Enhancing Financial Forecasting Accuracy through Spatiotemporal Transformer Models: A Novel Approach to EPS Surprise Prediction in the S&P 500

Adakole Ebute
Minerva University

Abstract

This paper presents a comprehensive analysis of financial forecasting models, specifically focusing on predicting Earnings Per Share (EPS) surprises using spatiotemporal feature representations in transformer architecture under a skew-normal distribution framework. Through rigorous experimentation, we assess the models' probabilistic forecasting capabilities, point estimate accuracy, and attention mechanisms across various configurations. Our findings reveal that while spatiotemporal models exhibit narrower confidence intervals, indicating a more precise probabilistic outlook, they encounter challenges with point estimate metrics such as MAE and MSE. This dichotomy underscores the models' nuanced trade-offs between achieving tight probabilistic forecasts and precise point estimates. Additionally, the analysis of attention mechanisms provides insights into how each model processes and prioritizes information, revealing minimal emphasis on categorical variables like industry classifications in forecasting outcomes. Despite their potential, the wide confidence intervals produced by both models highlight a significant limitation in their current implementation for practical financial applications. The study concludes that further refinement is necessary, including the integration of a broader spectrum of market dynamics indicators, to enhance predictive accuracy and informativeness. This research suggests that the developed frameworks could be adapted to other analytical domains where understanding asymmetry and bias is crucial, although the current model performance remains inconclusive regarding the nature of analyst bias in EPS predictions. This work lays the groundwork for future advancements in financial forecasting, emphasizing the need for ongoing innovation and exploration in this evolving field.

1 Introduction

The quest for a superior alternative to analysts' earnings forecasts has been a subject of considerable debate and research within the financial analysis community. A pivotal contribution to this discussion is evident in the prominence of model-based financial forecasting literature, which raises critical questions about the potential of machine learning models to provide less biased and more accurate forecasts than traditional analyst estimates (Harris and Wang, 2019). This research paper, drawing upon the foundational work of Liu et al. (2022), ventures into the realm of machine learning (ML) to forecast Earnings Per Share (EPS) surprise, a critical factor influencing investor sentiment and market dynamics. The EPS surprise metric, representing the divergence between actual and anticipated earnings, has been a pivotal focus due to its significant impact on market reactions and investor decisions. This research paper seeks to explore the extent to which advancements in machine learning, specifically through the lens of EPS Surprise, can bridge the gap between human analysts and automated systems in predicting earnings.

Analysts' forecasts have been consistently shown to exhibit an optimistic bias under certain conditions, often underreacting to negative earnings news while overreacting to positive news (Abarbanell and Bernard, 1992; De Bondt and Thale, 1990; Easterwood and Nutt, 1999; Lim, 2001; Wu et al., 2018). This tendency towards optimism has been attributed to a variety of factors, including cognitive biases and the presence of conflicts of interest (Easterwood and Nutt, 1999; Krolikowski, Chen and Mohr, 2016; Wang et al., 2017; Xu et al., 2013; Francis and Philbrick 1993; Dugar and Nathan 1995; Lin and McNichols 1998; Ramnath et al. 2008; Harris and Wang, 2019). Conversely, machine learning models, free from such biases, present a

unique opportunity to forecast financial outcomes based on historical data and quantitative analysis. The investigation is anchored in the hypothesis that these biases can be systematically analyzed and predicted as a function of historical company fundamentals and market dynamics, with minimal influence from external factors.

The selection of financial indicators, model architecture and hyperparameters, loss functions, and data imputation techniques employed in this study are deeply rooted in a comprehensive review of existing literature. This spans both ML-specific methodologies and broader financial analysis frameworks (Liu et al., 2022; Wasserman, Nohria, & Anand, 2010; Grigsby et al., 2021). Such a literature-backed approach has guided the adaptation and refinement of attention-based models, probabilistic forecasting, and spatiotemporal dynamics to tailor the predictions to the nuances of the EPS surprise metric. This interdisciplinary approach ensures that the proposed method is not only theoretically sound but also practically viable across various market conditions.

By focusing on predicting the bias present in earnings forecasts through EPS surprise, rather than directly forecasting earnings, this paper aims to address several challenges highlighted in the literature. These include the optimistic bias of analysts, the limitations of existing models in capturing the full spectrum of market and company-specific dynamics, and the inherent difficulty in accurately forecasting earnings using external information (Mendenhall, 1991; Ota, 2011; Tse & Yan, 2008; Lim, 2001; Kerl and Ohlert, 2015; Easterwood and Nutt, 1999; Wu et al., 2018). This paper also seeks to bolster these advantages by focusing on a model that is generalizable across industries, thereby addressing the analysts' need for industry-specific specialization. The methodology attempts to establish a robust yet nuanced understanding of earnings predictions by leveraging attention-based models and the latest advancements in sequence-to-sequence machine learning, thereby proposing an innovative approach to financial forecasting that is both more accurate and less susceptible to bias.

In essence, this paper is motivated by the limitations observed in both analyst-driven forecasts and existing machine-learning models. By focusing on short-run predictions, there's a hope to provide a viable alternative that combines the best of both worlds: the nuanced understanding of market dynamics possessed by analysts and the analytical precision of machine learning models. The contributions of this paper are as follows:

1. The introduction of a novel ML-based approach, centralized on the Spacetimeformer Architecture (Grigsby et al., 2021), for forecasting EPS surprise, aimed at reducing the systematic bias prevalent in analysts' forecasts.

2. The application of transformer models to financial forecasting, leveraging their capability to discern intricate patterns in sequential data and communicate the time and/or indicators in the sequence that were critical to any observed values.

3. The development of a model that is not only capable of rivaling analyst forecasts in the short run but is also generalizable across different industries, thereby broadening its applicability.

The remainder of this paper is organized as follows: Section II provides a comprehensive background on the financial indicators critical to understanding EPS surprise and reviews the state-of-the-art in time-series forecasting, emphasizing the novel application of transformer and attention-based models within this domain. Section III details the proposed methodology, from the selection of financial indicators and the adjustments made to standard transformer models to better suit financial forecasting, to the innovative approach to handling missing data. The experimental setup, described in Section IV, outlines the data preparation processes, hyperparameter selection rationale, and the multifaceted loss definition designed to refine forecast accuracy. Section V showcases the results of preliminary and main

experiments, illustrating the efficacy of the approach through comparative analysis and statistical evaluation. Finally, Section VI concludes the paper with a discussion of the implications of the findings, the limitations encountered, and prospective directions for future research.

2 Background and Related Work

2.1 Background

2.1.1 Financial Indicators in Time-series

The exploration of financial indicators and their impact on EPS surprise forms a foundational element of this research paper. Financial indicators such as revenue, net income, and various cost metrics are integral components that, when analyzed collectively, provide a comprehensive view of a company's financial health and performance. These indicators are not standalone metrics but are interconnected in ways that directly influence the calculation of EPS and, subsequently, the EPS surprise.

Financial Indicators and EPS Calculation

EPS is a widely regarded profitability measure, calculated as the net income divided by the outstanding shares of a company. This basic formula encapsulates the essence of financial performance distilled into a per-share basis, making it a critical metric for investors. The calculation of EPS surprise, then, involves comparing the actual EPS reported by a company to the EPS estimates provided by financial analysts prior to earnings announcements. The formula can be expressed as follows:

$$\text{EPS Surprise} = \frac{\text{Actual EPS} - \text{Estimated EPS}}{\text{Estimated EPS}}$$

Given that net income—a key variable in the EPS formula—is influenced by revenue, costs, and other income statement items, all these factors are inherently interconnected within the EPS surprise metric.

Time-Series Nature of EPS Surprise

For a variable to be classified as a time series, it must demonstrate autocorrelation, meaning it is correlated with its own past and future values. This can apply to variables that are inherently autocorrelated, such as revenue or costs, which tend to rise or fall over time, or to variables that are functions of autocorrelated variables. EPS, being a function of net income (itself a result of autocorrelated variables like revenue and costs), is inherently autocorrelated. Applying the same definition, EPS surprise also qualifies as a time-series variable since it is derived from the fluctuations and trends in actual EPS compared to estimated EPS over time.

Limitations of Financial Data Sourcing

Access to comprehensive and accurate financial data poses a significant challenge in financial forecasting. Unlike many datasets that are freely available, detailed financial data is often proprietary and sold at a premium. This limitation is particularly acute for individual-level forecasting, where access to sell-side data, which includes detailed analyst forecasts and proprietary financial metrics, is restricted. The advent of crowdsourcing websites represents an innovative response to this challenge, providing a platform for aggregating and sharing financial forecasts and estimates.

Accuracy of Analyst Estimates

It is well-documented that analyst estimates tend to be more accurate than baseline models, owing to the vast amount of information utilized in building these models and the complex interplay among various pieces of financial and market data (Harris & Wang, 2019). However, this paper posits that the bulk of market dynamics and information interactions are encapsulated within historical market data, which can be leveraged to effectively predict analysts' biases without the need to include market sentimental factors. Empirical evidence supports the notion that analysts' superior performance is attributed to their access to a broader spectrum of frequently observed information (Fried and Givoly 1982; Kross et al. 1990; Alford and Berger 1999; Sougiannis and Yaekura 2001; Ball & Ghysels, 2018; Groysberg et al., 2011).

Overarching Financial Goal
This paper navigates the inquiry: Can earnings forecasts be fully automated by machine-driven models, or can these models provide complementary information to analysts, thereby enhancing the quality of earnings forecasts? (Ball & Ghysels, 2018) The ambition of this research is to affirm the latter by offering insights into analysts' biases through the analysis of financial indicators and EPS surprise. By forecasting EPS surprise rather than directly predicting EPS, the model aims to encapsulate the level of analysts' bias, accounting for the randomness in company performance and the imperfect information available to analysts. This approach not only highlights the practical benefits of predicting EPS surprise but also underscores the limitations inherent in sourcing and analyzing financial data.

In conclusion, the research advances the understanding of financial indicators in forecasting EPS surprise, providing a nuanced perspective on the time-series nature of financial metrics and the limitations faced in financial data sourcing. Through this exploration, the paper aspires to enrich the dialogue on machine learning's role in financial forecasting and its capacity to offer complementary insights alongside traditional analyst estimates.

