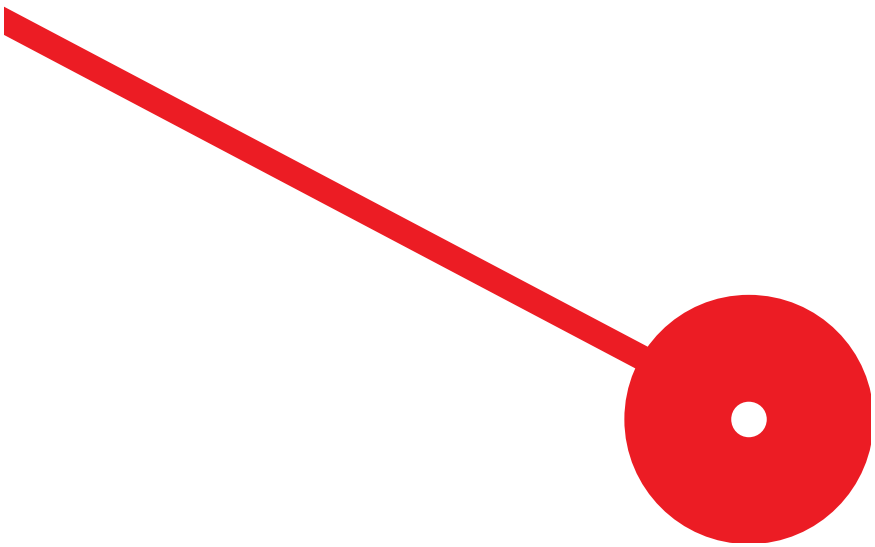
INSTITUTO
SUPERIOR
DE CONTABILIDADE
E ADMINISTRAÇÃO
DO PORTO
POLITÉCNICO
DO PORTO

M MESTRADO
BUSINESS INTELLIGENCE AND ANALYTICS

Transfer Learning Approaches for Business Forecasting

Fábio Miguel Reis de Azevedo

10/2024



Fábio Miguel Reis de Azevedo. Transfer Learning Approaches for Business Forecasting
10/2024

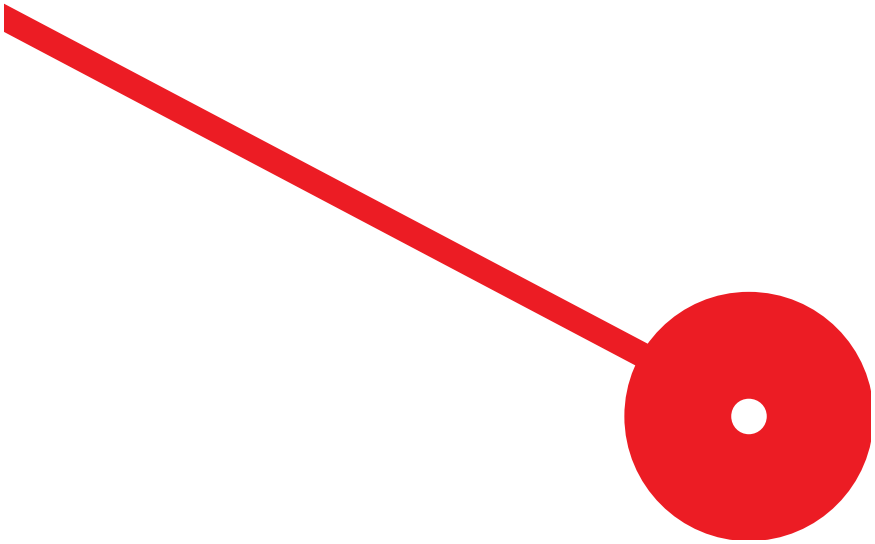
INSTITUTO
SUPERIOR
DE CONTABILIDADE
E ADMINISTRAÇÃO
DO PORTO
POLITÉCNICO
DO PORTO

M MESTRADO
BUSINESS INTELLIGENCE AND ANALYTICS

Transfer Learning Approaches for Business Forecasting

Fábio Miguel Reis de Azevedo

Dissertação de Mestrado apresentada ao Instituto Superior de Contabilidade e Administração do Porto para a obtenção do grau de Mestre em Business Intelligence and Analytics, sob orientação da Doutora Patrícia Alexandra Gregório Ramos e do Doutor José Manuel Soares Oliveira.



Fábio Miguel Reis de Azevedo. Transfer Learning Approaches for Business Forecasting
10/2024

Resumo:

Esta dissertação examina a forma como a aprendizagem por transferência pode melhorar a previsão do negócio de retalho, melhorando a precisão e a eficiência dos modelos de previsão da procura, projeção de vendas e otimização de inventário. Aborda dois desafios principais na previsão de retalho: escassez de dados e generalização. A aprendizagem por transferência utiliza modelos pré-treinados de áreas relacionadas para construir modelos de previsão mais robustos, integrando conhecimentos de diversos conjuntos de dados para criar soluções adaptáveis a vários ambientes de retalho. O estudo compara cinco modelos de aprendizagem por transferência: Chronos da AWS, TimeGPT da Nixtla, Lag-Llama da Invenia Technical Computing, TimesFM da Google Research e Uni2TS da SalesforceAI, avaliados utilizando o conjunto de dados M5, que integra dados reais de vendas a retalho da Walmart. Este conjunto de dados apresenta desafios devido às suas estruturas hierárquicas, dependências temporais e factores externos como promoções e flutuações de preços. O desempenho de cada modelo é avaliado utilizando o erro médio absoluto escalado (MASE) e a pontuação de probabilidade classificada contínua (CRPS). O MASE mede a exatidão das previsões pontuais, permitindo comparações entre séries cronológicas com sazonalidade variável, enquanto o CRPS avalia a exatidão das previsões probabilísticas, crucial para compreender a incerteza na procura no retalho. Os modelos são testados em dois cenários: Zero-Shot, em que os modelos são aplicados sem treino no conjunto de dados M5, e Fine-Tuning, em que os modelos são especificamente ajustados ao mesmo. Os resultados mostram que o TimesFM supera consistentemente os outros modelos em ambos os cenários, demonstrando robustez e versatilidade. O Lag-Llama tem um bom desempenho nas definições Zero-Shot, captando eficazmente os padrões gerais das séries temporais. Os modelos Chronos melhoram com o tamanho, mas permanecem menos competitivos, enquanto o Uni2TS enfrenta desafios no Fine-Tuning, indicando um potencial sobreajuste.

Palavras-chave: *Transfer Learning*, Previsão no Negócio, Processamento de Linguagem Natural, Arquitetura *Transformer*, Redes Neurais Profundas, Modelos baseados no Mecanismo de Atenção.

Abstract:

This dissertation examines how transfer learning can enhance retail business forecasting, improving the accuracy and efficiency of models for demand forecasting, sales projection, and inventory optimization. It addresses two main challenges in retail forecasting: data scarcity and generalization. Transfer learning leverages pre-trained models from related fields to build more robust forecasting models, integrating insights from diverse datasets to create adaptable solutions for various retail environments. The study compares five transfer learning models: Chronos by AWS, TimeGPT by Nixtla, Lag-Llama by Invenia Technical Computing, TimesFM by Google Research, and Uni2TS by SalesforceAI, evaluated using the M5 dataset, which contains real-world retail sales data from Walmart. This dataset poses challenges due to its hierarchical structures, temporal dependencies, and external factors like promotions and price fluctuations. Model performance is assessed using Mean Absolute Scaled Error (MASE) and Continuous Ranked Probability Score (CRPS). MASE measures point forecast accuracy, allowing comparisons across time series with varying seasonality, while CRPS evaluates probabilistic forecast accuracy, crucial for understanding uncertainty in retail demand. The models are tested in two settings: Zero-Shot, where they are applied without training on the M5 dataset, and Fine-Tuning, where they are specifically adjusted to it. Results show that TimesFM consistently outperforms the other models in both settings, demonstrating robustness and versatility. Lag-Llama performs well in Zero-Shot settings, capturing general time series patterns effectively. Chronos models improve with size but remain less competitive, while Uni2TS faces challenges in Fine-Tuning, indicating potential overfitting.

Keywords: Transfer Learning, Business Forecasting, Natural Language Processing, Transformer Architecture, Deep Neural Networks, Attention-based Models.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A10G – NVIDIA A10G GPU

AI – Artificial Intelligence

AR – Autoregressive

ARIMA – AutoRegressive Integrated Moving Average

CIFT – Crowd-Informed Fine-Tuning

CNN – Convolutional Neural Network

CRPS – Continuous Ranked Probability Score

DeepAR – Deep Autoregressive Networks

DPO – Direct Preference Optimization

ETS – Exponential Smoothing State Space

Freq-Mask – Frequency Mask Augmentation

Freq-Mix – Frequency Mix Augmentation

GPT – Generative Pretrained Transformer

GPT-3 – Generative Pretrained Transformer 3

GPU – Graphics Processing Unit

ICLR – International Conference on Learning Representations

LLaMA – Large Language Model Meta AI

LoRA – Low-Rank Adaptation

LOTSA – Large-scale Open Time Series Archive

LLMs – Large Language Models

LSTM – Long Short-Term Memory

MAE – Mean Absolute Error

MASE – Mean Absolute Scaled Error

ML – Machine Learning

MLM – Masked Language Modeling

MLP – Multi-Layer Perceptron

MOIRAI – Model not explicitly defined

MSSE – Mean Squared Scaled Error

NeurIPS – Neural Information Processing Systems

N-BEATS – Neural Basis Expansion Analysis for Time Series Forecasting

NHITS – Neural Hierarchical Interpolation for Time Series Forecasting

PFN – Probabilistic Forecasting Network

ReLU – Rectified Linear Unit

RLHF – Reinforcement Learning from Human Feedback

RoPE – Rotary Positional Embeddings

RMSNorm – Root Mean Square Layer Normalization

RNN – Recurrent Neural Network

rMAE – Relative Mean Absolute Error

rRMSE – Relative Root Mean Square Error

Seq2Seq – Sequence to Sequence Model

SGD – Stochastic Gradient Descent

SKU – Stock-Keeping Unit

t-distribution – Student's t-distribution

VAR – Vector Autoregression

WQL – Weighted Quantile Loss

1 Introduction

The primary aim of this research is to explore how transfer learning can improve retail business forecasting. By utilizing pre-trained models from related fields, we hope to boost the accuracy and efficiency of predictions specific to retail, thus enabling better decision-making processes. Transfer learning offers a way to tackle the common problem of data scarcity in retail time series forecasting. Our goal is to develop innovative techniques that use transfer learning to make forecasting models more robust, even when there's a lack of labeled data.

In this study, we compare different transfer learning models, including Chronos by AWS, TimeGPT by the Nixtla Team, Lag-Llama by Invenia Technical Computing, TimesFM by Google Research, and Uni2TS by SalesforceAI. We evaluate how these models perform in enhancing prediction accuracy for various retail forecasting tasks, such as demand forecasting, sales projections, and inventory optimization. By applying the latest transfer learning techniques, we aim to improve the precision and reliability of these predictions, helping retailers optimize their resources and streamline operations, addressing Goal 1 (G1) of conducting a comprehensive comparison of these models to assess their relative performance and generalization capabilities.

Another important goal of the work is to enhance the generalization ability of forecasting models in the retail sector. By incorporating insights from diverse datasets through transfer learning, we aim to create models that can identify patterns and trends across different retail environments. This directly ties to Goal 2 (G2), which focuses on assessing the adaptability and scalability of the selected models when applied to complex retail datasets, such as the M5 dataset.

