
ITEM_B	2025-05-03	19.00	+5	49	54

F CALIBRATION

To potentially improve alignment between the explanation and the original model's prediction $\hat{y}_{AG}$, especially where the surrogate f_{LGBM} deviates significantly, a post-processing calibration step can be considered. Note this is an optional step and lacks any theoretical support. It is meant as a patch for a system requirements that must bridge the gap between the surrogate model's forecast and that of the AutoGluon TS. Acknowledging the SHAP additivity property ($f_{LGBM}(x) = \phi_0 + \sum_j \phi_j$, where ϕ_0 is the base value from the surrogate explanation and ϕ_j are the surrogate's SHAP values [17]), a more appropriate calibration aims to rescale the feature contributions (ϕ_j) so their sum matches the deviation of the original prediction from the base value ($\hat{y}_{AG} - \phi_0$).

This can be achieved heuristically by calculating a scaling factor s:

$$s = \frac{\hat{y}_{AG} - \phi_0}{f_{LGBM}(x) - \phi_0} = \frac{\hat{y}_{AG} - \phi_0}{\sum_j \phi_j}$$

assuming the denominator ($\sum_j \phi_j$) is non-zero. The calibrated SHAP values are then $\phi_{j,calib} = s \times \phi_j$. This factor s naturally handles potential sign inversions when ϕ_0 lies between $f_{LGBM}(x)$ and $\hat{y}_{AG}$.

Handling edge cases, such as $\hat{y}_{AG} = 0$ while $f_{LGBM}(x) \neq 0$, requires careful consideration:

- **If surrogate predicts base value ($f_{LGBM}(x) = \phi_0$):**
  - If the target model also predicts the base value ($\hat{y}_{AG} = \phi_0$), both models agree on zero deviation from the baseline. No calibration is needed; the original (zero-sum) surrogate SHAP values ϕ_j can be used.
  - If the target model predicts differently ($\hat{y}_{AG} \neq \phi_0$), calibration is mathematically impossible (division by zero: $\sum_j \phi_j = 0$). This indicates a significant local mismatch where the surrogate fails to capture the target's deviation. Action: Issue a warning about the fidelity failure and either skip displaying calibrated values or show the original ϕ_j with a strong caveat.
- **If target model predicts zero ($\hat{y}_{AG} = 0$):**
  - Calibration can generally proceed using the formula $s = -\phi_0/(f_{LGBM}(x) - \phi_0)$, as long as the surrogate prediction is not exactly the base value (i.e., $f_{LGBM}(x) \neq \phi_0$).
  - However, consider the practical context: for zero forecasts often driven by specific business rules (e.g., stock-outs, discontinued items), providing a rule-based explanation might be more insightful than displaying potentially complex calibrated SHAP values that sum to $-\phi_0$.

We acknowledge that any such calibration remains a heuristic adjustment applied post-hoc [18]. It lacks strong theoretical grounding within the SHAP framework, as it modifies explanation values based on outcomes rather than directly reflecting the (surrogate) model's internal logic that generated the original ϕ_j. It should be applied cautiously, primarily serving as a pragmatic way to reconcile explanation additivity with the target prediction when fidelity is imperfect [21, 22]. Hence it is optional depends on the XAI system requirements.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. doi:10.1109/ACCESS.2018.2870052

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems 31*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). Curran Associates, Inc., 9505–9515. https://papers.nips.cc/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf

[3] Angel Angelov, Carlos Perez-Ferrer, Yoana Krasteva, and Stefano Nichele. 2023. Evaluating Correctness and Robustness of Local Anomaly Explanations Using Ground Truth. *Algorithms* 16, 12 (2023), 545. doi:10.3390/a16120545

[4] Goran Antic and Vladimir Mihajlovic. 2021. Principles and Practice of Explainable Machine Learning. *Information* 12, 8 (2021), 324. doi:10.3390/info12080324

[5] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608 (Position paper at ICML 2017 Workshop on Human Interpretability in Machine Learning). arXiv:1702.08608 https://arxiv.org/abs/1702.08608

[6] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. arXiv preprint arXiv:2003.06505. arXiv:2003.06505 https://arxiv.org/abs/2003.06505

[7] Georg M. Goerg. 2013. Forecastable Component Analysis. *Journal of Machine Learning Research* 14 (2013), 2967–3000. http://jmlr.org/papers/v14/goerg13a.html

[8] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5, Article 93 (2019), 42 pages. doi:10.1145/3236009

[9] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. arXiv preprint arXiv:1812.04608. arXiv:1812.04608 https://arxiv.org/abs/1812.04608

[10] Giles Hooker and Lucas Mentch. 2019. Please Stop Permuting Features: An Explanation and Alternatives. arXiv preprint arXiv:1905.03151. arXiv:1905.03151 https://arxiv.org/abs/1905.03151