2.1.2 Probabilistic Forecasting

Probabilistic forecasting represents a significant paradigm shift from deterministic forecasting methods. Instead of providing a single value as a prediction, probabilistic forecasts generate a range of possible outcomes along with their associated probabilities. This approach is particularly valuable in financial forecasting, where uncertainty is a constant factor, and the stakes of decisions are high.

In the context of EPS surprise forecasting, employing confidence intervals or prediction broadbands allows stakeholders to understand not just what the model predicts will happen, but also the model's confidence in those predictions. For instance, a forecast might predict an EPS surprise of 5% with a 95% confidence interval ranging from 2% to 8%, indicating that there is a 95% probability the actual EPS surprise will fall within that range. Present literature indicates that even a 10% forecast error magnitude, or in this case an EPS surprise of 0.1, is enough to trigger significant stock reactions, emphasizing the critical role of conveying uncertainty in forecasts (Dreman & Berry, 1995).

The adoption of probabilistic forecasting in financial analysis has profound practical implications. It enables investors, analysts, and other stakeholders to make more informed decisions by considering not just the most likely outcome but also the range of possible outcomes. This approach is especially beneficial in risk management, portfolio planning, and strategic decision-making processes, where understanding the breadth of potential scenarios is as crucial as predicting the most likely future state.

2.1.3 Attention-based Models

Attention-based models, notably Transformers (Vaswani et al.; 2017), have revolutionized the landscape of machine learning, impacting various fields such as natural language processing, image recognition, and, increasingly, financial forecasting. The introduction of the Transformer architecture marked a significant departure from traditional recurrent and convolutional neural networks, particularly for sequence-to-sequence tasks. This innovation is driven by the model's capacity to process sequential data in parallel, thereby capturing long-range dependencies more effectively—a capability of paramount importance in financial forecasting, especially for predicting EPS surprise that is extrinsically dependent on several other indicators.

Background on Attention Mechanism
The core innovation of the Transformer model lies in its self-attention mechanism. This mechanism allows the model to evaluate the sequence in its entirety, assigning varying degrees of importance or "attention" to different elements based on their relevance to the task at hand. In the domain of financial forecasting, this means that the Transformer can distinguish between various financial indicators' impacts on

EPS surprise, considering factors like revenue trends, industry effect, or market dynamics in their entirety rather than in isolation.

Formulas for Attention Calculation
The self-attention mechanism's mathematical foundation is built upon calculating attention scores, which determine the influence of each sequence element on the others. Given a sequence of vectors X, where each vector x_i represents the features for a specific time period (e.g., a financial quarter), the attention score between two elements x_i and x_j is computed using the formula:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

, where:
- Q, K, V are the query, key, and value matrices derived from the input sequence, respectively.
- d_k represents the dimensionality of the key vectors, providing a scaling factor to ensure the stability of the softmax function.

This formulation allows the model to assess the relevance of all elements within the sequence, prioritizing those with higher relevance to the forecasting task.

Multiple Layers and Attention Heads
The Transformer architecture employs multiple layers and attention heads to refine its understanding of the sequence. Each layer allows the model to process the information at a different level of abstraction, while multiple attention heads enable it to focus on different parts of the sequence simultaneously. This multi-headed approach allows the Transformer to capture a wide array of relationships within the data, from the most direct to the more subtle and complex ones.

The implications of employing multiple layers and attention heads are significant. They allow the Transformer model to:
- Handle multivariate predictors effectively, crucial for financial forecasting where the prediction of EPS surprise requires analyzing a plethora of interconnected financial indicators.
- Capture long-range dependencies and complex temporal patterns, overcoming the limitations of traditional time-series models that may struggle with such complexity.

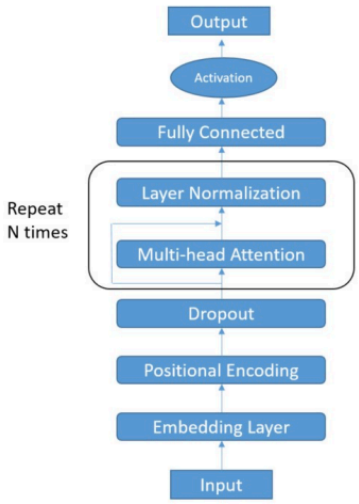

Figure 1: Vanilla architecture for Attention-based model as highlighted in Liu et al. (2022).

Application in Financial Forecasting
By leveraging the self-attention mechanism, models can prioritize relevant financial indicators and trends over various time periods, offering a more detailed and comprehensive forecast of EPS surprise. This ability to discern and emphasize crucial pieces of information makes the Transformer exceptionally suited for financial forecasting tasks where understanding the interplay between different indicators is key to predicting market movements and financial outcomes.

2.2 Related Work

2.2.1 Architectural and Modular Variants of the Vanilla Transformer for Time Series Forecasting

Recent advances in machine learning have witnessed the introduction and evolution of Transformer models beyond their initial application in natural language processing to various domains, including time series forecasting. Wen et al. (2023) systematically review these developments, particularly

emphasizing the architectural and modular innovations tailored for time series analysis. Here, the variants of the vanilla Transformer as mentioned in the survey are summarized, focusing on their application to financial forecasting.

Positional Encoding Innovations

- Learnable Positional Encoding: To overcome the limitations of fixed encoding, learnable embeddings were introduced, allowing the model to adapt positional information to the specific characteristics of the time series data, leading to more effective forecasting models (Zerveas et al., 2021; Lim et al., 2021).

- Timestamp Encoding: Acknowledging the rich information in time series timestamps, some models have integrated timestamp data as additional positional encodings, significantly enhancing the model's ability to interpret and utilize temporal information (Zhou et al., 2021; Wu et al., 2021; Zhou et al., 2022).

Application-Specific Transformers

- Time Series Forecasting: Tailored Transformer variants have been developed to address the unique challenges of time series forecasting, focusing on capturing temporal dependencies and seasonality, with models like Autoformer (Wu et al., 2021) and FEDformer (Zhou et al., 2022) demonstrating significant advancements.

- Spatio-Temporal Forecasting: Crossformer (Zhang and Yan, 2023) utilizes cross-dimension dependency for multivariate time series forecasting by embedding the variates into a 2D vector array through a novel dimension-segment-wise embedding to preserve time and dimension information. Spacetimeformer (Grigsby et al., 2021) flattens the time-separated multivariates into one long stream of variables thus allowing the transformer to better learn the distinct temporal and spatial relationships.

This comprehensive exploration of Transformer variants for time series underscores the model's versatility and its potential to revolutionize financial forecasting. By addressing specific challenges such as long-range dependency modeling, efficiency, and multi-resolution analysis, these innovations pave the way for more accurate, robust, and scalable forecasting models.

2.2.2 Baseline Model Choice: SpaceTimeformer

The SpaceTimeFormer, as described earlier, represents a notable advancement in the application of Transformer models to the field of time series forecasting, particularly in financial contexts. This model, through its innovative architecture, addresses two fundamental challenges in time series analysis: capturing intricate temporal dynamics and understanding spatial relationships among multiple variables over time.

Dynamics of the SpaceTimeFormer
The SpaceTimeFormer distinguishes itself by effectively flattening spatio-temporal data into a unified sequence of variables. This architecture deviates from traditional models that might treat spatial and temporal dimensions separately or require complex mechanisms to intertwine these dimensions. By presenting the data as a single stream, the SpaceTimeFormer simplifies the model's task of learning dependencies, allowing for a more straightforward and potentially more powerful analysis of the relationships within the data.

This approach leverages the inherent strengths of the Transformer architecture—namely, its ability to capture long-range dependencies and model interactions across the entire input sequence. The model's attention mechanism can thus operate over the combined spatiotemporal data, identifying relevant patterns and relationships that influence the forecasting task.

Why is the SpaceTimeFormer a Good Choice?

- Simplicity and Clarity: One of the primary advantages of the SpaceTimeFormer is its simplicity. By reducing the complexity involved in handling separate spatial and temporal modules, the model can focus on learning the core dynamics within the data. This simplicity also translates to clarity in the model's learned representations, making it easier to interpret the influence of various factors on the forecasted outcomes.

- Enhanced Visibility of Spatio-Temporal Relationships: The unified sequence approach enhances the model's ability to discern both spatial and temporal relationships within the data. In financial forecasting, where the interplay between different variables (such as market indicators, industry factors, and economic factors) over time is crucial, this ability can lead to more accurate and insightful predictions.

- Efficiency and Scalability: By flattening the spatio-temporal dimensions, the SpaceTimeFormer can potentially offer improvements in computational efficiency, making it well-suited for handling large-scale financial time series data. This efficiency does not come at the expense of model performance, as the Transformer's attention mechanism remains highly capable of modeling complex dependencies.

- Adaptability to Financial Forecasting: The characteristics of financial time series data—marked by volatility, complex dependencies, and the influence of external factors—demand a model capable of understanding nuanced relationships. The SpaceTimeFormer's design aligns well with these requirements, offering a promising avenue for capturing the unpredictable nature of financial markets and providing insights into future movements.

To summarize, the SpaceTimeFormer introduces a novel method for handling spatiotemporal data, rendering it an attractive option for financial forecasting. By streamlining and elucidating the examination of intricate interrelations within time series data, alongside leveraging the Transformer architecture's inherent capabilities, it emerges as an essential instrument for those in research seeking an understanding of the dynamics in financial markets.

2.2.3 Seasonal Decomposition in Time-series Forecasting

Recent research, including studies by Wu et al. (2021), Zhou et al. (2022), Lin et al. (2021), and Liu et al. (2022a), have started to acknowledge the importance of incorporating seasonal-trend decomposition techniques, as originally proposed by Cleveland et al. (1990) and further examined by Wen et al. (2020), into Transformer-based models. This addition is not merely a supplemental enhancement but a core component that significantly amplifies the model's predictive accuracy.

Wen et al. (2023) demonstrate through experimental evidence that implementing a simple moving average seasonal-trend decomposition architecture can lead to substantial improvements in the model's forecasting capability—specifically, a performance boost ranging from 50% to 80%. This dramatic increase in efficacy underscores the decomposition's role in enabling the Transformer to more effectively capture and utilize the underlying patterns within time series data, such as seasonality and trend components that are prevalent in financial markets.