Ultimately, our overarching objective is to advance forecasting practices in the retail industry. By exploring how transfer learning can address data scarcity and improve model generalization, we hope to provide new insights and methods that significantly improve inventory management and supply chain optimization. This aligns with Goal 3 (G3), which is to address retail-specific challenges, such as variability in sales patterns, and to evaluate how robustly these models generalize across diverse product categories and store environments. Through this research, we aim to demonstrate the practical benefits of applying transfer learning to business forecasting, contributing to the ongoing evolution of the retail landscape.

The outline of the thesis begins with an introductory chapter that establishes the research focus on exploring how transfer learning can enhance retail business forecasting. This section emphasizes the potential of transfer learning to tackle issues related to data scarcity and improve the generalization capabilities of forecasting models.

Following the introduction, a literature review examines the evolution of transfer learning alongside advancements in natural language processing. This review highlights the significant role of transfer learning in recommendation systems, the transformative effects of attention-based models like Transformers, and the growing integration of artificial intelligence technologies across various sectors.

The methodology chapter presents the Transformer architecture, detailing its essential components, including the encoder-decoder structure, multi-head attention, feed-forward networks, and positional encoding. This section clarifies how these elements interact to process sequential data both efficiently and accurately. Next, the thesis introduces five foundation models specifically designed for point and probabilistic forecasting: Chronos, Lag-Llama, TimeGPT, MOIRAI (Uni2TS), and TimesFM. Each model is discussed comprehensively, addressing its unique features, architecture, training processes, and advantages in managing retail time series data. Additionally, the chapter provides an overview of benchmark models used for comparison, which include ARIMA, ETS, Seasonal Naïve, and Naïve models.

An empirical study chapter follows, presenting findings from the application of these forecasting models to the M5 dataset. This section characterizes the M5 dataset, emphasizing its hierarchical structure, temporal dependencies, and external factors that enhance its realism as a benchmark for retail sales forecasting. It also explains the concept and importance of fine-tuning in adapting pre-trained models to the M5 dataset, thereby improving their capacity to capture specific retail patterns. Performance metrics for model evaluation are introduced, focusing on the Mean Absolute Scaled Error (MASE) for point forecast accuracy and the Continuous Ranked Probability Score (CRPS) for assessing probabilistic forecast accuracy. The chapter thoroughly discusses the results, analyzing the performance of each model under both Zero-Shot and Fine-Tuning scenarios, highlighting the strengths and weaknesses of each approach. Notably, TimesFM consistently stands out as the top performer, showcasing high accuracy, generalization capability, and efficiency, while Lag-Llama and TimeGPT exhibit potential for particular forecasting needs. Areas for improvement in Chronos and Uni2TS are also identified.

The concluding chapter summarizes the key findings of the research, underscoring the practical implications of the models for enhancing retail decision-making, optimizing inventory management, and improving supply chain operations. It acknowledges the limitations of the study and suggests directions for future research, including the exploration of additional models and methodologies, scaling to more complex retail networks, and testing for real-world implementation. The document concludes with a comprehensive list of references cited throughout the thesis.

2 Literature Review

In the evolving landscape of computer science, recent years have marked a significant leap forward, particularly in transfer learning and natural language processing (NLP). These advancements redefine our approach to machine learning and lay the groundwork for a future where artificial intelligence seamlessly integrates into various aspects of human life. At the heart of this transformation lies the principle of transfer learning, a method that has proven to be a game-changer in enhancing Click-Through Rate (CTR) prediction in recommendation systems, thereby personalizing user experiences and optimizing online platform efficiency (Pan & Yang, 2010).

Transfer learning leverages knowledge acquired in one domain to improve performance in another, a crucial strategy for overcoming the "domain shift" challenge where the data distribution in the source domain differs significantly from that in the target domain. As highlighted by (He et al. 2014), this approach enables recommendation systems to effectively utilize accumulated knowledge from related domains, fine-tuning pre-trained models and developing specific neural architectures to bridge the gap between disparate data distributions. The surge in interest around transfer learning methodologies, explored by (Zhang et al. 2023) and (Covington et al. 2016), underscores their potential in crafting more insightful and user-centric recommendation engines.

A pivotal advancement in transfer learning has been the adoption of Deep Neural Networks (DNNs), which can learn high-dimensional latent representations of users and items, facilitating a more nuanced transfer of knowledge between domains (Rendle et al., 2010). Moreover, the integration of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) has significantly advanced our ability to capture sequential and temporal patterns in user interactions, which is particularly relevant for CTR prediction (Tan et al., 2019).

However, the journey of transfer learning from theory to practice in recommendation systems is just one facet of a broader narrative. In the wider domain of AI, attention-based models have heralded a new era in NLP and time series modeling. The introduction of the Transformer architecture by (Vaswani et al. 2017) marked a significant departure from traditional RNNs and convolutions. By relying solely on attention mechanisms, the Transformer achieves unparalleled computational efficiency and performance in NLP tasks, showcasing an exceptional ability to capture long-range dependencies without the

need for sequential processing. This breakthrough has set new benchmarks in machine translation tasks and demonstrated the versatile applicability of the Transformer across various NLP endeavors.

The last decade has witnessed significant progress in various areas of artificial intelligence (AI) and machine learning (ML), largely driven by the advent of deep learning (LeCun, Bengio, & Hinton, 2015; Goodfellow, Bengio, & Courville, 2016). These advancements span from the understanding and generation of natural language (Devlin, Chang, Lee, & Toutanova, 2019; Brown et al., 2020) to image recognition (Krizhevsky, Sutskever, & Hinton, 2012) and beyond. Deep learning has revolutionized how machines learn, providing the ability to automatically extract relevant features from raw data, an approach known as representation learning (Bengio, Courville, & Vincent, 2013). Deep Neural Networks (DNNs) have been successfully applied to various tasks, including speech recognition and machine translation (Hinton, Osindero, & Teh, 2006; Sutskever, Vinyals, & Le, 2014), demonstrating the flexibility and power of deep learning.

Building on the foundational work of Vaswani et al. (2017), subsequent research has expanded attention-based models beyond NLP. Studies like "A Time Series is Worth 64 Words" and "Unified Training of Universal Time Series Forecasting Transformers" delve into the application of time series as compact representations, advocating for unified training approaches for time series forecasting models. These studies highlight the potential for generalization and efficiency enhancements across multiple domains, underscoring the transformative impact of attention-based models in understanding and processing data sequences.

The implications of these advancements are profound. In recommendation systems, integrating transfer learning techniques has ushered in a new era of personalized user experiences, where systems are not only responsive but also anticipatory, adapting to user preferences with remarkable accuracy. Similarly, the advent of the Transformer architecture and its successors has pushed the boundaries of NLP, enabling machines to understand, interpret, and generate human language with previously unimaginable efficacy.

The synergy between transfer learning and attention-based models is driving the evolution of AI technologies. The potential for these methodologies to be applied across

a broader spectrum of AI applications—from healthcare diagnostics to autonomous systems—is immense. With ongoing research and development, we are on the cusp of unlocking even greater capabilities, making AI more adaptable, efficient, and, ultimately, more human-centric.

This convergence between sophisticated transfer learning methodologies and the revolutionary Transformer model in AI and machine learning sets the stage for an unprecedented era of innovation. The synergy of these technologies is not merely academic curiosity but a leap towards realizing AI systems that can learn more efficiently, adapt more dynamically, and perform more complex tasks with precision and understanding previously unattainable.

The transformative impact of these advancements extends beyond theory to address practical challenges and real-world applications. In recommendation systems, the nuanced understanding and application of transfer learning have led to marked improvements in personalization and efficiency. Systems are now better equipped to understand user preferences, even with limited data, by drawing parallels from related domains, thereby enhancing user experience and engagement. This capability is critical in an era where digital platforms vie for user attention, and the ability to deliver personalized content swiftly and accurately can significantly influence platform loyalty and user satisfaction.

Similarly, in NLP, the introduction of the Transformer architecture has catalyzed a reevaluation of how machines understand language. The model's ability to process words in relation to all other words in a sentence, without the constraints of sequence, allows for a deeper understanding of context and nuance. This leap forward has broad implications, from improving machine translation systems to creating more responsive and conversational AI agents. The Transformer's influence extends to areas such as sentiment analysis, where its capacity to grasp language subtleties can lead to more accurate interpretations of human emotions and opinions, opening new pathways for AI applications in social media, customer service, and beyond.

Moreover, applying attention-based models in time series forecasting presents a novel approach to predicting future events based on past data. This capability is crucial across various sectors, including finance, where accurate market predictions can lead to more informed investment decisions, and meteorology, where it can contribute to more

accurate weather forecasts. The versatility of these models, as evidenced in studies like "A Time Series is Worth 64 Words" and "Unified Training of Universal Time Series Forecasting Transformers," showcases their potential to revolutionize predictive modeling across disciplines.

As we venture into the future, the ongoing evolution of transfer learning and attention-based models promises to create AI systems that are not only more intelligent but also more intuitive and responsive to human needs. The potential for these technologies to be customized and applied across myriad domains underscores the advent of a new chapter in AI, where machines can learn from a broader array of experiences and insights, mimicking human learning processes more closely than ever before.

In NLP, models like BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) have marked milestones, showing that models pre-trained on large text datasets can perform language tasks with few or no task-specific training samples. These models, based on the Transformer architecture (Vaswani et al., 2017), stand out for their ability to understand the context of a text in depth. Transfer learning, as explored by Caruana (1997) and Pan and Yang (2010), allows the knowledge gained in one task to be applied to another, different but related task. This approach has proven particularly effective in recommendation systems (Covington, Adams, & Sargin, 2016), where pre-trained models can be fine-tuned to offer personalized recommendations from limited data.

In the context of recommendation systems, the prediction of Click-Through Rate (CTR) has significantly evolved with the adoption of deep neural networks (He & McAuley, 2014; Tan, Xu, & Liu, 2019), allowing for a more nuanced understanding of user preferences and item-user interaction. Computer vision has also been transformed by deep learning, with convolutional networks (CNNs) setting new standards in tasks such as image classification (Krizhevsky et al., 2012) and medical image segmentation (Zhou et al., 2018). Machine translation and text generation have benefited from introducing attention-based models, which have improved the quality of translation by capturing contextual nuances (Luong, Pham, & Manning, 2015). Models like the Transformer (Vaswani et al., 2017) have demonstrated impressive capabilities in understanding and generating natural language coherently and relevantly.

AI and ML are at a turning point, with deep learning, NLP, and transfer learning leading a wave of innovations that transcend previous technology boundaries. These advances

offer promises not only in the academic field but also in practical applications ranging from the improvement of recommendation systems to significant advancements in healthcare (Wei, Wang, & Zhou, 2019), cybersecurity (Yang, Chen, Choi, & Lin, 2019), and much more. As we continue to explore these technologies, AI is becoming increasingly integrated into the fabric of daily life, opening new paths for innovation and discovery.

In conclusion, the swift progress within computer science, driven by advancements in transfer learning and NLP, signals a transformative shift in our engagement with AI and machine learning. The fusion of these technologies not only elevates our existing systems but also carves out fresh pathways for exploration and implementation, foreshadowing an era where the full spectrum of AI's capabilities becomes integral to every facet of our society. As we delve deeper into these innovations, we stand on the brink of unlocking an unprecedented breadth of possibilities, promising to redefine the landscape of AI applications and their impact on the world around us.

3 Methodology

3.1 Transformer Architecture

The Transformer architecture represents a groundbreaking approach for sequence transduction tasks, eliminating the need for recurrent or convolutional neural networks, which were previously dominant models in this domain (Vaswani et al., 2017; Bahdanau et al., 2014; Sutskever et al., 2014). Prior sequence transduction models relied heavily on these architectures, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks (Bahdanau et al., 2014). However, these models faced challenges, particularly their inability to parallelize, which significantly impacted their efficiency in handling long sequences (Sutskever et al., 2014). The introduction of the Transformer, based solely on attention mechanisms, overcomes these limitations by replacing recurrence and convolution with a model that relies on multi-head self-attention mechanisms to capture dependencies across sequences (He et al., 2016).