[11] Rob J. Hyndman, Roman A. Ahmed, George Athanasopoulos, and Han Lin Shang. 2011. Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis* 55, 9 (2011), 2579–2589. doi:10.1016/j.csda.2011.03.006

[12] Yanfei Kang, Rob J. Hyndman, and Kate Smith-Miles. 2017. Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting* 33, 2 (2017), 345–358. doi:10.1016/j.ijforecast.2016.09.004

[13] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). Curran Associates, Inc., 3146–3154. https://papers.nips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[14] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking Explainability as a Dialogue: A Practitioner's Perspective. arXiv preprint arXiv:2201.08164 (Presented at NeurIPS 2022 Workshop on Human-Centered AI). arXiv:2201.08164 https://arxiv.org/abs/2201.08164

[15] Xuhong Li, Mengnan Du, Jiamin Chen, Yekun Chai, Haoyi Xiong, and Himabindu Lakkaraju. 2023. $\mathcal{M}^4$: A Unified XAI Benchmark for Faithfulness Evaluation of Feature Attribution Methods across Metrics, Modalities and Models. In *Advances in Neural Information Processing Systems 36*, Amos Storkey and Po-Ling Loh (Eds.). https://proceedings.neurips.cc/paper_files/paper/2023/hash/05957c194f4c77ac9d91e1374d2def6b-Abstract-Datasets_and_Benchmarks.html

[16] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex John DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence* 2, 1 (2020), 56–67. doi:10.1038/s42256-019-0138-9

[17] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). Curran Associates, Inc., 4765–4774. https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[18] Christoph Molnar. 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). Self-published. https://christophm.github.io/interpretable-ml-book/

[19] Fotios Petropoulos, Daniele Apiletti, Vassilios Assimakopoulos, Mohamed Zied Babai, Devon K. Barrow, Souhaib Ben Taieb, Christoph Bergmeir, Ricardo J. Bessa, Jakub Bijak, John E. Boylan, Jethro Browell, Claudio Carnevale, Jennifer L. Castle, Pasquale Cirillo, Michael P. Clements, Clara Cordeiro, Fernando Luiz

Cyrino, Shari De Baets, Alexander Dokumentov, Joanne Ellison, Piotr Fiszeder, Philip Hans Franses, David T. Frazier, Michael Gilliland, M. Sinan Gonul, Paul Goodwin, Luigi Grossi, Yael Grushka-Cockayne, Mariangela Guidolin, Massimo Guidolin, Ulrich Gunter, Xiaojia Guo, Renato Guseo, Nigel Harvey, David F. Hendry, Ross Hollyman, Tim Januschowski, Jooyoung Jeon, Victor Richmond R. Jose, Yanfei Kang, Anne B. Koehler, Stephan Kolassa, Nikolaos Kourentzes, Sonia Leva, Feng Li, Konstantia Litsiou, Spyros Makridakis, Gael M. Martin, Andrew B. Martinez, Sheik Meeran, Theodore Modis, Konstantinos Nikolopoulos, Dilek Onkal, Alessia Paccagnini, Anastasios Panagiotelis, Ioannis Panapakidis, Jose M. Pavia, Manuela Pedio, Diego J. Pedregal, Pierre Pinson, Patrícia Ramos, David E. Rapach, James Reade, Bahman Rostami-Tabar, Michał Rubaszek, Georgios Sermpinis, Han Lin Shang, Evangelos Spiliotis, Aris Syntetos, Priyanga Dilini Talagala, Thiyanga S. Talagala, Len Tashman, Dimitrios Thomakos, Thordis Thorarinsdottir, Ezio Todini, Juan Ramon Trapero Arenas, Xiaoqian Wang, Robert L. Winkler, Alisa Yusupova, and Florian Ziel. 2022. Forecasting: theory and practice. *International Journal of Forecasting* 38, 3 (2022), 705–871. doi:10.1016/j.ijforecast.2021.11.001

[20] Vikas C. Raykar, Arindam Jati, Sumanta Mukherjee, Nupur Aggarwal, Kanthi Sarpatwar, Giridhar Ganapavarapu, and Roman Vaculin. 2023. TsSHAP: Robust model agnostic feature-based explainability for univariate time series forecasting. arXiv:2303.12316 [cs.LG]

[21] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215. doi:10.1038/s42256-019-0048-x

[22] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David Inouye, and Pradeep Ravikumar. 2019. On the (In)fidelity and Sensitivity of Explanations. In *Advances in Neural Information Processing Systems 32*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett (Eds.). Curran Associates, Inc., 10967–10978. https://papers.nips.cc/paper/2019/file/1e1f446d42a2c439b897afe3a2494724-Paper.pdf