The significance of seasonal-trend decomposition lies in its ability to disentangle the complex data into more manageable components, thereby allowing the Transformer model to focus on and learn from these distilled elements separately. This methodological approach enhances the model's ability to identify and react to the specific dynamics of each component, whether it be the recurring patterns of seasonality or the overarching directionality of trends.

Why is this important for the paper? The incorporation of seasonal trend decomposition aligns with the objective of advancing financial forecasting methods. Financial markets are characterized by pronounced seasonal effects and long-term trends, influenced by cyclical economic factors, fiscal policies, and market sentiment. By dissecting these elements and analyzing them within the Transformer framework, the paper aims to achieve a more nuanced and accurate modeling of financial time series.

3 Proposed Method

3.1 Financial Variable Selection

The approach to variable selection in this paper draws significant inspiration from the pioneering work by Liu, Ouyang, & Xu (2022), which focused on predicting Earnings Per Share (EPS) Surprise within Asian markets utilizing deep learning techniques. The methodology outlined in their study provides a solid foundation for this endeavor to forecast EPS Surprise, albeit with a strategic shift toward companies listed on the US Stock Exchange. This adaptation necessitates a careful selection and substitution of financial indicators to ensure relevance and applicability to the distinct characteristics of the US financial landscape.

3.1.1 Adapting Indicators for the US Stock Exchange

The indicators identified by Liu et al. (2022) as possessing the highest predictive power for EPS Surprise include a mix of financial fundamental factors, consensus estimate data, and market indicators. To tailor these indicators for the US market, a detailed analysis is undertaken to align them with available data points and the specific nuances of US companies. This process involves a substitution process where direct equivalents of the indicators used in the Asian markets are not applicable or available for US companies. Close substitutes are sought that can offer comparable predictive insights. For example, if certain financial metrics are based on indexes or characteristics of the Asian markets, adjustments are made to their US counterparts, ensuring consistency with US accounting standards and practices.

3.1.2 Rationale for Indicator Selection

The rationale behind adopting and adapting the financial indicators from Liu et al.'s study stems from the proven predictive efficacy of these variables in a deep-learning context (2022). By leveraging indicators that have already demonstrated significant predictive power in Asian markets, the aim is to construct a robust model that can effectively identify potential EPS Surprises in the US market. Moreover, through thoughtful substitutions and additions, the model is finely tuned to the specificities and idiosyncrasies of US companies, thereby enhancing its relevance and accuracy.

3.2 Probabilistic Distribution Output: The Skew-Normal Distribution

In the pursuit of refining the predictive accuracy and relevancy of the model for forecasting EPS Surprise, a pivotal decision was made to employ the skew-normal distribution as the probabilistic output. This choice was motivated by two primary considerations: the necessity to capture the inherent asymmetry in financial predictions, and the specific aim to delve into the optimistic bias often observed in earnings forecasts. The skew-normal distribution, with its ability to model data that deviates from a symmetrical distribution, emerges as an ideal fit for these objectives.

3.2.1 Asymmetry in Financial Predictions

Financial data, especially metrics like EPS Surprise, often exhibit asymmetrical patterns. Traditional normal distributions, while useful in many scenarios, fall short of accurately representing these patterns due to their inherent symmetry about the mean. The skew-normal distribution, however, incorporates a shape parameter that allows for the modeling of data

skewness, making it adept at capturing the asymmetric nature of financial predictions.

The probability density function (PDF) of the skew-normal distribution is given by:

$$f(x;\xi,\omega,\alpha) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\left(\frac{x-\xi}{\omega}\right)\right)$$

where:
- x is the variable of interest (EPS Surprise in this case),
- ξ denotes the location parameter (mean),
- ω represents the scale parameter (standard deviation),
- α is the shape parameter that introduces skewness,
- ϕ is the standard normal probability density function, and
- Φ is the standard normal cumulative distribution function.

This formulation allows the skew-normal distribution to adeptly model datasets that are not symmetrically distributed, providing a more nuanced and accurate representation of the actual financial phenomena.

3.2.2 Investigating Optimistic Bias in Earnings Forecasts

The second driving force behind selecting the skew-normal distribution is its utility in investigating the optimistic bias inherent in earnings forecasts. The shape parameter α not only facilitates the representation of skewed data but also enables a quantitative analysis of the direction and magnitude of this bias. By analyzing the skewness parameter, researchers and practitioners can gauge the extent to which forecasts are skewed positively, indicating an optimistic bias, or negatively, suggesting a pessimistic bias.

3.3 Categorical Variable Representation

The representation of categorical variables within the embedding layers is crucial for enhancing the model's ability to discern patterns and make accurate predictions using identifier information. This process involves embedding discrete categorical data into continuous vector spaces, enabling the model to efficiently process and utilize this information in forecasting tasks. The integration process for these variables differs based on whether the model operates in a temporal-only or spatio-temporal feature representation. Below, the implementation process for both cases is detailed, emphasizing the SpaceTimeFormer approach.

3.3.1 Temporal Feature Representation

In the temporal-only scenario, categorical variables are embedded separately and then concatenated with continuous variables before being passed to the linear layer for dimensionality matching so as to ensure a single feature vector. This process is illustrated by the following:

Let $\mathbf{y}$ be the input vector containing continuous variables and $\mathbf{c}_i$ represent the i-th categorical variable. The embedding of $\mathbf{c}_i$ through the embedding layer E_i is given by:

$$\mathbf{e}_i = E_i(\mathbf{c}_i)$$

The final input vector $\mathbf{y}'$ that combines continuous and categorical embeddings is:

$$\mathbf{y}' = \text{concat}(\mathbf{y}, \mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n)$$

where n is the number of categorical variables.

This vector $\mathbf{y}'$ is then passed through a linear transformation $\mathcal{L}$ to match the model's dimensionality:

$$\mathbf{y}'' = \mathcal{L}(\mathbf{y}')$$

3.3.2 Spatio-Temporal Feature Representation

In the spatio-temporal scenario, particularly with the implementation of the SpaceTimeFormer, the process involves flattening the continuous variables into a long stream alongside the

categorical variables. Given the disparate dimensionalities, continuous vectors are upscaled with zeros to match the dimensionality of categorical embeddings. This operation can be expressed as follows:

Given a continuous vector $\mathbf{y}$ and categorical embeddings $\mathbf{e}_i$ as defined above, the spatio-temporal input for each time-step $\mathbf{y}_{st}$ before linear transformation is obtained by:

$$\mathbf{y}_{st} = \text{flattened}(\mathbf{y} \oplus \mathbf{0}, \mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n)$$

where $\mathbf{0}$ is a zero vector of dimensionality such that $\dim(\mathbf{y} \oplus \mathbf{0}) = \dim(\mathbf{e}_i)$, ensuring dimensionality matching between continuous and categorical data.

Each feature vector $\{w \mid w \in y_{st}\}$ is then subjected to a linear transformation $\mathcal{L}_{st}$ to scale to the model's dimensionality:

$$\mathbf{w}'' = \mathcal{L}_{st}(\mathbf{w})$$

3.4 Missing Data Imputation: Reconstruction

In addressing the perennial challenge of missing data within time series forecasting, particularly in financial markets, the model adopts an innovative approach centered on the use of "given" embeddings. This methodology enables the dynamic reconstruction of missing data points during model inference, integrating this capability directly into the forecasting process. The inclusion of a reconstruction loss component in the loss function further refines this imputation strategy, ensuring the model's proficiency not only in forecasting but also in accurately infilling missing data.

3.4.1 Given Embedding and Data Reconstruction

The "given" embedding operates under the premise that certain data points within the input may be missing or unknown at the time of prediction. By assigning a unique embedding to denote the presence or absence of data, the model can differentiate between observed and missing values. This distinction allows it to apply specialized attention to the reconstruction of these missing points, leveraging the context provided by adjacent and related data points within the time series.

During model inference, portions of the input are intentionally masked to simulate the missing data scenario. The model is then tasked with reconstructing these masked points based on the surrounding context and its learned understanding of the data's underlying patterns. This process not only aids in mitigating the impact of missing data on forecast accuracy but also enhances the model's robustness and adaptability to incomplete datasets.

3.4.2: Reconstruction Loss

The reconstruction loss plays a pivotal role in optimizing the model's capacity for data imputation. Evaluated by comparing the model's reconstructed values against the actual (masked) data points, this loss metric guides the model toward more accurate and plausible reconstructions. The loss function utilized for this purpose mirrors that of the forecast loss, maintaining consistency in the evaluation criteria applied across different aspects of the model's output.

Given the probabilistic nature of the forecasting model, which employs a skew-normal distribution to capture the asymmetry and potential bias in financial predictions, the reconstruction loss is similarly approximated using a skew-normal distribution. This congruence ensures that the model remains aligned with the probabilistic framework underlying its forecasting methodology, providing a coherent and unified approach to handling uncertainty, both in predictions and in data imputation.

3.4.3 Implementation and Impact

Incorporating the reconstruction loss into the overall loss function, with a weighted amount reflecting its relative importance $\mathbf{recon}_{imp}$, allows for a balanced optimization process. This

process ensures that the model's performance in forecasting EPS Surprise is not compromised by its additional focus on data reconstruction. The dual objective of achieving high forecast accuracy while effectively imputing missing data elevates the model's utility and applicability, particularly in financial contexts where data completeness and reliability are often concerns.

4 Experimental Setup

4.1 Data Preparation

The foundation of this research relies heavily on the comprehensive assembly and refinement of data, specifically tailored to company financials within the US stock market. This endeavor necessitated a methodical approach to data collection, ensuring the incorporation of pertinent financial indicators and consensus estimates, despite the inherent challenges associated with data completeness and reliability. The finalized predictors and substitutions are shown in Figure 2.

4.1.1 Dataset Construction

The dataset was meticulously constructed with a primary focus on companies listed in the S&P 500, leveraging the index's representation of the broader US market. This selection criterion addressed the initial hurdle of data volume management, ensuring a focused yet representative sample of the market. The dataset amalgamates a variety of financial indicators, including company fundamentals, industry classifications, consensus estimates, and market performance metrics, aligning with the predictor variables proposed by Liu et al. (2022) substituting indicators where necessary to fit the US context. Most tickers employed the specific quarterly periods of 2013Q3 up to 2023Q3 with just under 20 tracing back to 2010Q4.