In the Transformer, the model follows an encoder-decoder structure. The encoder processes the input sequence into a continuous representation, while the decoder generates the output sequence from this representation. Both the encoder and decoder consist of multiple layers, each containing two key sub-layers: a multi-head self-attention mechanism and a fully connected feed-forward network (Vaswani et al., 2017). Each sub-layer in the encoder is connected by a residual connection, followed by layer normalization (He et al., 2016). This use of residual connections ensures the gradient flow and allows deeper networks to train more efficiently, while layer normalization helps stabilize the training process.

The decoder mirrors the encoder's structure, but with an additional sub-layer for encoder-decoder attention, which attends to the output of the encoder stack (Bahdanau et al., 2014). This allows the decoder to focus on relevant parts of the input sequence while generating the output sequence. The Transformer decoder is autoregressive, meaning it predicts one symbol at a time, conditioning future predictions on the previously generated symbols. To prevent the model from accessing future tokens during training, a masking mechanism is employed to ensure that each prediction depends only on previous outputs (Sutskever et al., 2014).

A critical aspect of the Transformer is the attention mechanism, which models dependencies between sequence elements irrespective of their position. The **scaled dot-product attention** mechanism calculates the relevance between different elements of the sequence using query, key, and value vectors (Bahdanau et al., 2014). These vectors allow the model to capture relationships between tokens at varying distances from each other, which is crucial for tasks such as machine translation. In traditional models like convolutional neural networks, the number of operations required to relate signals from distant positions grows linearly or logarithmically based on the distance between those positions (Gehring et al., 2017). In contrast, the Transformer reduces this complexity, allowing it to process all tokens in parallel, leading to much faster training times (He et al., 2016).

To further enhance the model's capability, the Transformer employs **multi-head attention**. This mechanism projects the queries, keys, and values into multiple subspaces, allowing the model to attend to different parts of the sequence simultaneously from various perspectives (Wu et al., 2016). Instead of performing a single attention function, multi-head attention applies the attention mechanism in parallel across multiple heads, each operating on a different subspace. This increases the model's ability to capture nuanced relationships between sequence elements and better represents information at multiple levels of abstraction (Bahdanau et al., 2014).

In addition to the attention mechanism, each layer in the encoder and decoder contains a fully connected feed-forward network that applies two linear transformations with a ReLU activation in between. This feed-forward network operates on each token independently, enhancing the overall ability of the model to process complex patterns (Sutskever et al., 2014). Despite the simplicity of these networks, they play a vital role in further processing the representations generated by the attention mechanism, improving the model's ability to learn intricate relationships in the data.

One challenge with the Transformer is the lack of inherent mechanisms to handle the sequential nature of input data, as it does not rely on recurrence or convolution. To address this, the Transformer incorporates **positional encodings** to provide the model with information about the relative positions of tokens in the sequence (Gehring et al., 2017). Positional encodings, based on sine and cosine functions, are added to the input embeddings to allow the model to learn relationships between tokens based on their positions. These encodings enable the model to account for the order of words, which is

critical in many natural language processing tasks (Press & Wolf, 2016). By providing this information, the Transformer can effectively capture the sequential structure of the data without relying on traditional recurrence or convolutional operations.

In conclusion, the Transformer architecture has revolutionized sequence transduction tasks by replacing recurrent and convolutional models with attention mechanisms. This shift has led to significant improvements in computational efficiency and parallelization, making it possible to capture long-range dependencies across sequences with greater speed and accuracy. The Transformer's combination of multi-head attention, feed-forward networks, and positional encoding has set a new standard in natural language processing, establishing its superiority in tasks like machine translation.

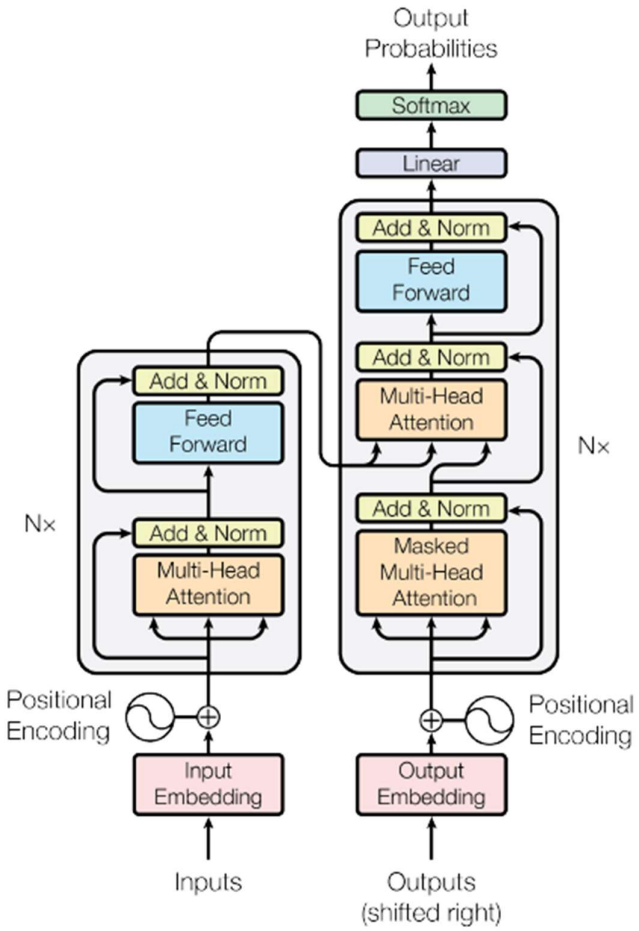


Figure 1 - The Transformer - model architecture. Reprinted from (Vaswani, A., et al., 2017).

3.2 Foundation Models for Point and Probabilistic Forecasting

The selection of the models Chronos, TimeGPT, Lag-Llama, TimesFM, and Uni2TS was carefully based on their relevance, innovation, and contribution to the field of transfer learning applied to time series. These models were chosen for representing different approaches and technological capabilities to address specific challenges in retail forecasting, including data scarcity, the need for generalization, and probabilistic forecasting. Each model offers a distinct solution to critical problems, enabling a comparative analysis that explores the advantages and limitations of modern techniques.

Chronos, for example, utilizes time series quantization and tokenization, facilitating the application of attention mechanisms to continuous temporal data. TimeGPT, with its ability to make predictions in Zero-Shot scenarios, stands out for its generalization capabilities, eliminating the need for extensive adjustments to new datasets. Lag-Llama integrates innovations such as Rotary Positional Encoding (RoPE) and attention mechanisms tailored for time series, focusing on multivariate scenarios and complex temporal interdependencies. TimesFM, with its architecture based exclusively on decoders, provides an efficient approach to long-term forecasting, reducing the need for frequent data segmentation. Finally, Uni2TS emphasizes multivariate and hierarchical adaptability, enabling the capture of temporal and contextual patterns across multiple levels of granularity.

Moreover, the choice of these models reflects their relevance in current scientific literature and their ability to address real-world challenges in the retail sector, such as temporal dependencies, seasonality, hierarchical relationships among products and locations, and the influence of external factors like prices and promotions. By incorporating modern architectures based on Transformers and attention mechanisms, these models allow for the exploration of efficiency and accuracy under diverse conditions, ranging from limited data scenarios to large-scale applications. Thus, the selection of these models ensures a comprehensive and robust evaluation of the most advanced techniques currently available in the domain of time series forecasting.

3.2.1 Chronos

The **Chronos** framework adapts existing language model architectures for probabilistic time series forecasting by leveraging techniques typically used in natural language

processing. Unlike natural language, where data consists of words or tokens from a finite vocabulary, time series data is real-valued and continuous. To bridge this gap, Chronos applies a series of adaptations designed to transform time series data into a format that language models can process.

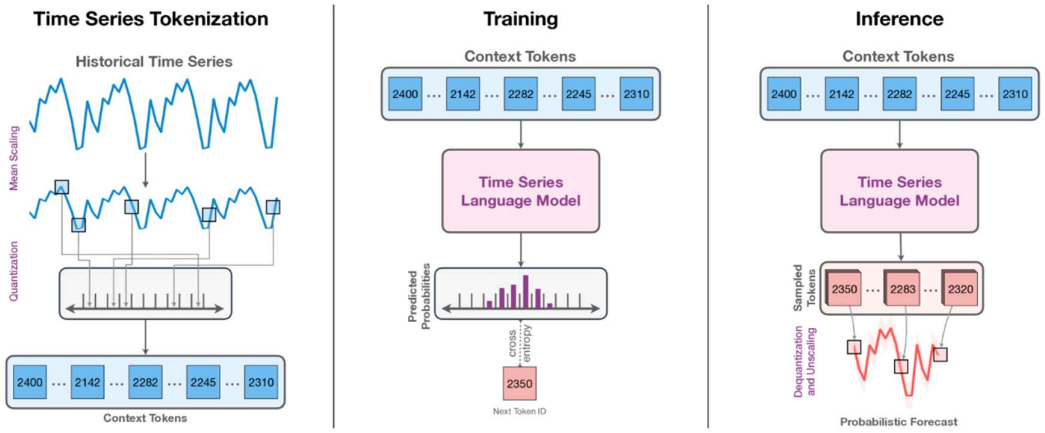


Figure 2 - High-level depiction of Chronos. Reprinted from (Ansari, A. F., et al., 2024).

The first crucial step is **scaling** the time series data. Time series often come from various domains (finance, climate, healthcare, etc.), and the values can vary greatly in scale, which presents challenges during optimization in deep learning models. To address this, Chronos employs **mean scaling**, where each value in the time series is normalized by the mean of its absolute values from the historical context (Salinas et al., 2020). This normalization helps bring all time series into a comparable range, which facilitates more stable and efficient training across diverse datasets.

After scaling, the continuous time series values are converted into discrete tokens through **quantization**. Quantization involves mapping these scaled values into a fixed number of discrete categories or **bins** (Rabanser et al., 2020). This process is essential because language models, like transformers, operate on discrete tokens from a finite vocabulary. By quantizing the time series, Chronos effectively tokenizes continuous data into a form that these models can process in the same way they handle language tokens.

Chronos adopts **uniform binning** for quantization, which involves dividing the data into evenly spaced intervals. This method is chosen for its simplicity and ability to generalize to new, unseen datasets. In some cases, other binning methods, such as quantile binning, could be employed, but uniform binning ensures that the model doesn't overfit to the specific distribution of a training set (Dooley et al., 2023). Once the real values are

converted into tokens, **dequantization** is used during inference to map the predicted tokens back into their original real-valued form, ensuring that the forecasts generated by the model are interpretable and useful.

The next key adaptation in Chronos is the use of a **categorical output distribution**. In typical time series models, forecasting might involve predicting continuous values directly. However, Chronos treats the prediction task as a **regression via classification** problem, where the model learns to predict which bin (or token) a future time point will fall into. This approach enables the use of a **cross-entropy loss function**, commonly employed in language modeling tasks (Torgo & Gama, 1997). This setup allows Chronos to seamlessly leverage existing language modeling infrastructure and libraries, making it easier to integrate with standard transformer models, like those used in natural language processing (Raffel et al., 2020). Importantly, this method allows Chronos to model more complex and flexible output distributions, including multimodal distributions, which can capture diverse patterns in time series data (Salinas et al., 2020).

During **forecasting**, Chronos is capable of generating **probabilistic forecasts**. Instead of providing a single point forecast, it generates multiple possible future trajectories by **sampling** from the predicted token distribution. These samples represent different possible futures, providing a distribution of outcomes rather than a single deterministic prediction (Raffel et al., 2020). This probabilistic nature of Chronos makes it particularly valuable in real-world applications, where uncertainty and variability are inherent to time series data, such as in financial markets or weather forecasting.

Chronos stands out by demonstrating **zero-shot forecasting** capabilities, meaning that it can generalize to new time series datasets that it has never seen during training. This is achieved through its minimalist approach to adapting language models for time series, avoiding heavy reliance on task-specific features or fine-tuning, which are common in other models (Dooley et al., 2023). By focusing on tokenization and scaling techniques, Chronos effectively transforms time series into a language-like structure that existing language models can understand and process.

In conclusion, the **Chronos** framework offers a powerful, efficient, and flexible approach to time series forecasting by leveraging the strengths of language models. Through its novel tokenization, quantization, and probabilistic forecasting mechanisms, it enables the

application of transformers to time series data with minimal modifications, positioning it as a strong contender for general-purpose time series modeling.

3.2.2 Lag-Llama

The Lag-Llama model, developed by Rasul et al. (2023), represents a significant step towards creating foundation models specifically designed for probabilistic time series forecasting. Unlike traditional models that often require domain-specific tuning or task-specific architectures, Lag-Llama is built to generalize across a wide variety of time series datasets. This makes it particularly effective in zero-shot scenarios, where the model is expected to perform on unseen datasets without additional fine-tuning. The key innovation lies in how the model leverages lag features to extract temporal patterns from time series data, thus enabling it to predict future values based on historical trends and patterns. These lag features are constructed from a set of lag indices corresponding to different time frequencies—such as quarterly, monthly, weekly, daily, hourly, and even second-level intervals (Salinas et al., 2020). By capturing the periodicity and temporal dependencies inherent in time series data, the lag features allow the model to learn and generalize effectively.

In terms of architecture, Lag-Llama draws inspiration from the LLaMA transformer model (Touvron et al., 2023), but it introduces several key modifications to tailor the model specifically for time series data. One of the main modifications is the use of RMSNorm (Zhang & Sennrich, 2019) for pre-normalization, which improves the stability of training. Additionally, the model employs Rotary Positional Encoding (RoPE) (Su et al., 2021), which encodes positional information within the attention mechanism, helping the model to better understand the sequential nature of time series. This combination of attention mechanisms and positional encoding enables Lag-Llama to process and learn from long sequences of time series data more efficiently, making it well-suited for handling datasets with varying temporal resolutions.

A unique aspect of the Lag-Llama architecture is its focus on probabilistic forecasting. The model does not just output point estimates of future values; instead, it predicts the parameters of a probabilistic distribution, allowing for uncertainty quantification. The distribution head in Lag-Llama projects the model's internal features into the parameters of a chosen probability distribution. For their initial experiments, the authors selected a Student’s t-distribution (Student, 1908), which provides flexibility in modeling heavy-

tailed distributions that are often present in real-world time series data. By predicting the degrees of freedom, mean, and scale of the distribution, the model is capable of generating probabilistic forecasts that offer a range of possible outcomes, not just a single deterministic prediction. This feature is particularly useful in fields such as finance or climate forecasting, where uncertainty is a critical factor in decision-making.

One of the challenges that Lag-Llama addresses is the variability in the magnitude of time series data. For instance, time series in one dataset might represent stock prices in dollars, while another could track energy usage in kilowatt-hours. To handle these discrepancies, Lag-Llama employs value scaling (Salinas et al., 2020). This technique involves normalizing each univariate window of data by its mean and variance, thus standardizing the data across different scales. This scaling ensures that the model can train more effectively on diverse datasets and still maintain generalization capabilities. During training, the time series data is standardized, and during inference, the predictions are de-standardized, ensuring that the outputs remain interpretable and in line with the original data units.

To further improve generalization and robustness, data augmentation techniques are incorporated into the training process. Lag-Llama utilizes Freq-Mix and Freq-Mask (Chen et al., 2023), which are augmentation methods that operate in the frequency domain. These techniques modify the frequency components of the time series data, simulating variations in the underlying signals. This not only helps the model avoid overfitting to the training data but also improves its ability to handle noise and unexpected patterns in the test data. Additionally, stratified sampling is employed to ensure that datasets with varying time series frequencies and lengths are proportionally represented during training. This prevents over-representation of certain datasets, ensuring that the model does not disproportionately learn from a few dominant datasets.

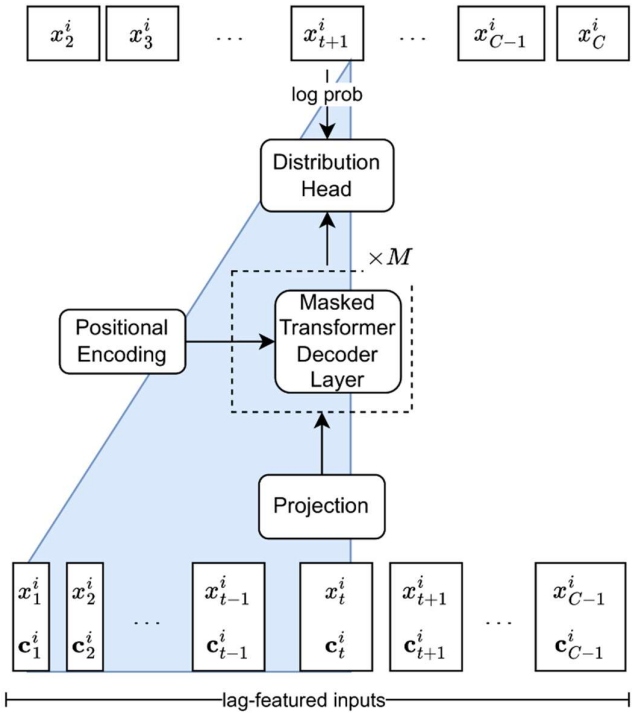


Figure 3 - The Lag-Llama architecture. Reprinted from (Rasul, K., et al., 2024).

Another strength of Lag-Llama lies in its ability to perform well in zero-shot settings, where the model is evaluated on datasets it has never encountered during training. This zero-shot capability is particularly valuable in real-world applications, where new datasets emerge continuously, and retraining or fine-tuning may not be feasible. The model's ability to generalize across datasets from different domains—such as finance, energy, and healthcare—without task-specific tuning highlights its potential as a versatile and scalable solution for time series forecasting.

Overall, Lag-Llama combines the strengths of transformer-based architectures with novel adaptations tailored to time series forecasting. By leveraging lag features, probabilistic forecasting mechanisms, value scaling, and frequency-based augmentations, the model sets a new standard for handling univariate time series data at scale. Its robust performance in zero-shot scenarios positions it as a strong contender in the evolving landscape of foundation models for time series forecasting, capable of addressing challenges across diverse domains.

3.2.3 TimeGPT

The TimeGPT model, developed by Garza and Mergenthaler-Canseco (2023), introduces a foundation model specifically designed for time series forecasting. Unlike traditional

models that are tailored to specific datasets, TimeGPT leverages self-attention mechanisms inspired by transformer architectures (Vaswani et al., 2017) to generalize across a wide range of time series data from different domains. The model is designed to handle the complexities of diverse datasets, each with unique temporal patterns, such as varying frequencies, trends, and levels of noise. By training on over 100 billion data points from sectors like finance, healthcare, and energy, TimeGPT is equipped to manage a wide variety of forecasting tasks with minimal retraining (Challu et al., 2023). Its encoder-decoder structure, which allows for the processing of both short-term and long-term dependencies, makes the model highly adaptable and capable of generating accurate forecasts across diverse applications.

The core of TimeGPT's architecture relies on its ability to manage time series data with various characteristics by utilizing local positional encoding to maintain the temporal order of the data, even when dealing with irregular or missing values (Salinas et al., 2020). This feature enhances the model's ability to work with real-world datasets where data collection may be inconsistent. The residual connections and layer normalization applied throughout the layers of the architecture ensure the model can be trained effectively, avoiding vanishing gradients and improving generalization across different datasets (Vaswani et al., 2017). These architectural adjustments make TimeGPT particularly robust when applied to datasets with varying degrees of complexity and noise.

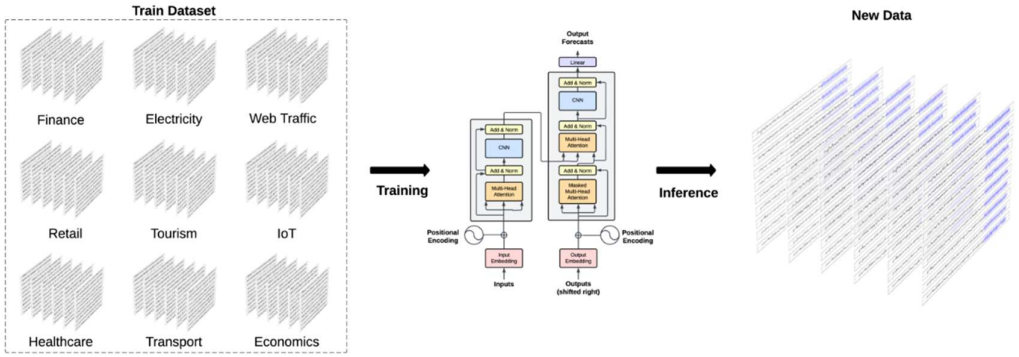


Figure 4 - TimeGPT architecture and training process. Reprinted from (Garza, A., et al., 2024).

A major strength of TimeGPT lies in its capacity for probabilistic forecasting. Instead of generating single-point predictions, the model can provide a range of possible future outcomes with associated confidence intervals. This is achieved using conformal prediction methods, which allow the model to produce flexible, non-parametric prediction

intervals that are not constrained by traditional distributional assumptions (Shafer & Vovk, 2008). Probabilistic forecasting is especially valuable in industries such as finance and energy, where uncertainty must be accounted for in decision-making processes. TimeGPT's ability to provide both accurate predictions and associated uncertainties ensures that it can support more informed, data-driven decisions across various sectors.

The training process for TimeGPT was conducted over multiple days using a cluster of NVIDIA A10G GPUs, allowing for extensive hyperparameter optimization. Key parameters, such as learning rates and batch sizes, were fine-tuned to maximize the model's performance. The authors observed that using a larger batch size in combination with a smaller learning rate resulted in more effective training, a finding consistent with other large-scale models such as GPT-3 (Brown et al., 2020). The training dataset encompassed over 100 billion data points, capturing a wide variety of temporal patterns, from regular, cyclical patterns to highly noisy and irregular data, ensuring that TimeGPT can generalize effectively to both structured and chaotic time series data (Taylor & Letham, 2018).

Zero-shot inference is one of TimeGPT's standout capabilities. This means that the model can be applied to new, unseen datasets without any additional fine-tuning or training. In extensive testing, TimeGPT outperformed traditional statistical models as well as cutting-edge deep learning models across a wide range of domains (Lim et al., 2021). This zero-shot performance is particularly advantageous in real-world applications, where the need to continually retrain or fine-tune models can be both time-consuming and resource-intensive. TimeGPT's ability to generate accurate forecasts across different industries—such as finance, healthcare, energy, and retail—without requiring additional optimization makes it a versatile tool for forecasting tasks on a large scale.

Moreover, TimeGPT's efficiency in terms of computational resources is notable. In tests comparing inference speed, TimeGPT was able to generate forecasts in milliseconds per series, a performance that rivals even the simplest statistical models. This speed, combined with its zero-shot capability, positions TimeGPT as a highly efficient model that can be used for large-scale forecasting with minimal computational overhead (Challu et al., 2023). This efficiency is critical for industries that need to process vast amounts of data in real-time, such as financial markets or IoT applications, where timely forecasts are essential for decision-making.