Company Fundamentals and Industry Classifications: The core of the dataset comprises company fundamentals and market performance indicators. Given the high reliance on accurate and comprehensive data, there was an initial engagement with dataJockey.io for company fundamentals. Despite its utility, issues with data completeness and consistency were encountered, prompting a shift towards combining data from Zacks.com and dataJockey.io for a more robust dataset. This amalgamation offered a deeper historical horizon and improved consistency crucial for the modeling needs.

Industry Classifications: For industry classifications, the Bloomberg Industry Classification System (BICS) was substituted with the Global Industry Classification Standard (GICS), utilizing the detailed Wikipedia listing of S&P 500 companies. This choice ensured a comparable depth of industry insight.

Consensus Estimates: The Seeking Alpha platform served as the primary source for consensus estimates, offering a quarterly breakdown of analysts' expectations. This choice was informed by the platform's comprehensive coverage and the granularity of the data, which aligned with the model's requirements. However, it's noteworthy that while Seeking Alpha provides a rich dataset, it inherently carries biases observed in crowdsourced investment research. Despite the potential for reduced bias compared to traditional sell-side research, the reliance on such estimates introduces a layer of uncertainty, particularly given the varying motivations and analytical rigor behind each contribution (Jame et al., 2016). Alongside these were some missing variables which were mitigated through alternative measures like revenue estimates.

Market Performance Metrics: For market performance data, the SHASHR index was substituted with the S&P 500 60-day trailing return, calculated from publicly available pricing data from the Wall Street Journal, thereby maintaining relevance to a focus on the US stock market.

4.1.2 Challenges and Limitations

The construction of the dataset was not without its challenges. Despite rigorous data collection

efforts, notable gaps and inconsistencies are present, primarily attributable to variations in the 10-Q reporting formats across companies and the limitations of the data sources. Furthermore, the significant presence of missing or erroneous data, particularly in SEC filings that could not be effectively parsed, necessitated the implementation of a reconstruction loss within the model. This approach aims to account for and mitigate the impact of these data gaps on forecasting accuracy. Despite the rich dataset, significant performance determinants like industry characteristics, company specifics, and leadership traits, which often rely on qualitative or sentimental insights, remain unaccounted for. This limitation is crucial, considering such factors collectively drive over 60% of company performance and are not fully represented in quantitative data alone.

4.1.3 Dataset Metrics

- The average quarter length per ticker was approximately 39.7, indicating a substantial timeframe for both prediction and context separation, based on the 18-quarter context benchmark.

- The mean deviation between the highest and lowest consensus EPS estimate was observed to be 0.2862 or approximately 28.62%, indicating a considerable variance in EPS estimates, reflective of the inherent uncertainty in financial forecasts.

- There were 19967 observations with a context length of 18 quarters and a prediction length of 8 quarters yielding around 9600 viable samples.

4.1.4 Seeking Alpha as a Data Source: Pros and Cons

Seeking Alpha, renowned for its crowdsourced investment research, presents a unique blend of opportunities and challenges as a data source for consensus estimates. While professional forecasts often exhibit a conservative bias to ensure they are easily surpassed, crowdsourcing platforms like Seeking Alpha might mitigate this through a broader base of contributors with varied motivations. However, the platform's crowdsourced nature may also introduce a layer of bias, albeit potentially less pronounced than that of traditional sell-side forecasts. While Seeking Alpha offers a more democratized perspective on investment analysis, the diversity in analytical depth and motivation among contributors can vary widely, affecting the consistency and reliability of the data. 4.1.5 Train/Validation/Test Split

To maintain chronological integrity, the dataset was divided into three distinct segments: training, validation, and testing, with 0.15 of the data allocated to both the testing and validation splits, respectively.

This division was meticulously executed to ensure that for each ticker within the dataset, the samples designated for testing were extracted from the latter portions of the available time series. Following this, samples for validation were selected from the subsequent earlier sequence, while the remaining earlier portions of the dataset were allocated for training purposes.

Such an arrangement guarantees that the data segments used for both validation and testing purposes are chronologically positioned in the future relative to the data on which the model was initially trained. This setup is crucial for simulating a realistic forecasting scenario where the model must predict future values based on historical data, thereby enhancing the model's practical applicability and ensuring its predictive performance is evaluated under conditions that closely mimic real-world financial forecasting challenges.

Available | Unavailable/Not Applicable but used Alternative | Unavailable with no Alternative

Input Factors	Applied Alternative
Revenue*	
Gross Profit*	
Operating Income*	
Pretax Income*	
Net Income*	
Gross Margin*	
Net Profit Margin*	
Book Value per share*	
EBITDA*	
Cash Flow from Operations*	
Cash Flow from Investing Activities*	
Net Debt*	
Cash Flow from Financing Activities*	
Estimate Comparable EPS Adjusted*	
Cash Flow per share	Free Cash Flow*
BICS Level 1 Classification	GICS Sector
BICS Level 2 Classification	GICS Industry Group
BICS Level 3 Classification	GICS Industry
BICS Level 4 Classification	GICS Sub-Industry
Dividend per share	Dividend*
Current Ratio	Debt–to-Equity Ratio*
SHASHR index 60-day trailing return	S&P 500 index 60-day trailing return*
Bloomberg Consensus estimate for adjusted EPS	Seeking-Alpha Normalized Consensus EPS Estimate*
Standard deviation of consensus EPS	Seeking-Alpha number of Consensus EPS estimates*
Median of Consensus Estimates for Adjusted EPS	Seeking-Alpha Mean Normalized Consensus EPS Estimate*
Analysts' low EPS estimate	Seeking-Alpha Low Normalized Consensus EPS Estimate*
Analysts' high EPS estimate	Seeking-Alpha High Normalized Consensus EPS Estimate*
Consensus estimate for EBITDA	Seeking-Alpha Consensus Funds-From-Operations (FFO) Estimate*
Consensus estimate for net income	
Consensus estimate for gross margin	Seeking-Alpha Consensus Revenue Estimate*
Percentage change of Consensus Estimates for Adjusted EPS over prior estimate	
Change of Consensus Estimates for Adjusted EPS in the four weeks prior to present date	

Figure 2: Data predictors following data collation and substitutions. Predictors with (*) were given imputation markers within the dataset as detailed in the conclusion.

4.2 Loss Function Composition for Forecasting Model

The loss function of this EPS Surprise forecasting model is meticulously designed to encapsulate the multifaceted nature of financial time series forecasting. It integrates three primary components: forecast loss, reconstruction loss, and classification loss. This composite structure ensures that the model not only predicts future values with accuracy but also effectively imputes missing data and correctly categorizes the type of financial indicators being forecasted.

4.2.1 Forecast Loss

The forecast loss quantifies the accuracy of the model's predictions against actual observed values. This component is critical for assessing the model's performance in predicting future financial indicators, such as EPS Surprise. The lower the forecast loss, the closer the model's predictions are to the actual outcomes, indicating a higher forecasting accuracy.

4.2.2 Reconstruction Loss

Introduced earlier in the paper, the reconstruction loss addresses the model's ability to handle missing data within the input series. It measures the discrepancy between the model's

reconstructions of intentionally masked portions of the input data and their true values. The inclusion of this loss component encourages the model to develop a nuanced understanding of the underlying patterns in the data, facilitating the accurate imputation of missing values.

4.2.3 Classification Loss

Given the normalization of data columns in the dataset, the classification loss component gains importance. It evaluates the model's proficiency in correctly identifying the category to which a specific prediction belongs, such as distinguishing between revenue and income. This capability is crucial, especially considering that the input features are normalized to standardize their scale across different indicators.

Normalization poses a challenge for the model to recognize the specific category of a prediction without relying on the magnitude difference between variables like revenue and income, which would be apparent in non-normalized data. In a normalized dataset, predictions are adjusted relative to the mean of their respective indicators, thus obfuscating direct comparisons based on magnitude alone. The classification loss, therefore, serves as a measure of the model's confidence in its predictions, indicating how effectively it can attribute normalized predictions to the correct financial indicators. The model's ability to excel in this aspect is indicative of its comprehensive understanding of the data's structure and dynamics. This not only enhances the accuracy of predictions but also ensures that the forecasts are meaningful and correctly interpreted in the context of financial analysis.

4.2.4 Composite Loss Function

The composite loss function is articulated as:

$$\text{Total Loss} = \text{Forecast Loss} + (\text{Reconstruction Loss} \times \text{Reconstruction Importance}) + (\text{Classification Loss} \times \text{Classification Importance})$$

Here, "Reconstruction Importance" and "Classification Importance" are weighting factors that adjust the influence of the reconstruction and classification losses, respectively, on the total loss. This allows for a balanced optimization process, ensuring that the model equally prioritizes forecasting accuracy, data imputation, and the correct classification of financial indicators.

By integrating these three components, the loss function not only drives the model towards higher forecasting precision but also ensures robust handling of missing data and accurate categorization of predictions. This comprehensive approach underpins the model's ability to deliver actionable insights for financial forecasting, particularly in predicting EPS Surprise with a nuanced understanding of the underlying financial data.

5 Experiments

5.1 Preliminary Experiment

5.1.1 DeepAR vs. Temporal-only Transformer

This experimental design meticulously evaluates the predictive prowess of a temporal-only Transformer model against a widely acknowledged autoregressive model, DeepAR. This preliminary test is crucial for validating the Transformer's utility as a competent predictor in financial forecasting, particularly in the realm of EPS surprise predictions.

Hyperparameter Configuration
The model configuration adopted in this study is tailored to balance computational efficiency and robust model performance, while also safeguarding against overfitting. Key hyperparameters for both models include but are not limited to:

- Batch size per epoch: Set at 32, optimizing for space-time efficiency, ensuring a manageable computational load.

- Number of epochs: Capped at 20 to prevent overfitting while allowing sufficient training depth.

Transformer Model	DeepAR Model	Both Models
Number of Activation Heads: 8	Number of Recursive Neural Network (RNN) Layers: 2	Prediction Length: 8 quarters
Applied Activation Function: softrelu	RNN Cells per Layer: 8	Context length: 18 quarters
	Cell type: LSTM[1]	Batch size per epoch: 32
	Dropout Cell type: Zoneout Cell	Dropout rate: 0.2
		Number of epochs: 20
		Embedding dimension for categoricals: 8
		Number of samples for probabilistic forecasting: 100
		Point Estimate Central Tendency: Median
		Distribution Output: Student-T
		Loss: Negative Log-Likelihood

Table 1: Table showing hyperparameter configuration for both the DeepAR model and Temporal-only Transformer.