To summarize, TimeGPT represents a major advancement in the field of time series forecasting. By leveraging the flexibility and power of transformer-based architectures, TimeGPT can handle a wide variety of forecasting tasks with minimal retraining. Its ability to produce both point forecasts and probabilistic forecasts, along with its efficient zero-shot inference capabilities, makes it a powerful tool for industries that rely on time series data. As a foundation model, TimeGPT highlights the potential for transformer architectures to expand beyond natural language processing into other domains, offering a promising future for large-scale time series forecasting models.

3.2.4 MORAI - Uni2TS

The MOIRAI model, as outlined in the work of Woo et al. (2024), addresses the challenge of creating a universal time series forecasting model capable of handling a wide range of downstream tasks and datasets. Time series data is often heterogeneous, presenting unique issues such as multivariate forecasting, cross-frequency learning, and dealing with different distributional properties. Traditional models are typically trained for specific tasks, limiting their generalization across diverse scenarios (Salinas et al., 2020). MOIRAI overcomes these limitations with its innovative architecture and training methodology, designed to scale across various time series frequencies and domains.

One of the defining features of MOIRAI is its capability to process multivariate time series through Any-variate Attention, which allows multiple variates to be treated as part of a unified sequence, rather than being modeled independently (Nie et al., 2023). This not only captures the relationships between different variates but also enhances forecasting accuracy in more complex, interconnected datasets. Additionally, MOIRAI utilizes Rotary Position Embeddings (RoPE) and binary attention biases, further improving the model's ability to capture temporal patterns and distinguish between different variates (Su et al., 2024). This combination allows MOIRAI to handle both univariate and multivariate time series data without significant changes to the architecture, making it a highly flexible and scalable model.

To effectively manage time series data with varying frequencies, MOIRAI introduces multi-patch size projection layers. Traditional time series models often use a fixed patch size, which can limit their ability to generalize across datasets with different temporal sampling rates. MOIRAI addresses this by utilizing multiple projection layers with different patch sizes, each optimized for specific frequency ranges (Oreshkin et al., 2020).

This design allows the model to handle both high-frequency and low-frequency data effectively, making it adaptable to a wide range of time series datasets, from those with fine-grained daily observations to those with coarser, monthly or yearly data.

Probabilistic forecasting is another crucial aspect of MOIRAI. Unlike models that produce only point estimates, MOIRAI generates probabilistic forecasts by modeling a mixture distribution. This approach incorporates multiple distributions—such as the Student's t-distribution, negative binomial, and log-normal distributions—to better capture the variability and uncertainty inherent in time series data (Salinas et al., 2020). By predicting the parameters of these distributions, MOIRAI can produce a range of possible outcomes, providing a more nuanced and flexible forecast compared to models that assume a single, fixed distribution. This makes MOIRAI particularly useful in applications where uncertainty is critical, such as financial forecasting or risk management.

The training process for MOIRAI is powered by the Large-scale Open Time Series Archive (LOTSA), a dataset containing over 27 billion observations across nine different domains. This large and diverse dataset exposes the model to a wide variety of time series patterns, improving its ability to generalize across different domains and tasks. During training, the model employs a masked encoder strategy, where sections of the time series are masked, encouraging the model to learn the underlying structure of the data in an unsupervised manner (Woo et al., 2024). This approach, combined with the extensive training dataset, allows MOIRAI to perform well in both in-distribution and out-of-distribution tasks, making it a powerful tool for zero-shot forecasting.

To conclude, MOIRAI represents a significant advancement in universal time series forecasting models by integrating innovative features such as Any-variate Attention, multi-patch size projections, and probabilistic mixture distributions. Its ability to handle diverse time series data, combined with its scalable and flexible architecture, makes it an ideal solution for a wide range of forecasting tasks. Leveraging the LOTSA dataset, MOIRAI demonstrates strong performance in zero-shot forecasting, highlighting its potential as a foundation model for time series analysis.

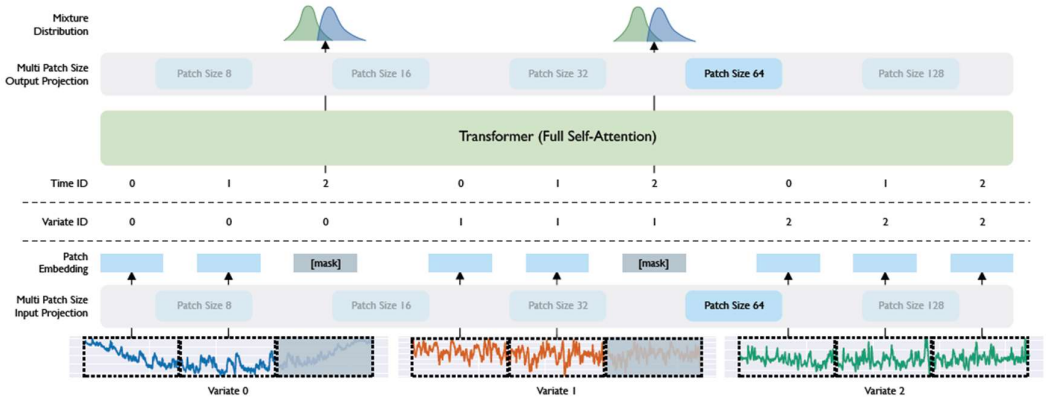


Figure 5 - Overall architecture of MOIRAI, showing a 3-variate time series where variates 0 and 1 are target variables (to be forecasted), and variate 2 is a dynamic covariate (with known values in the forecast horizon). Reprinted from (Woo et al., 2024).

3.2.5 TimesFM

The TimesFM model, developed by Das et al. (2024), introduces a decoder-only architecture for time-series forecasting that leverages large-scale pretraining on diverse datasets. Unlike traditional time-series models, which often require extensive fine-tuning or supervision, TimesFM is designed to provide zero-shot forecasting across a variety of previously unseen datasets. The goal of this model is to produce state-of-the-art performance without the need for dataset-specific retraining. TimesFM's architecture is inspired by the success of language models in NLP, but it incorporates key innovations to address the unique challenges of time-series data (Vaswani et al., 2017).

The model architecture of TimesFM is fundamentally different from traditional encoder-decoder structures in time-series forecasting. It is a decoder-only architecture, meaning that the model predicts future values based solely on past data, without needing an external encoder to process the input sequence (Das et al., 2024). The architecture uses input patching, where the time-series data is broken down into contiguous patches. These patches are treated similarly to tokens in language models, allowing the model to process sequences more efficiently (Nie et al., 2022). This patch-based approach is inspired by recent advances in long-horizon forecasting, which have demonstrated that breaking data into patches can improve both performance and inference speed. By reducing the number of tokens fed into the transformer, TimesFM significantly accelerates its training and inference processes.

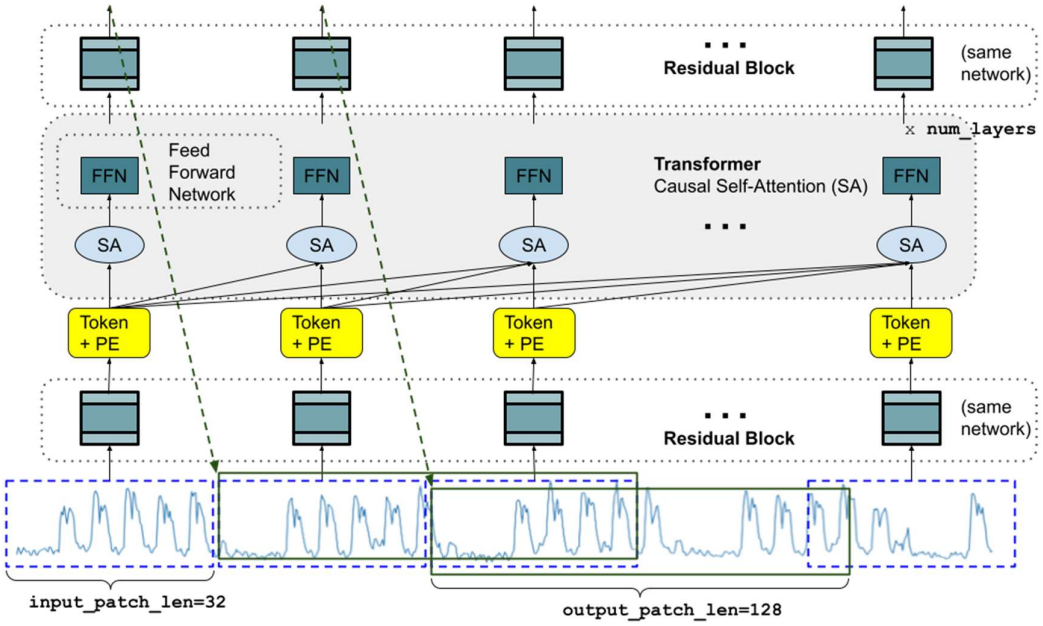


Figure 6 - Illustration of the model architecture during training. Reprinted from (Das, A., et al., 2024).

An important feature of TimesFM's architecture is the use of longer output patches. While typical autoregressive models generate one token at a time, TimesFM predicts larger chunks of future values in a single step. For instance, given an input patch of 32 time points, the model might predict 128 future points at once. This approach reduces the number of steps required for long-horizon forecasts, resulting in more efficient and accurate predictions. However, there is a trade-off: predicting too many points at once can lead to performance degradation, especially for shorter time-series (Das et al., 2024). To address this, the model uses a flexible patching strategy that adapts to the specific requirements of each forecasting task.

The model also employs a unique patch masking strategy during training. This strategy ensures that the model learns to handle variable context lengths by randomly masking parts of the input patches, forcing the model to learn from incomplete data. This masking technique is critical for enabling the model to generalize across datasets with varying lengths and granularities (Vaswani et al., 2017). Additionally, TimesFM uses causal attention, meaning that each prediction can only attend to past data points, ensuring that the model respects the temporal structure of time-series data.

A key advantage of TimesFM is its ability to handle zero-shot forecasting. Unlike other models that require extensive fine-tuning on new datasets, TimesFM is pretrained on a

large corpus of time-series data, allowing it to make accurate predictions on previously unseen datasets without additional training. The pretraining data includes both real-world datasets (such as Google Trends and Wikipedia pageviews) and synthetic time-series data. This diverse training set ensures that the model learns a wide variety of temporal patterns, improving its ability to generalize across domains (Das et al., 2024).

Overall, TimesFM represents a significant advancement in time-series forecasting by introducing a decoder-only architecture that is optimized for efficiency and scalability. Its use of input patching, causal attention, and flexible output patch lengths enables it to handle long-horizon forecasting tasks with high accuracy. Furthermore, its zero-shot capabilities make it a powerful tool for real-world applications where retraining on new data is not feasible. By leveraging both real-world and synthetic data during pretraining, TimesFM demonstrates strong performance across a wide range of time-series datasets, setting a new standard for foundation models in this field.

3.3 Benchmarks

The Benchmark Models discussed include ARIMA (Autoregressive Integrated Moving Average), ETS (Exponential Smoothing State Space), Seasonal Naïve, and Naïve models. These models serve as key baseline techniques for evaluating advanced forecasting models such as transformers.

ARIMA is a traditional model used extensively for stationary time series forecasting. It combines autoregressive terms with moving averages and differencing to stabilize the time series (Box & Jenkins, 1970). While effective for univariate time series, ARIMA is less ideal for handling complex seasonal patterns or multivariate datasets. The model's ability to model long-term dependencies through autoregression makes it a popular choice, though it can become computationally expensive as complexity increases (Hyndman & Athanasopoulos, 2018). In the study conducted by Ramos, Santos, and Rebelo (2015), ARIMA was used in the retail sector for consumer sales forecasting, where its performance was compared to state-space models. Their results showed that ARIMA models were effective in short-term retail sales forecasting, but struggled with long-term predictions, particularly when seasonality or complex trends were involved. This highlights the limitations of ARIMA when applied to sectors with dynamic and irregular sales cycles.

ETS (Exponential Smoothing State Space) is another powerful model that handles time series data with error, trend, and seasonal components. It smooths data through exponential functions, making it highly adaptive to recent changes in trends and seasonality. The flexibility of the ETS model, particularly in handling both additive and multiplicative seasonality, makes it suitable for a wide range of forecasting applications. However, like ARIMA, it may struggle with highly complex or irregular multivariate datasets. According to the research by Makridakis, Spiliotis, and Assimakopoulos (2020), which evaluated 61 forecasting methods over 100,000 time series datasets in the M4 competition, ETS emerged as one of the stronger traditional models, particularly when dealing with datasets that exhibited clear seasonal trends. However, the study also indicated that ETS could not outperform newer machine learning and hybrid approaches on datasets with irregular trends or nonlinear patterns.

The Naïve and Seasonal Naïve models represent simple benchmarks. The Naïve model assumes that the next data point will be identical to the last observed point, which is useful for series with no significant trend or seasonality (Makridakis et al., 1998). On the other hand, the Seasonal Naïve model replicates values from the same season in the previous cycle, making it an effective and computationally efficient method for data with strong seasonal patterns. These simple models are critical for comparison, as they provide a lower-bound performance metric, highlighting whether more complex models deliver genuine improvements in forecasting accuracy.

These benchmark models are fundamental in evaluating the improvements introduced by advanced forecasting techniques. As highlighted in the M4 competition (Makridakis et al., 2020), newer techniques, including machine learning and hybrid methods, generally outperform traditional methods like ARIMA and ETS on complex datasets. Nevertheless, these benchmarks offer transparency and simplicity, making them indispensable for gauging the effectiveness of newer, more complex models in practical forecasting tasks. The comparison by Ramos et al. (2015) between ARIMA and state-space models in consumer sales also emphasized the importance of domain-specific evaluation, as ARIMA's performance varied with the complexity and nature of the data.

4 Empirical Study

This chapter presents an empirical study on retail sales forecasting using the M5 dataset. Effective forecasting is critical in retail for inventory management, demand planning, and promotions. By analyzing historical data, this study evaluates five models—TimeGPT, Lag-Llama, TimesFM, Uni2TS, and Chronos—for their ability to deliver accurate and actionable forecasts. The study was conducted using the Google Colab environment, recognized for its flexibility, accessibility, and ability to integrate data science and machine learning tools. Google Colab was chosen due to its availability of free computational resources, native integration with advanced libraries, and compatibility with GitHub repositories, which were essential for implementing and analyzing the forecasting models.

Each model underwent fine-tuning to adapt to the dataset's unique patterns, capturing key elements such as seasonality and regional demand variations. The environment setup involved installing necessary dependencies via automated commands in the Colab terminal, enabling the use of libraries such as GluonTS, Nixtla, TimeGPT, and AutoGluon, as well as data manipulation tools like pandas and NumPy. GitHub repositories were directly integrated into the environment to access the latest implementations of the models. Among the repositories used were those for the five models evaluated: TimeGPT (*https://github.com/Nixtla/nixtla*), Lag-Llama (*https://github.com/time-series-foundation-models/lag-llama*), TimesFM (*https://github.com/google-research/timesfm*), Uni2TS (*https://github.com/SalesforceAIResearch/uni2ts*), and Chronos (*https://github.com/amazon-science/chronos-forecasting*). These repositories provided the latest implementations and documentation necessary for the study.

The study assesses performance using Mean Absolute Scaled Error (MASE) for point forecast accuracy and Continuous Ranked Probability Score (CRPS) for probabilistic forecast quality, comparing results in both Zero-Shot and Fine-Tuning settings. Each model was carefully configured to analyze its adaptability and predictive capabilities, exploring different hyperparameter configurations to evaluate the impact of fine-tuning on performance. Fine-tuning allowed the models to better capture the unique characteristics of the M5 dataset, including seasonality and regional demand variations.

This approach provides insight into the effect of fine-tuning on model effectiveness. The evaluation was based on quantitative metrics such as MASE and CRPS, which allowed a detailed comparison between zero-shot and fine-tuning approaches. To further complement the analysis, visualizations of the forecasts were created, highlighting qualitative differences and enabling a comprehensive assessment of the models' capabilities.

In summary, this chapter highlights the practical role of fine-tuned forecasting models—TimeGPT, Lag-Llama, TimesFM, Uni2TS, and Chronos—in retail, underscoring their importance in achieving reliable, data-driven decision-making. By integrating computational tools, advanced predictive models, and quantitative analysis methodologies within the Google Colab environment, the study demonstrates the effectiveness of these approaches in addressing the challenges of retail sales forecasting.

4.1 Dataset M5

The use of the M5 dataset in this research is justified by its meticulous construction and comprehensive nature, which mirrors the complexities of real-world retail sales. Released as part of the M5 Forecasting Competition, the dataset quickly gained traction among researchers and practitioners in the field of demand forecasting. Its rich structure and diverse features make it one of the most comprehensive publicly available datasets, providing a challenging testbed for time series forecasting.

The dataset offers a realistic representation of challenges faced in the retail sector, such as hierarchical relationships between products and regions, temporal dependencies, seasonality, and external factors like promotions and price fluctuations. These characteristics allow for the evaluation of forecasting models under conditions that closely mirror real-world scenarios, making it particularly suitable for this comparative analysis of advanced forecasting techniques.

Additionally, the M5 dataset provides a diverse and hierarchical framework, encompassing data from multiple levels of granularity, such as individual products, departments, and geographical regions. This enables the testing of models' ability to generalize across different contexts and scales. Its temporal depth, capturing daily sales over several years, facilitates the examination of long-term trends, seasonal patterns, and the impact of promotional events and external variables.

Its inclusion of real-world complexities—such as sparse sales data, periods of zero demand, and irregular spikes due to promotions—ensures that models are tested against scenarios that are both challenging and representative of retail dynamics. By integrating these aspects, the M5 dataset serves as an essential benchmark for evaluating how different forecasting methods, including statistical models, machine learning techniques, and deep learning architectures, can perform in practical business applications. Its adoption in this study underscores its value as a gold standard for demand forecasting research.

4.1.1 Historical Sales Data for the M5 Dataset

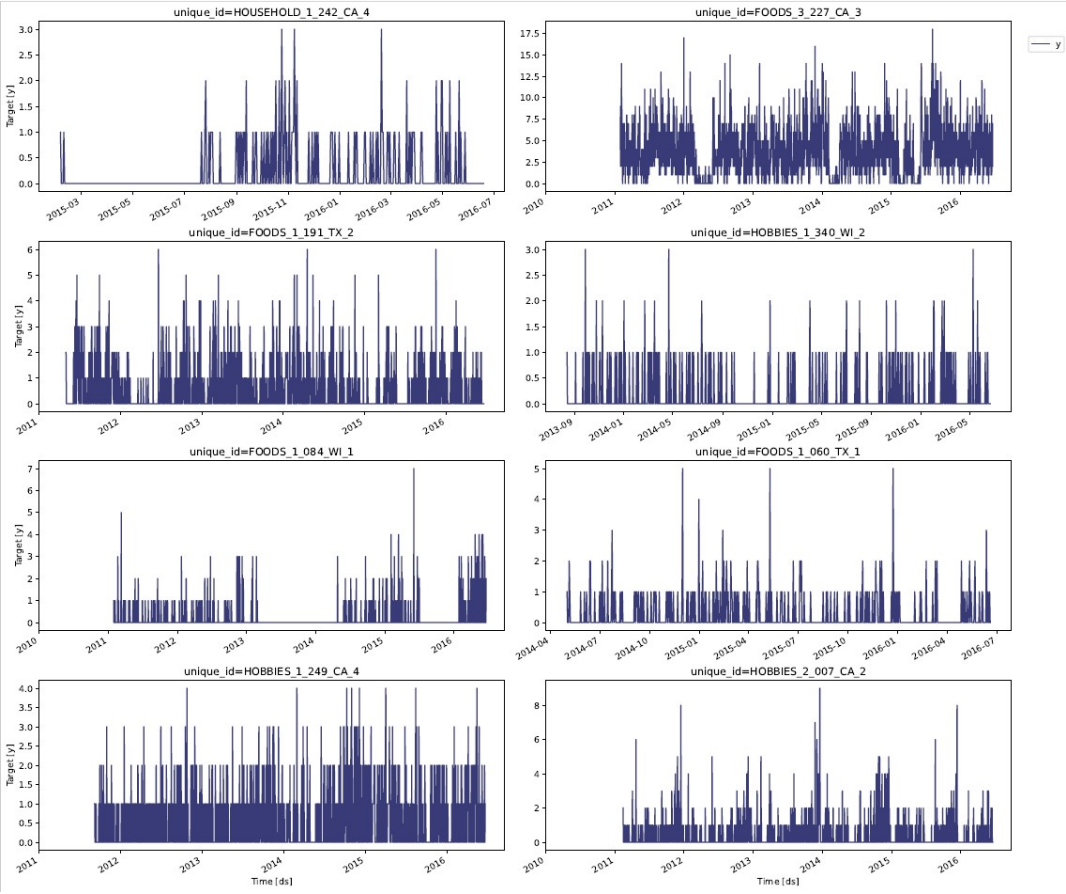


Figure 7 - Historical sales data for various product-region combinations in the M5 dataset across categories such as Household, Foods, and Hobbies.

Figure 7 displays the historical sales data for various products in different categories (e.g., Household, Foods, Hobbies) across multiple regions in the M5 dataset. Each subplot represents a unique product-region combination, showcasing the daily sales volume over time.

X-Axis (Time): Spanning from 2011 to mid-2016, indicating the historical period covered by the dataset.

Y-Axis (Target Sales): Shows the sales quantity, where each point represents the sales on a specific day.

Product and Region Labels: Each title identifies a unique product and store location, using identifiers from the M5 dataset.

- **Key Observations**:

Seasonality and Patterns: Some products display seasonal trends, with regular peaks and troughs that could be attributed to seasonality or specific sales events.

Sales volatility: Certain items, particularly in categories like Hobbies and Foods, show irregular sales patterns, potentially indicating responses to promotions, holidays, or varying consumer demand.

Sparse sales data: Some products have intermittent sales, indicating days with zero sales, which is common in non-essential or specialty categories like Hobbies.

This historical data is crucial for training and validating forecasting models in retail demand prediction, as it contains the underlying patterns, trends, and anomalies typical in retail sales. These insights help in developing models that can anticipate future sales and manage inventory effectively.

4.1.2 Realism in Retail Sales Simulation

The dataset is sourced from Walmart, one of the world's largest retailers, ensuring that the data reflects realistic sales patterns observed across a diverse range of products and locations. Covering approximately 30,000 unique products sold across 10 stores in 3 distinct U.S. states (California, Texas, and Wisconsin), the dataset spans a period of 1,969 days (approximately five years), offering both breadth and depth in terms of product diversity and geographic coverage.

The realism of the dataset comes not just from its sales data but also from its detailed supplementary information, including calendar data, price fluctuations, and promotional event data, all of which are crucial for understanding the dynamic nature of retail demand. Retailers face continuous changes in consumer behavior, influenced by external factors such as holidays, price changes, and promotional campaigns. These influences are

meticulously captured in the M5 dataset, allowing forecasters to model not only product-specific trends but also external events that affect purchasing behavior.