- Prediction and Context length: The context length is set to 18 quarters in line with Liu et al. (2022) and a forecasting length of 8 quarters, reflecting a strategic choice to focus on near-to-mid-term forecasting relevancy.

- Embedding dimension for categoricals: Fixed at 8, to sufficiently capture the nuances of categorical variables without overwhelming the model with excessive dimensionality.

- Distribution Output: A Student-T distribution was applied to allow for higher degrees of uncertainty through the inclusion of the degrees of freedom.

These choices underline a strategic emphasis on model efficiency and performance, demonstrating a thoughtful approach to handling the extensive computations typical of Transformer models and the intricate dynamics of financial data. The rest of the applied hyperparameters are presented in Table 1. Comparatively, the DeepAR model yielded a slightly higher MSE of 3.1648, indicating a slightly less accurate performance compared to the Transformer models. However, it is essential to interpret these results in practical terms, considering the specific context of financial forecasting. The MSE quantifies the average squared difference between the predicted and actual EPS surprise values. In this context, a lower MSE suggests that the model's predictions are closer to the actual values, signifying higher predictive accuracy.

[1] LSTM - Long short-term memory network

Figure 3: Top - EPS surprise predictions for Google using Amazon's base DeepAR model; Bottom - EPS surprise predictions for Netflix using Amazon's base DeepAR model. Both plots show the 95% confidence interval for the predictions (green band) as well as the median of the predictions supplied as a point estimate (green trendline) to be compared against the actual surprise values (blue trendline). These predictions are for the last 8 quarters before 2023Q3 applying information only up to 2021Q2.

Figure 4: Top - EPS surprise predictions for Google using the vanilla Transformer model; Bottom - EPS surprise predictions for Netflix using the vanilla Transformer model. Both models were run under the model dimensionality of 32 and feed-forward dimensionality of 4. Both plots show the 95% confidence interval for the predictions (green band) as well as the median of the predictions supplied as a point estimate (green trendline) to be compared against the actual surprise values (blue trendline). These predictions are for the last 8 quarters before 2023Q3 applying information only up to 2021Q2.

Table 2

	Vanilla Transformer		
Model Dimension ⟶ Feed-forward Dimension	32 — 4	128 — 16	DeepAR
MSE	2.7590	**2.7587**	3.1648
MASE	**1.0440**	1.1770	1.1051
MAPE (%)	**10.8493**	14.2848	16.5374
sMAPE (%)	**1.0541**	1.0954	1.0885
MSIS	20.1714	**19.5291**	19.9387
RMSE	1.6610	**1.6609**	1.7789
NRMSE	7.4396	**7.4391**	7.9678

Further examining the quantitative results, it's evident that while both models provide reasonably accurate predictions, the Transformer models exhibit superior performance across multiple metrics. Its lower MSE, MASE, MAPE, sMAPE, and MSIS values across both model setups indicate better predictive accuracy and interval calibration compared to DeepAR. However, while these provide insight into overall predictive accuracy, it is essential to consider additional factors such as the width of prediction intervals and the model's ability to capture uncertainty. In the case of probabilistic forecasting, the width of prediction intervals reflects the model's uncertainty about future values. Although the confidence interval bands may appear wider for some predictions like for Netflix versus Google, comparing them based on the standard deviation of the prior actual EPS surprise values for both companies provides a more meaningful assessment of uncertainty. A wider confidence interval indicates higher uncertainty in the predictions, potentially stemming from volatile market conditions or limited historical data availability.

Upon analyzing the predictions for the last 8 quarters ending in 2023 Q3 for Google and Netflix, it was observed that the 95% prediction intervals generated by both models encompassed the actual EPS surprise values. However, the median of the prediction intervals, used as the point estimate, did not effectively track the actual values, indicating potential inaccuracies in point predictions. This suggests that while the models captured the overall variability in EPS surprise, their point estimates may lack precision.

In figures 3a and 3b, representing the predictions for Google and Netflix using DeepAR, respectively, the predicted EPS surprise values are presented along with the 95% prediction intervals. Despite the intervals containing the actual values, the median point estimates deviate from the true values, indicating a discrepancy between predicted and actual EPS surprise.

Conversely, in figures 4a and 4b, illustrating the predictions for Google and Netflix using the Transformer model, respectively, a similar trend is observed. While the prediction intervals encompass the actual values, the median point estimates exhibit deviations from the true values, suggesting potential limitations in point prediction accuracy.

Practically, these results imply that while both models can capture the variability in EPS surprise and provide probabilistic forecasts, their point estimates may not always align with the actual values. Decision-makers should consider the uncertainty inherent in the predictions and interpret them in conjunction with other relevant information when making strategic decisions.

It is important to note that the quantitative performance comparisons between the Transformer model and DeepAR may be influenced by the specific training setup employed in this study. DeepAR may rely on a more long-term setup, which could lead to better performance given a different training setup. Therefore, these comparisons should be interpreted cautiously and may vary based on the chosen training configuration and dataset characteristics.

5.2 Main Experiments

5.2.1 Normal Distribution vs. Skew-Normal Distribution Output

This main experiment focuses on comparing the performance of a temporal-only Transformer model utilizing either a skew-normal distribution output or a normal distribution output. This comparison aims to ascertain the impact of accounting for asymmetry in the distribution of EPS surprise predictions. The following hyperparameter choices underscore a tailored approach to optimize model performance for this specific test.

Hyperparameter Configuration

- Batch Size and Epochs: A batch size of 16 and a total of 15 epochs are chosen, striking a balance between computational efficiency and adequate model training without overfitting. However, the inclusion of checkpoint monitoring and early stopping with patience of 3 epochs made for a model that terminated after 9 epochs.

- Model Architecture: The model is configured with a dimensionality of 96, ensuring sufficient capacity to capture complex patterns in the data. The architecture includes 8 activation heads, 2 encoder and decoder layers, and a feed-forward dimension of 800, optimizing the model's ability to process and learn from the temporal data.

- Dropout Rates: Dropout is strategically applied with a rate of 0.2 for embedding dropout and feed-forward dropout, minimizing the risk of overfitting by regularizing the model training process. Attention-related dropouts are set to 0, maintaining focus on the model's attention mechanisms.
- Positional Embedding: The "time2vec" positional embedding type is selected to incorporate rich temporal information into the model, enhancing its understanding of time-related patterns (Grigsby et al., 2021).

- Attention Mechanisms: Given the baseline temporal-only embedding method, self-attention and cross-attention mechanisms were configured to use the Performer attention mechanism (Choromanski et al., 2020), optimizing computational efficiency while maintaining the model's focus on relevant temporal relationships.

- Learning Rate and Optimization: An initial learning rate of 1e-8, a base learning rate of 2e-4, and a warmup step count of 20 are set to ensure a gradual adjustment toward optimal learning. The decay factor is 0.9, facilitating a controlled reduction in the learning rate to fine-tune model performance (Penguin, 2021).

- Embedding Dimensions: Time and categorical embedding dimensions are both set to 8, ensuring that temporal and categorical features are succinctly represented within the model.

- Loss Function Configuration: The model utilizes a negative log-likelihood (`nll`) loss function, with weighted components for classification (importance=0.05) and reconstruction (inportance=0.1) losses, reflecting the emphasis on both predictive accuracy and data imputation capabilities while avoiding placing undue effect on predicting the past.

- Additional Configurations: The model leverages several advanced settings, including the use of "performer" kernels for efficient attention computation, seasonal decomposition to capture cyclical patterns, and a "gelu" activation function to introduce non-linearity into the model's computations.
This experimental setup, with its careful selection of hyperparameters, is designed to rigorously evaluate the efficacy of incorporating skew-normal distribution outputs in financial forecasting models.

Results

Table 3 presents the performance metrics for the Temporal-only Transformer with a Skew-Normal distribution output versus a Normal distribution output. The normal distribution model outperforms the skew-normal model in nearly all evaluated metrics except for the Mean Absolute Percentage Error (MAPE). Specifically, the normal model demonstrates a

lower forecast loss, overall loss, Mean Absolute Error (MAE), Mean Squared Error (MSE), reconstruction loss, Root Squared Error (RSE), and symmetric Mean Absolute Percentage Error (sMAPE). The skew-normal model, however, records a significantly lower MAPE compared to the normal model, an anomaly when juxtaposed with its performance across other metrics.

- Forecast Loss and Overall Loss: The normal model exhibits a lower forecast loss (58.20) and overall loss (1456.84) compared to the skew-normal model, which reports a forecast loss of 70.51 and an overall loss of 4191.74. This indicates that the normal model more accurately predicts EPS surprises, reflecting a more efficient handling of forecast errors.

- Mean Absolute Error (MAE): With MAE values of 3.43 for the normal model and 3.67 for the skew-normal model, the normal model demonstrates slightly better accuracy in its predictions, suggesting a closer alignment with the actual data points.

- Mean Absolute Percentage Error (MAPE): Interestingly, the skew-normal model records a lower MAPE (83049.97) compared to the extraordinarily high MAPE of the normal model (348374.88), suggesting a divergence in performance in this specific metric. This anomaly could indicate instances where the skew-normal model's predictions are proportionally closer to actual values in certain cases, despite its overall poorer performance.

- Mean Squared Error (MSE) and Reconstruction Loss: The MSE for the normal model (3.79) is lower than that of the skew-normal model (4.39), reinforcing the former's superior predictive accuracy. Additionally, the reconstruction loss is significantly higher in the skew-normal model (41212.35) than in the normal model (13986.36), indicating challenges in accurately reconstructing missing data.

- Root Squared Error (RSE) and Symmetric MAPE (sMAPE): Both RSE and sMAPE are lower for the normal model, signifying its better consistency and accuracy in prediction across various test samples.

The overall superior performance of the normal distribution model across most metrics, except for MAPE, suggests a more robust and consistent predictive capability. The lower model performance could be attributed to the increased degrees of freedom and its effect on a higher tendency of the skew-normal model to overfit the training data. This could be deduced from its poorer performance on the test data, as overfitting typically results in models that perform well on training data but fail to generalize to unseen data effectively.