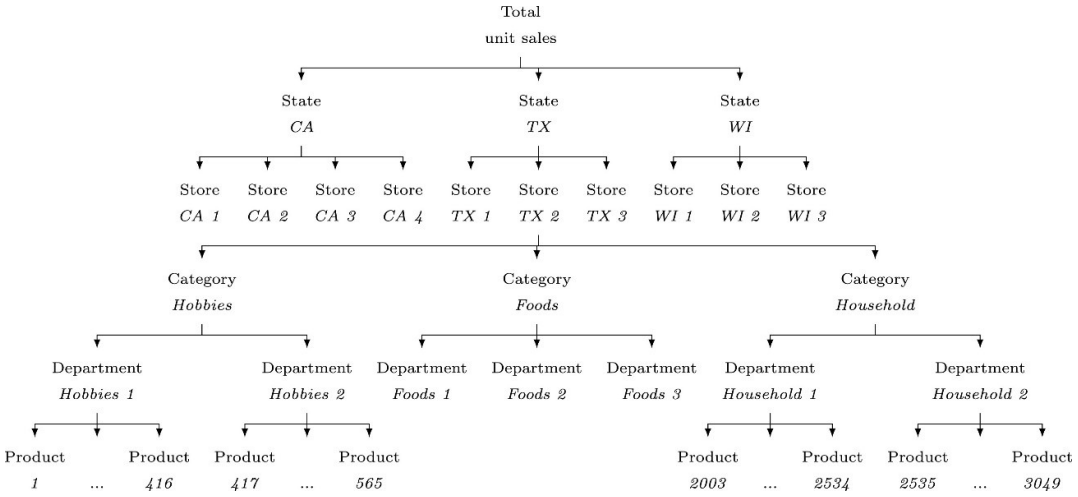


Figure 8 - Grouped time series used in the M5 competition. The data can be aggregated in 12 different levels using either location (state and store) or product-related information (category and department).

4.1.3 Complexity Arising from Hierarchical Relationships

One of the most notable features of the M5 dataset is its hierarchical structure, which introduces a high degree of complexity. This hierarchical nature means that the sales data can be viewed from multiple perspectives, each offering unique challenges for forecasters:

Product hierarchy: Products are categorized into various levels of detail, beginning with the individual SKU (stock-keeping unit) level and then aggregating upward into broader categories such as departments and product families. This nested structure is essential for retailers, as forecasts are often needed at different levels of granularity, depending on the decision being made (e.g., predicting demand for individual items versus entire departments).

Geographic hierarchy: In addition to product categories, the dataset includes a geographic hierarchy, where sales are recorded at the store level but can also be aggregated to represent state-level or even national-level sales. This hierarchy is critical for understanding regional variations in demand, as consumer preferences and purchasing patterns may differ across states or even individual stores.

These multi-level hierarchies introduce a level of complexity that is not present in simpler time series datasets. Forecasting models must be able to accurately predict sales across different hierarchical levels, which may have varying levels of aggregation and granularity. The ability to capture both high-level trends (e.g., seasonal sales spikes at the department or category level) and fine-grained patterns (e.g., daily sales fluctuations for a specific product in a specific store) is key to achieving high forecasting accuracy.

4.1.4 Temporal Dependencies and Seasonality

The time series data in the M5 dataset captures daily sales over multiple years, which means that it encompasses a wide range of temporal dependencies. One of the main challenges of forecasting retail demand is accounting for seasonality, as consumer purchasing behavior often follows cyclical patterns influenced by external factors such as holidays and weather changes.

The dataset captures these seasonal trends in great detail. For example, sales tend to surge during holiday periods such as Thanksgiving, Christmas, and Black Friday, while demand for certain products (e.g., winter clothing or school supplies) may fluctuate based on seasonal cycles. Additionally, the dataset includes non-traditional holidays or region-specific events that affect consumer behavior. Forecasters must model these temporal dependencies accurately to ensure that the model can capture both short-term fluctuations and long-term trends.

Furthermore, the dataset records promotional events and price discounts, which add another layer of complexity. Promotions can cause sharp spikes in demand for particular products, but these spikes are often short-lived. Predicting when and how these promotions will affect sales is a crucial component of demand forecasting in the retail industry. Models that do not take these factors into account may either overestimate or underestimate sales, leading to inaccurate predictions and suboptimal inventory management.

4.1.5 External Factors and Contextual Information

In addition to internal factors such as product characteristics and historical sales, the M5 dataset includes a wealth of external contextual information that can have significant impacts on demand. This includes:

Calendar data: The calendar dataset provides essential information about the dates of holidays and special events that may influence consumer behavior. The presence of national holidays like Independence Day or state-specific holidays allows forecasters to factor these into their predictions. By linking sales data to specific calendar events, models can account for predictable variations in demand tied to these external events.

Price data and promotions: Prices fluctuate over time due to promotions, markdowns, and other retailer-driven events. The dataset includes detailed information on product prices for each day, allowing models to analyze how price elasticity affects demand. For instance, certain products may exhibit higher sensitivity to price changes, leading to sharp increases in sales during promotions. Forecasting models must incorporate this price sensitivity into their predictions to generate more accurate and actionable insights.

Promotional campaigns: Promotions are often designed to drive short-term sales increases. However, their impact can vary depending on the timing, the product category, and the extent of the discount. The dataset records the occurrence of promotions, providing a rich source of information for understanding how temporary price cuts and special offers affect consumer behavior.

4.1.6 Real-world Retail Challenges Reflected in M5

The challenges presented by the M5 dataset closely resemble those faced by retailers in the real world, making it highly relevant for the retail industry. Inventory management, supply chain logistics, and financial planning all depend on accurate demand forecasting, especially in a dynamic and competitive market like retail.

For example:

- Sparse sales data: Many products in the dataset exhibit sparse demand, with long periods of no sales followed by short bursts of activity. This intermittent demand pattern is common in retail, particularly for non-essential or seasonal products. Forecasting such demand accurately requires models that can handle sparsity without overfitting or underpredicting.
- Stockouts and overstocking: Retailers aim to minimize both stockouts (when products are unavailable) and overstocking (when excess inventory leads to wasted resources). Inaccurate demand forecasting can lead to significant costs in

both cases, making the ability to predict future sales with precision a crucial competitive advantage.

- Promotional strategy optimization: Understanding the impact of price promotions and markdowns is another significant challenge reflected in the M5 dataset. Retailers must optimize their promotional strategies to balance short-term sales increases with long-term profitability. Forecasting models that can accurately predict the impact of promotional events allow retailers to design more effective promotional campaigns and maximize their return on investment.

4.1.7 Opportunities for Forecasting with LLMs

The use of large language models (LLMs) for time series forecasting in the retail sector offers several key advantages. Models such as Chronos, TimeGPT, Lag-Llama, Uni2TS, and TimesFM are specifically designed to tackle the complexity and challenges presented by the M5 dataset. Some of the major opportunities provided by LLMs include:

Multivariate modeling: LLMs excel at processing and modeling multivariate time series, meaning they can predict sales while considering not only historical sales data but also external factors like prices, promotions, holidays, and special events. This is critical for the M5 dataset, which includes variables such as price fluctuations and calendar data.

Capturing seasonal and long-term patterns: LLMs like TimeGPT are particularly adept at capturing long-range dependencies and complex seasonal patterns, both of which are crucial for retail forecasting. They can model long-term effects and recurring events such as holiday sales spikes or the impact of seasonal promotions.

Hierarchical temporal and spatial relationships: Models like Chronos and Uni2TS are designed to handle hierarchical time series, which allows them to capture relationships between sales data at various levels, such as individual store sales versus state-level or national-level aggregates. This helps in modeling the interactions between different levels of data aggregation.

Contextual learning: With the inclusion of calendar data and promotions, LLMs can forecast demand in different contexts, adjusting predictions based on special events or price fluctuations. Lag-Llama stands out in capturing these contextual variations, especially when dealing with sparse and intermittent demand patterns.

Scalability: The scalability of LLMs, especially in the Base and Large versions of models like Chronos, makes them suitable for processing large datasets without compromising accuracy. This is essential for handling high-complexity datasets like M5, which involve thousands of interdependent time series.

More accurate and fine-tuned forecasts: LLMs can consider complex factors and nonlinear interactions between variables, resulting in more accurate forecasts, particularly in complex retail scenarios such as predicting the impact of promotions across multiple stores or regions.

4.2 Fine-tuning

Fine-tuning is a critical process in machine learning, especially for large language models (LLMs), which are pre-trained on vast amounts of general data. The purpose of fine-tuning is to adapt these pre-trained models to specific tasks or domains by training them further on specialized datasets. This process helps models produce more accurate and relevant outputs for the target application (Pan & Yang, 2010). Fine-tuning allows these models to retain the general knowledge acquired during pre-training while specializing in a more focused area, improving performance on particular tasks.

The core process of fine-tuning involves several stages. First, the model is pre-trained on a large, diverse dataset that covers a wide range of topics. Once the model achieves a good level of generalization, fine-tuning is performed by exposing the model to a smaller, task-specific dataset. This dataset contains information specific to the problem the model needs to address. During fine-tuning, the model's weights are slightly adjusted, allowing it to focus on features that are more relevant to the new data. This process helps the model retain the benefits of general pre-training while specializing in the new domain.

Transfer learning plays an essential role in fine-tuning. It enables the knowledge learned from one task (or general dataset) to be transferred to another, related task (Pan & Yang, 2010). This technique is particularly useful when the data available for the new task is limited, as the pre-trained model already possesses a strong foundational understanding. Fine-tuning, in this context, is a way to transfer that general knowledge to a specialized domain, improving the model's performance on specific tasks.

Instruction-tuning is one of the most prominent methods used to enhance the zero-shot and few-shot learning abilities of large models (Zhang et al., 2023). In this approach,

models are fine-tuned on task-specific instructions formatted in natural language. By training the model with datasets presented as instructions, it becomes better at understanding and responding to user prompts. This method significantly improves the model's ability to generalize across unseen tasks, making it more flexible and efficient in handling a variety of different problems. For instance, instruction-tuned models can perform tasks like text summarization, translation, and sentiment analysis with minimal task-specific training data.

Instruction-tuning is particularly valuable for improving multi-task learning capabilities, allowing a single model to handle multiple tasks simultaneously. For example, models such as GPT-3 have been fine-tuned to follow instructions more effectively by learning from examples of task descriptions and prompts (Zhang et al., 2023). This improves their accuracy in completing tasks even when they have not been explicitly trained on those tasks during the pre-training phase.

However, fine-tuning presents several challenges. One of the most significant challenges is the risk of catastrophic forgetting, where the model loses some of its general knowledge while adapting to the new, specialized data. To mitigate this risk, techniques such as parameter-efficient fine-tuning are employed. For instance, Low-Rank Adaptation (LoRA) allows the model to update only a subset of its parameters, significantly reducing the risk of forgetting while minimizing computational costs (Tian et al., 2023). This method enables large models to be fine-tuned effectively without requiring the massive computational resources typically associated with full-model fine-tuning.

Fine-tuning is also essential in aligning large language models with human values and preferences. Alignment-tuning, often achieved through Reinforcement Learning from Human Feedback (RLHF), ensures that models generate ethical, safe, and helpful responses. RLHF involves training models based on feedback from human evaluators, who rank the quality of the model's output. The model is then fine-tuned to optimize for these preferences, reducing the chances of generating biased, harmful, or inaccurate text (Zhang et al., 2023). This process is crucial for applications where the reliability and safety of the model's output are paramount, such as in healthcare or legal systems.