An examination of the prediction plots for three distinct companies (A, AAPL, and ABBV) shown in Figure 5, reveals notable differences in the performance of models utilizing skew-normal versus normal distribution outputs. A key observation from these plots is the very similar confidence interval associated with both distribution outputs. This characteristic suggests that, despite earlier metrics indicating the superior performance of the normal distribution model for point estimates, the skew-normal distribution model’s performance relative to the normal distribution at modeling probabilistic uncertainty is effectively equal.

Figure 5: Prediction plots for temporal-only Transformer with varying distribution outputs. Left - Normal Distribution Output; Right - Skew-Normal Distribution Output. The confidence intervals were obtained by applying a quantile formula over 200 samples from the distribution output. The point estimates apply the distribution mean.

Both models' confidence intervals successfully encapsulate the true EPS surprise values, although their point estimates do not closely track the actual outcomes. This phenomenon underscores a potential misalignment between the models' probabilistic forecasting and their precision in point estimation. However, the substantial width of these intervals, in some instances extending to EPS Surprise values of ±4 (or 400%), highlights a significant analytical challenge. Such wide CIs suggest an underlying uncertainty that could be deemed excessively large, potentially rendering the forecasts less actionable due to the implied analyst error magnitude.

Furthermore, the skew-normal distribution model's shape parameter reveals a tendency towards right-skewness in its predictions. However, the proximity of these values to zero and the inclusion of zero within the 95% CI suggest a larger tendency towards a

predominantly symmetric distribution of EPS surprise within this modeling context. This observed symmetry might reflect the mitigating influence of Seeking Alpha's consensus estimate platform, suggesting its capacity to reduce bias through its inherent diversity of inputs.

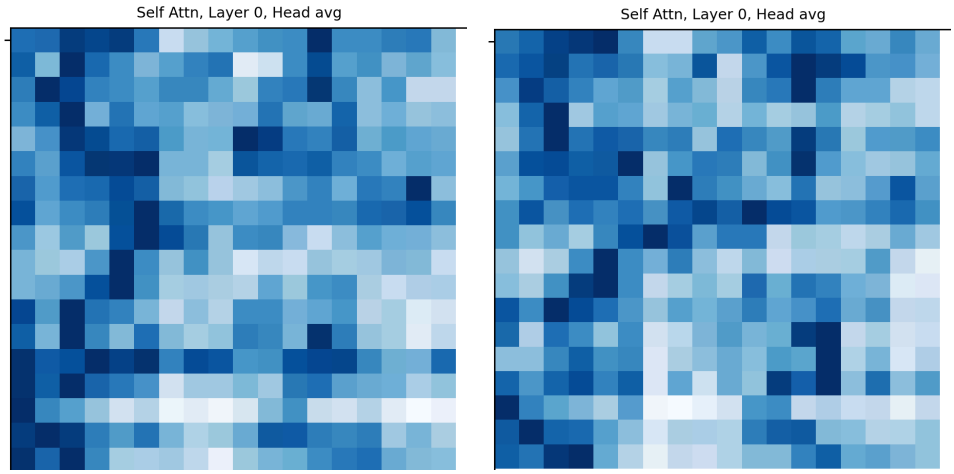

Figure 6: Encoder self-attention matrices for temporal-only Transformer with varying distribution outputs. Left - Normal Distribution Output; Right - Skew-Normal Distribution Output. Both matrices assume a context of 18 quarters hence the matrix is of dimension 18 x 18. Both matrices present the average attention across all 8 attention heads used within the encoder layer. The sequence order in each matrix is presented from top-to-bottom and left-to-right.

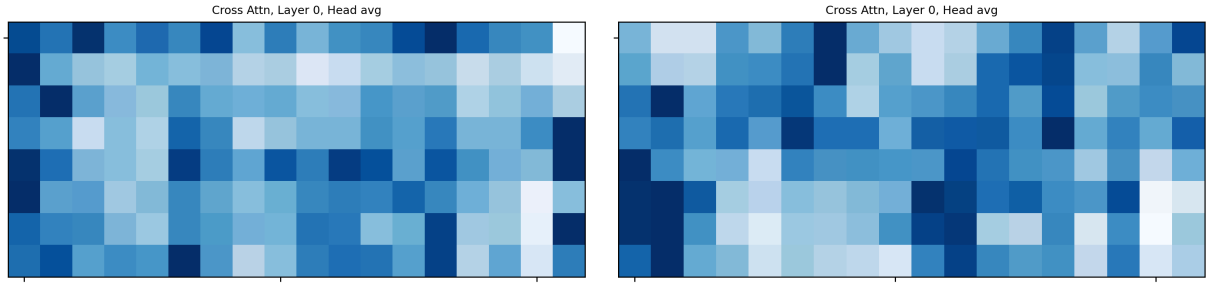

Figure 7: Decoder cross-attention matrices for temporal-only Transformer with varying distribution outputs. Left - Normal Distribution Output; Right - Skew-Normal Distribution Output. Both matrices assume a context of 18 quarters and a forecast of 8 quarters hence the matrix is of dimension 18 x 8. Both matrices present the average attention across all 8 attention heads used within the decoder layer. The sequence order in each matrix is presented from top-to-bottom and left-to-right.

A comparative analysis of the average head encoder self-attention (Figure 5) and cross-attention matrices (Figure 6) across both models yields another intriguing insight: the choice of distribution output exerts minimal influence on the identification of relevant points within the sequence. This similarity in attention patterns, irrespective of the distribution output, highlights that the inherent relationship between sequential data points remains consistent across different modeling approaches. Such consistency, ensured through controlled randomness via seed values, affirms that the structural dynamics captured by the model are inherently stable and unaffected by the nature of the distribution output employed.

In conclusion, while initial model metrics suggested a leaning towards the normal distribution model for point estimate performance, the probabilistic forecasting strength of the skew-normal distribution model, alongside its implications for understanding market dynamics and analyst behavior, offers

valuable insights. The narrower confidence intervals of the skew-normal model, coupled with the consistent attention patterns observed across both models, emphasize the nuanced benefits of considering distributional asymmetry in financial forecasting.

Contrast with Preliminary Tests
When contrasting the obtained performance metrics against those obtained from the preliminary tests, a notable decrease in performance is observed. This discrepancy can be attributed to the different test setups employed. The preliminary test utilized the entire dataset for testing, including samples previously exposed during training, inherently inflating the model's perceived performance by testing it on familiar data. In contrast, the main experiment employed a chronologically separated dataset for training, validation, and testing splits, ensuring the test set comprised only future, unseen data relative to the training set.

Table 3
(*) indicates variables that are limited to that configuration

Feature Representation ⟶ Distribution Output	Temporal — Normal	Temporal — Skew-Normal	Spatiotemporal — Skew-Normal
Loss	2.7590	4191.743	**206.48**
Forecast Loss	58.204	70.510	**11.748**
Reconstruction Loss	13986.362	41212.352	**1946.684**
MAE	**3.435**	3.672	7.296
MSE	**3.794**	4.392	10.3
sMAPE (%)	**28.451**	32.995	28.574
RSE	**28.929**	32.687	78.565
MAPE (%)	348374.875	**83049.969**	532562.688
Classification Loss			1.271*

5.2.2 Spatio-Temporal vs. Temporal Feature Representation

This main experiment focuses on comparing the performance of a spatio-temporal feature representation in the Transformer model against a temporal-only feature representation. The focus of this experiment centers on understanding how the inclusion of spatial variables, alongside temporal data, impacts model performance across various metrics. The experiment applies the same setup as the Skew-Normal versus Normal experiments with the only change being the embedding method on the Skew-Normal model. As such, these results are presented against the results obtained from the Skew-Normal model as the Skew-Normal applied a temporal-only feature representation.

Results
The comparison between the spatio-temporal and temporal feature representations unveils insightful distinctions in model performance across a range of metrics. The spatiotemporal

representation notably excels over the temporal-only representation in several aspects, although it does encounter setbacks in point estimate metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), and Root Squared Error (RSE) as evidenced in Table 3. The superior performance of the spatiotemporal representation in the symmetric Mean Absolute Percentage Error (sMAPE) metric underscores the complexities and, at times, the unreliability of point estimates within this experimental framework.
The spatiotemporal representation's performance is notably mixed, excelling in certain areas while facing challenges in others, particularly in point estimate metrics like MAE, MAPE, MSE, and RSE. The model's inferior performance in these metrics, being almost double that of the temporal-only representation, accentuates the inherent challenges in relying solely on point estimates for evaluating forecasting models. Conversely, the dramatic reduction in loss function values, by over a tenfold difference compared to the temporal-only model, indicates a profound improvement in the model's overall efficiency and accuracy in capturing the underlying patterns and nuances of financial data.

This dichotomy between point estimate performance and loss function efficiency suggests that the negative log-likelihood loss function, pivotal in optimizing the model's scale and shape parameters, plays a crucial role in enhancing the model's focus on achieving a higher probability of capturing true values. However, this optimization appears to come at the expense of the model's mean value accuracy, potentially leading to more erratic point estimates that deviate significantly from actual values.

An examination of the prediction plots for three distinct companies (A, AAL, and AAPL) shown in Figure 8, reveals more notable differences in the performance of the spatiotemporal representation against the temporal-only representation. The narrower 95% confidence intervals observed in the spatio-temporal model's prediction plots underscore its superior probabilistic forecasting capabilities. This enhanced performance is emblematic of the model's adeptness at quantifying and conveying uncertainty, a critical aspect of financial forecasting. The precision of these probabilistic forecasts, as evidenced by the more constrained confidence intervals, suggests an improved capacity to capture the complex variability inherent in financial time series.

This divergence in performance also underscores the inherent tension, earlier identified in the Skew-Normal vs. Normal experiment, between achieving precise point estimates and effectively capturing the range of possible outcomes. The less-wide 95% CIs of the spatiotemporal representation, coupled with its less accurate point estimates, underscore the model's shift towards a probabilistic understanding over point accuracy. This shift indicates a deeper analytical capability, where the model prioritizes capturing the breadth of potential EPS Surprise outcomes over pinpointing a single expected value.

As evidenced in Figure 9, the spatiotemporal representation, by design, incorporates a more complex data representation, weaving in spatial variables alongside temporal ones. This complexity is reflected in the dimensionality of its attention matrices, which are larger due to the model's attempt to account for the spatial relationships inherent in the data. Such an approach suggests a broader analytical scope, aiming to capture not just when events occur but also how they relate across different spatial dimensions—be it industry classifications, market dynamics, or financial fundamentals. However, despite this increased complexity, the self-attention matrices do not exhibit a discernible pattern in prioritizing specific temporal points over others. This indicates a sophisticated analytical process where the model dynamically focuses on various indicators across time, beyond mere chronological proximity.

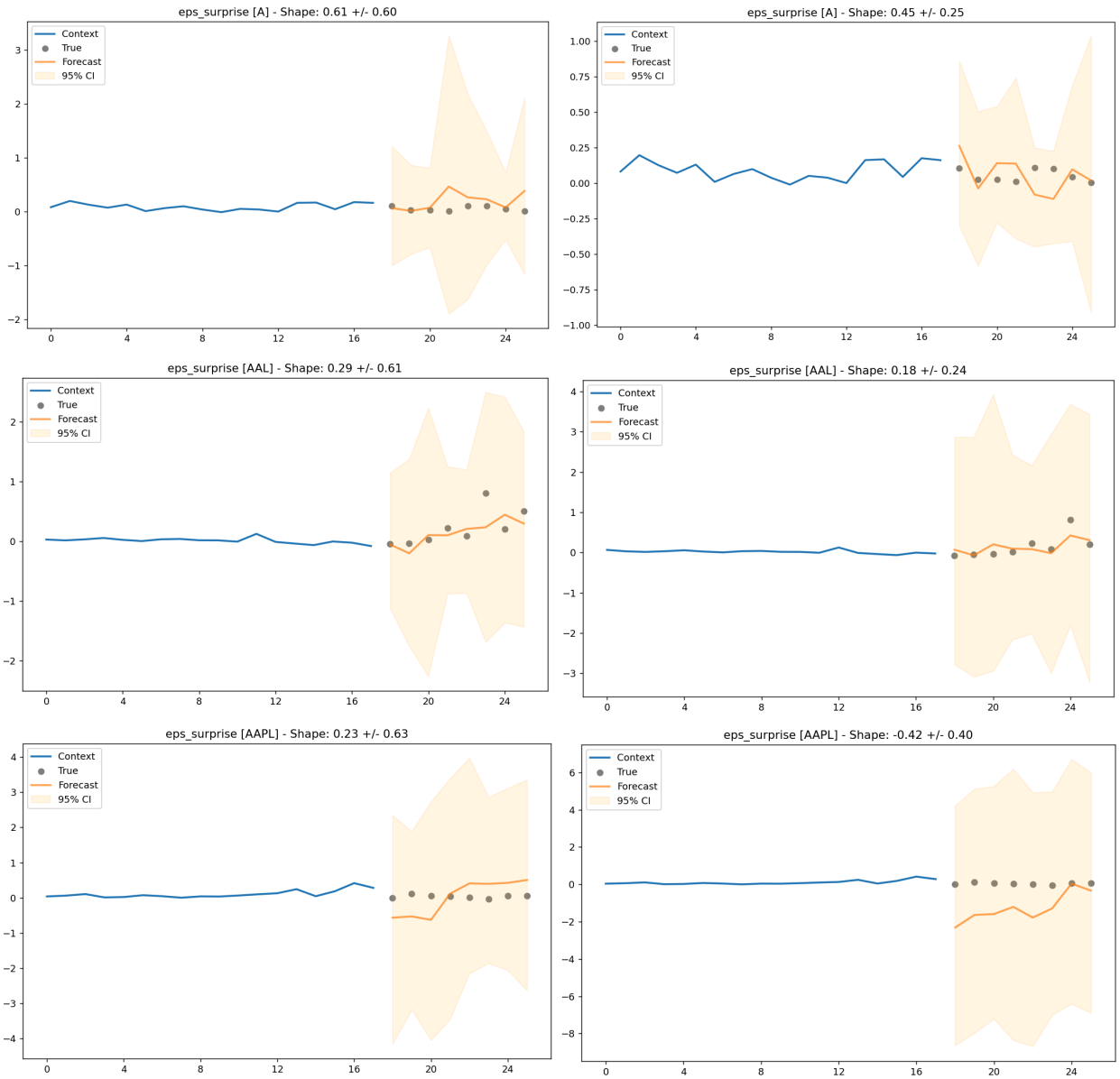

Figure 8: Prediction plots for Skew-Normal Distribution Transformer with varying feature representations. Left - Temporal-only Feature Representation; Right - Spatiotemporal Feature Representation. The confidence intervals were obtained by applying a quantile formula over 200 samples from the distribution output. The point estimates apply the distribution mean.

Interestingly, the minimal attention assigned to the right and bottom borders of the self-attention matrix, indicative of industry and sub-industry classifications, suggests that these categorical variables play a lesser role in the model's historical considerations. This observation implies that, while spatial variables are considered, their impact on the forecasting outcome is moderated, with the model finding limited predictive utility in the static classifications compared to the rich temporal dynamics captured in the data.

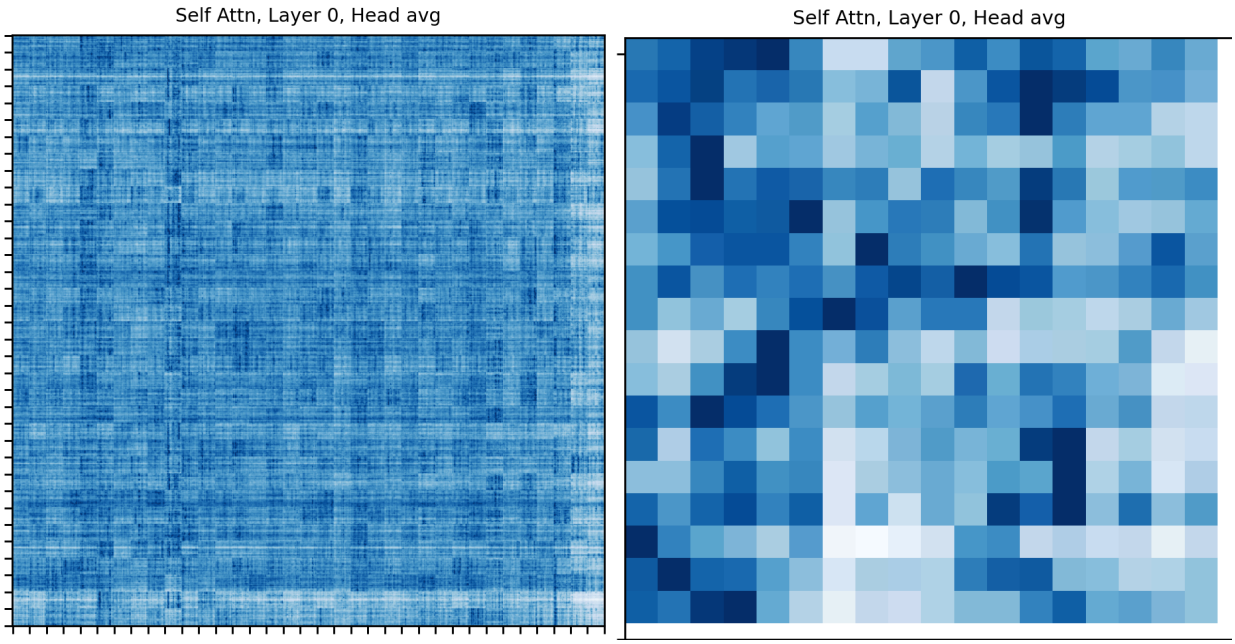

Figure 9: Encoder self-attention matrices for Skew-Normal Distribution Transformer with varying feature representations. Left - Spatiotemporal Feature Representation; Right - Temporal-only Feature Representation. Both matrices assume a context of 18 quarters but the spatial representation increases the dimensionality of the left matrix to (18 x 35) x (18 x 35) for 35 spatial variables while the right matrix is of dimension 18 x 18. Both matrices present the average attention across all 8 attention heads used within the encoder layer. The sequence order in each matrix is presented from top-to-bottom and left-to-right.

The cross-attention matrices, shown in Figure 10, further illustrate this point, with a visible demarcation indicating a diminished weight across the prediction sequence attributed to the industry and sub-industry classifications. Such a pattern suggests that in making EPS Surprise predictions, the model assigns greater importance to the temporal and spatial interactions within the data rather than relying heavily on static categorical representations like these. There is also the possibility that the model learns the static nature of these categorical variables and thus learns to attribute less weight to them past the earlier quarters of the prediction sequence. This strategic de-emphasis of these variables highlights the model's focus on extracting predictive signals from the intricate web of temporal relationships and spatial correlations.

Interestingly, despite the incorporation of spatial variables in the spatiotemporal model, the attention mechanism results indicate that the core relational dynamics—how different quarters relate to each other in the forecasting context—remain consistent across both model types. This consistency, ensured through controlled randomness via seed values, suggests that the addition of spatial variables does not fundamentally alter the temporal relationships that the model deems important. Instead, it enriches the model's analytical palette, allowing for a more nuanced interpretation of the financial time series data.

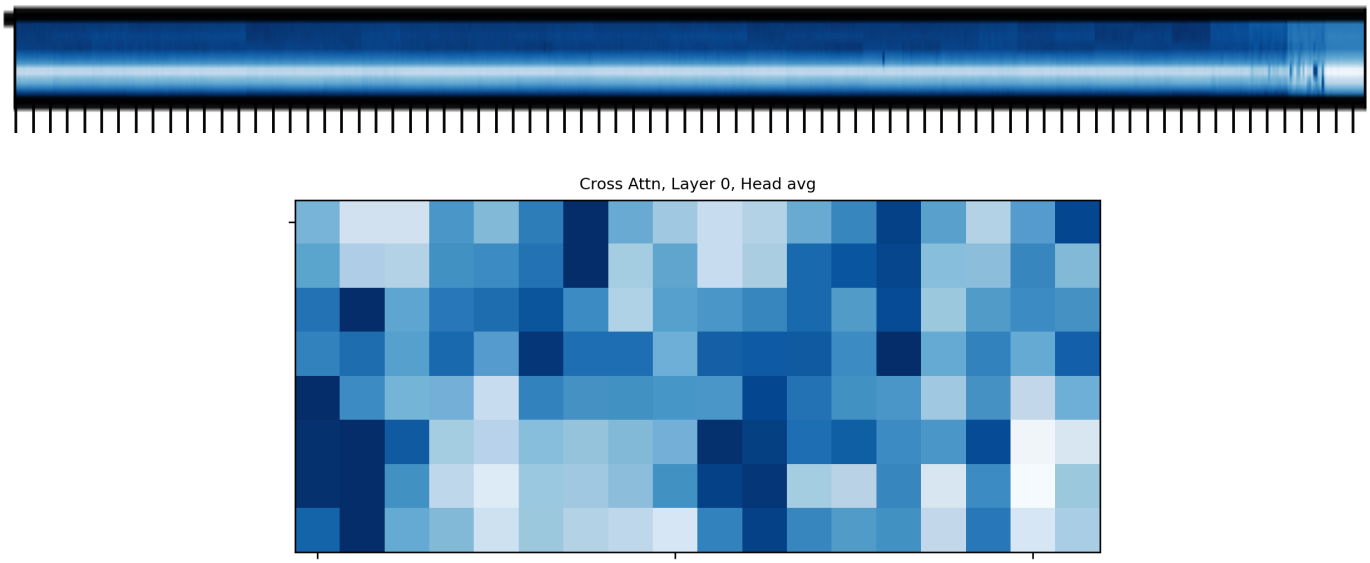

Figure 10: Decoder cross-attention matrices for Skew-Normal Distribution Transformer with varying feature representations. Top - Spatiotemporal Feature Representation; Bottom - Temporal-only Feature Representation. Both matrices assume a context of 18 quarters and a forecast of 8 quarters but the spatial representation increases the dimensionality of the top matrix to (18 x 35) x 8 for 35 spatial variables while the bottom matrix is of dimension 18 x 8. Both matrices present the average attention across all 8 attention heads used within the decoder layer. The sequence order in each matrix is presented from top-to-bottom and left-to-right.

6 Conclusion

In summarizing the insights gleaned from the extensive analysis conducted throughout this study, it becomes apparent that while the developed models hold significant potential, their current implementation yields confidence intervals that are overly broad and, consequently, not particularly informative for practical financial forecasting applications. This limitation underscores the inherent challenge in accurately predicting EPS surprises—a challenge exacerbated by the complexity of financial markets and the myriad factors influencing company performance.

The wide confidence intervals observed across both spatiotemporal and temporal models suggest an underlying difficulty in tightly constraining predictions within a narrow, highly probable range. This outcome highlights a crucial area for further refinement: the need for a more comprehensive synthesis of market dynamics indicators. As discussed, the incorporation of additional, potentially influential factors—such as industry characteristics, company specifics, leadership qualities, and broader economic indicators—could significantly enhance the models' predictive accuracy and the informativeness of their probabilistic forecasts. Addressing this need will require not only the development of more sophisticated data scraping and processing techniques but also an expansion of the model architecture to adeptly integrate and analyze a broader spectrum of data inputs.

Furthermore, the methodologies developed for this study, encompassing both financial data scraping techniques and the model architecture itself, offer versatile frameworks that can be adapted to other analytical contexts where asymmetry and bias are of primary concern. The potential applicability of these approaches extends beyond the specific realm of EPS surprise forecasting, suggesting utility in various domains where nuanced understanding and prediction of asymmetric outcomes are critical.

Despite the advancements presented in this research, the current performance of the models leaves us inconclusive regarding the nature and

extent of analyst bias within the scope of EPS prediction. This inconclusiveness, while reflective of the challenges inherent in modeling complex financial phenomena, also underscores the vast potential for future research in this area. Subsequent studies aimed at refining these models could benefit from deeper exploration into the factors driving analyst forecasts and their biases, leveraging advanced machine learning techniques to capture and analyze the subtle dynamics at play.

In conclusion, while the models developed in this study represent a meaningful step forward in the quest to enhance financial forecasting methodologies, their current limitations highlight the necessity for ongoing research and development. Future efforts should focus on expanding the datasets to include a more comprehensive array of market dynamics indicators, refining model architectures to more effectively capture and analyze these complexities, and exploring new applications where these innovative approaches can provide valuable insights. Through these endeavors, one may move closer to unlocking the full potential of machine learning in financial analysis, achieving more accurate, informative, and actionable forecasts.[2]

7 References

Abarbanell, J. S., & Bernard, V. L. (1992). Tests of analysts' overreaction/underreaction to earnings information as an explanation for anomalous stock price behavior. The Journal of Finance, 47(3), 1181. https://doi.org/10.2307/2328982

Alford, A. W., & Berger, P. G. (1999). A simultaneous equations analysis of forecast accuracy, analyst following, and trading volume. Journal of Accounting, Auditing & Finance, 14(3), 219–240. https://doi.org/10.1177/0148558x9901400303

Ball, R. T., & Ghysels, E. (2018). Automated earnings forecasts: Beat analysts or combine and conquer? Management Science, 64(10), 4936–4952. https://doi.org/10.1287/mnsc.2017.2864

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., Belanger, D., Colwell, L., & Weller, A. (2020, September 30). Rethinking attention with performers. arXiv.Org. https://arxiv.org/abs/2009.14794

De Bondt, W. F. M., & Thaler, R. H. (2002). Do analysts overreact? In Heuristics and Biases (pp. 678–685). Cambridge University Press. http://dx.doi.org/10.1017/cbo9780511808098.040

Dreman, D. N., & Berry, M. A. (1995). Analyst forecasting errors and their implications for security analysis. Financial Analysts Journal, 51(3), 30–41. https://doi.org/10.2469/faj.v51.n3.1903

Easterwood, J. C., & Nutt, S. R. (1999). Inefficiency in analysts' earnings forecasts: Systematic misreaction or systematic optimism? The Journal of Finance, 54(5), 1777–1797. https://doi.org/10.1111/0022-1082.00166

Francis, J., & Philbrick, D. (1993). Analysts' decisions as products of a multi-task environment. Journal of Accounting Research, 31(2), 216. https://doi.org/10.2307/2491271

Fried, D., & Givoly, D. (1982). Financial analysts' forecasts of earnings. Journal of Accounting and Economics, 4(2), 85–107. https://doi.org/10.1016/0165-4101(82)90015-5

Grigsby, J., Wang, Z., Nguyen, N., & Qi, Y. (2021, September 24). Long-Range transformers for dynamic spatiotemporal forecasting. arXiv.Org. https://arxiv.org/abs/2109.12218

Groysberg, B., Healy, P., Nohria, N., & Serafeim, G. (2011). What factors drive analyst forecasts? Financial Analysts Journal, 67(4), 18–29. https://doi.org/10.2469/faj.v67.n4.4

Harris, R. D. F., & Wang, P. (2019). Model-based earnings forecasts vs. financial analysts' earnings

[2] **AI Statement:** Artificial Intelligence was applied to modify the initial draft of the paper for better grammatical structure and adherence to research paper guidelines. These guidelines included but were not limited to the removal of first-person references and the use of abbreviations without an earlier definition. Lastly, it was applied to analyze the paper for logical inconsistencies or claims that were not backed with source citations but were critical to the analysis alongside a readership process of about four(4) reviewers.

forecasts. The British Accounting Review, 51(4), 424–437. https://doi.org/10.1016/j.bar.2018.10.002

Hribar, P., & McInnis, J. (2012). Investor sentiment and analysts' earnings forecast errors. Management Science, 58(2), 293–307. https://doi.org/10.1287/mnsc.1110.1356

Jame, R., Johnston, R., Markov, S., & Wolfe, M. C. (2016). The value of crowdsourced earnings forecasts. Journal of Accounting Research, 54(4), 1077–1110. https://doi.org/10.1111/1475-679x.12121

Johnson, T. E., & Schmitt, T. G. (1974). Effectiveness of earnings per share forecasts. Financial Management, 3(2), 64. https://doi.org/10.2307/3665292

Kazemi, S. M., Goel, R., Eghbali, S., Ramanan, J., Sahota, J., Thakur, S., Wu, S., Smyth, C., Poupart, P., & Brubaker, M. (2019, July 11). Time2Vec: Learning a vector representation of time. arXiv.Org. https://arxiv.org/abs/1907.05321

Krolikowski, M. W., Chen, G., & Mohr, J. E. (2016). Optimism pattern of all-star analysts. International Review of Financial Analysis, 47, 222–228. https://doi.org/10.1016/j.irfa.2016.08.003

Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

Lim, T. (2001). Rationality and analysts' forecast bias. The Journal of Finance, 56(1), 369–385. https://doi.org/10.1111/0022-1082.00329

Liu, Q., Ouyang, L., & Xu, G. (2022). Prediction of Earning Surprise using Deep Learning Technique.

Penguin. (2021, May 29). Setting a minimum learning rate on "Reduce On Plateau." Stack Overflow. https://stackoverflow.com/questions/67746083/setting-a-minimum-learning-rate-on-reduce-on-plateau

Sougiannis, T., & Yaekura, T. (2001). The accuracy and bias of equity values inferred from analysts' earnings forecasts. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.253033

Tse, S. Y., & Yan, H. (2008). Analysts' incentives and systematic forecast bias. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1107770

Wang, S., Wei, C., & Han, L. (2017). Do business relationships affect the accuracy of analyst earnings forecasts? Evidence from China. Asia-Pacific Journal of Financial Studies, 46(1), 155–177. https://doi.org/10.1111/ajfs.12164

Wasserman, N., Nohria, N., & Anand, B. N. (2001). When does leadership matter? The contingent opportunities view of CEO leadership. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.278652

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2023, August). Transformers in time series: A survey. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. http://dx.doi.org/10.24963/ijcai.2023/759

Wu, haixu, Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. NeurIPS.

Wu, Y., Liu, T., Han, L., & Yin, L. (2018). Optimistic bias of analysts' earnings forecasts: Does investor sentiment matter in China? Pacific-Basin Finance Journal, 49, 147–163. https://doi.org/10.1016/j.pacfin.2018.04.010

Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., & Eickhoff, C. (2021, August 14). A transformer-based framework for multivariate time series representation learning. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. http://dx.doi.org/10.1145/3447548.3467401

Zhang, Y., & Yan, J. (2023). Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. ICLR.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 35(12), 11106–11115. https://doi.org/10.1609/aaai.v35i12.17325

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022, January 30). FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. arXiv.Org. [https://arxiv.org/abs/2201.