Direct Preference Optimization (DPO) offers an alternative to RLHF by directly using human preferences to fine-tune models for factual accuracy and content quality. According to Tian et al. (2023), models fine-tuned with DPO showed a significant

reduction in factual errors compared to models fine-tuned using traditional methods. In complex tasks, such as generating biographies or answering medical queries, ensuring factual consistency is critical, and fine-tuning through DPO helps achieve greater reliability in the model's output.

Fine-tuning is widely applied across various industries to enhance the performance of large language models. For example, Raj et al. (2024) discuss how fine-tuning plays a crucial role in adapting LLMs for enterprise solutions. In fields like customer service, legal documentation, and code generation, fine-tuning on domain-specific data leads to improved accuracy and task-specific knowledge. This allows enterprises to deploy models that not only generate accurate responses but also handle proprietary or sensitive information appropriately.

In the financial sector, Jeong (2023) highlighted the importance of fine-tuning models to cater to the specific needs of the industry, such as market predictions, regulatory compliance, and fraud detection. Fine-tuned models in finance are designed to understand and process the unique terminology, trends, and datasets associated with stock markets, economic indicators, and customer behavior. For example, fine-tuned models have been employed for real-time stock price prediction, enabling firms to process financial news and prdict market movements with greater accuracy.

In highly regulated industries such as healthcare, fine-tuning helps models understand medical terminology and adhere to privacy and security standards. Models are fine-tuned with domain-specific medical data, enabling them to process patient information, diagnose diseases, or provide treatment recommendations more accurately and reliably.

Crowd-Informed Fine-Tuning (CIFT), introduced by Lalor et al. (2017), represents an innovative fine-tuning approach that uses human feedback and crowd data to inform model adjustments. This method incorporates data from human interaction patterns into the fine-tuning process, helping models better understand and respond to diverse user inputs. By simulating real-world behavior, CIFT helps improve model performance in handling a variety of user scenarios.

In conclusion, fine-tuning is a flexible and essential process in machine learning. It allows pre-trained models to adapt to specific tasks, improving their relevance and performance in practical applications. From instruction-tuning, which enhances the model's ability to follow complex prompts, to alignment-tuning, which ensures ethical and accurate

responses, fine-tuning plays a critical role in tailoring models to meet the demands of different industries. Techniques such as parameter-efficient tuning and reinforcement learning help address challenges in the fine-tuning process, ensuring that models perform well without compromising their general knowledge or computational efficiency.

4.3 Performance Metrics

In this study, we used two key metrics to evaluate the performance of our time series forecasting models: the Mean Absolute Scaled Error (MASE) and the Continuous Ranked Probability Score (CRPS). Each metric serves a distinct purpose, allowing us to assess both point and probabilistic forecasts, which are essential for capturing the complexity of retail demand patterns.

4.3.1 Mean Absolute Scaled Error (MASE)

As explained by Oliveira and Ramos (2024), MASE was chosen as the primary metric for evaluating point forecast accuracy. MASE is advantageous for its scale-independent properties, which make it ideal for comparing models across time series with varying seasonality and demand levels—a common feature in retail data. The MASE formula is defined as follows:

$$\mathrm{MASE}_{i,j} = \frac{1}{H} \sum_{t=L_j+1}^{L_j+H} \frac{|y_{i,t} - \hat{y}_{i,t}|}{\frac{1}{L_j-m} \sum_{t=m+1}^{L_j} |y_{i,t} - y_{i,t-1}|}$$

where:

- $y\,i,t$ is the observed value for series i at time t,
- $\hat{\gamma}\,i,t$ is the forecasted value,
- H is the forecast horizon (set at 28 days),
- Lj denotes the origin of each cross-validation window j,
- J is the number of cross-validation windows (1 in this study, which corresponds to the usual train/test setting),
- m is the seasonal period (7 days, to capture weekly seasonality), and
- N is the total number of time series analyzed (30,490 in this case).

MASE was selected because it accommodates the diverse patterns within large retail datasets by scaling forecast errors relative to the average seasonal error in the sample. This normalization ensures that MASE provides a robust, scale-independent metric for accuracy across various seasonalities and data scales, making it an excellent choice for retail forecasting where demand often fluctuates (Oliveira & Ramos, 2024). Hyndman and Koehler (2006) have recommended MASE for forecasting due to its scale-free properties, which support reliable comparisons across series with differing magnitudes. Additionally, Gneiting and Raftery (2007) have validated MASE's reliability in dynamic forecasting environments, including retail, where demand is highly variable.

4.3.2 Continuous Ranked Probability Score (CRPS)

To complement MASE and evaluate the accuracy of our probabilistic forecasts, we employed the Continuous Ranked Probability Score (CRPS). CRPS is especially valuable for probabilistic forecasting because it measures the accuracy of the full predicted distribution rather than just a single point estimate. In retail forecasting, where it's crucial to anticipate a range of possible outcomes rather than a fixed demand figure, CRPS provides insights that help manage risks associated with under- or over-stocking.

The CRPS metric is calculated as:

$$\mathrm{CRPS}(F_t, y_t) = \int_{-\infty}^{\infty} \left(F_t(x) - 1\{x \geq y_t\}\right)^2 \, dx$$

where:

- $Ft(x)$ represents the cumulative distribution function (CDF) of the forecast at time t.
- yt is the observed value,
- $1\{x \geq yt\}$ is an indicator function equal to 1 if $x \geq yt$ and 0 otherwise.

The CRPS penalizes errors across the entire predictive distribution, with a higher penalty for predictions that deviate significantly from observed values. This property makes CRPS ideal for assessing probabilistic forecasts in retail, where the costs of misestimating demand can vary significantly based on whether a product is over- or under-forecasted. Gasthaus et al. (2019) and Shchur et al. (2023) have also highlighted CRPS's

effectiveness in evaluating model performance in settings where data variability is substantial, further supporting its use in complex forecasting contexts like retail.

4.3.3 Combining MASE and CRPS for Comprehensive Model Evaluation

Using both MASE and CRPS provides a holistic evaluation of forecast performance by balancing the need for accurate point predictions with the requirement for reliable uncertainty estimates. MASE offers a straightforward measure of central accuracy, while CRPS evaluates the spread and reliability of the forecast distribution. This combination is particularly effective in retail forecasting, where precise demand predictions help optimize stock levels and probabilistic forecasts inform risk management and operational planning.

In summary, the dual use of MASE and CRPS provides a nuanced understanding of forecast quality, ensuring that models are not only accurate in their central estimates but also reliable in their probabilistic range. Together, these metrics guide decision-making in retail, supporting efficient inventory control and minimizing the risks associated with demand uncertainty.

4.4 Results and Discussion

4.4.1 Analysis of Chronos Performance (Zero-Shot Evaluation)

In the Zero-Shot setting, where the Chronos model has not been fine-tuned on the M5 dataset, the MASE (Mean Absolute Scaled Error) and CRPS (Continuous Ranked Probability Score) metrics provide insights into each model's accuracy and probabilistic forecasting ability, respectively. These metrics are assessed across five model sizes—Tiny, Mini, Small, Base, and Large—to evaluate how increasing model complexity impacts performance on the M5 dataset, a challenging retail dataset with hierarchical structures, seasonality, and promotional impacts:

Tiny: The smallest and lightest version, focused on efficiency. Although it delivers quick forecasts, it sacrifices some precision in exchange for lower computational requirements. It is ideal for real-time use cases or when hardware resources are limited.

Mini: Slightly more robust than Tiny, Mini remains lightweight but offers better capacity for capturing simple patterns and short-term fluctuations. It is suitable for smaller-scale forecasting tasks or for companies with less complex data needs.

Small: A balanced version between efficiency and accuracy. Small is ideal for moderately sized datasets with clear temporal dependencies. It is a good starting point for models that need more capacity without consuming too many computational resources.

Base: Designed to handle more complex and detailed time series data, the Base model is suitable for larger datasets like the M5, which involve capturing interactions between variables, seasonal patterns, and promotional effects with greater accuracy.

Large: The most powerful version, Large can handle large datasets with extreme complexity, such as data for thousands of products across different stores and locations. Large models are recommended for high-precision forecasts where large volumes of data and multiple interacting factors are involved.

Table 1 - Performance of the Chronos model under Zero-Shot evaluation settings across different model sizes (Tiny, Mini, Small, Base and Large), measured by MASE and CRPS.

Model	Model Size	MASE	CRPS
Chronos	**Tiny**	1.0920	0.5973
	Mini	1.1040	0.5858
	Small	1.0940	0.5817
	Base	1.1072	0.5726
	Large	1.1098	0.5706

4.4.1.1 MASE (Mean Absolute Scaled Error) Analysis

Chronos Tiny has a MASE of 1.0920, which is slightly above 1, indicating that the model performs close to a naive forecast. This score reflects a basic ability to capture general trends but suggests limited accuracy in exact point forecasting, particularly for complex retail patterns in the M5 dataset.

Chronos Mini exhibits a MASE of 1.1040, slightly worse than Tiny, showing that while Mini is marginally more complex, it does not significantly improve point forecast accuracy. This could be due to Mini's limited capacity to capture intricate patterns like seasonal fluctuations or promotional spikes.

Chronos Small shows some improvement, with a MASE of 1.0940, slightly better than Mini, suggesting that as model complexity grows, it starts to capture trends marginally

better, even in Zero-Shot settings. However, the improvement is small, and the MASE still reflects limited precision in point forecasting.

Chronos Base and Large have MASE values of 1.1072 and 1.1098 respectively, showing that, while they manage the larger dataset better, point forecast accuracy remains close to a naive baseline, perhaps due to the lack of fine-tuning. These results indicate that further tuning would be beneficial for achieving more precise point predictions, especially to handle the rich seasonal and promotional nuances in the M5 data.

4.4.1.2 CRPS (Continuous Ranked Probability Score) Analysis

The CRPS values improve consistently as model complexity increases, reflecting that larger models handle uncertainty and probabilistic forecasting more effectively. Lower CRPS values indicate that the model's predicted probability distribution is closer to actual outcomes, which is crucial in retail forecasting where demand uncertainty (e.g., around promotions) needs to be managed carefully.

Chronos Tiny starts with a CRPS of 0.5973, indicating basic probabilistic performance but limited capability in predicting the full range of demand variations seen in the M5 dataset.

Chronos Mini and Small improve to 0.5858 and 0.5817, respectively, showing incremental improvements in probabilistic forecasting. This suggests that these versions better capture demand variability, such as regional differences or small promotions, providing more accurate probabilistic estimates than Tiny.

Chronos Base and Large demonstrate the best CRPS scores, with 0.5726 and 0.5706 respectively. These results suggest that Base and Large versions of Chronos are more reliable for uncertainty management, making them preferable for scenarios where predicting the upper and lower bounds of demand is critical for inventory planning and risk management.

4.4.1.3 Probabilistic and Point Forecasting for Retail Sales Using Chronos Large Model

Figure 9 - Probabilistic and point forecasts of retail product sales generated by the Chronos Large model across various categories and regions.

Figure 9 displays the probabilistic and point forecasts generated by the Chronos Large model for a selection of retail products across various categories and regions. The comparison is between observed (actual) sales data and the model's forecasted values, showcasing both the accuracy of point predictions and the uncertainty captured in probabilistic forecasts.

Observed Sales (Blue Line): The blue line in each subplot represents actual sales data over time, covering the period from early March to mid-June 2016. This data serves as the baseline to assess the model's predictive performance.

Point Forecast (Orange Line): The orange line represents the model's point forecast—the single best estimate of sales for each future time point. This line allows for direct comparison with observed sales to gauge how closely the model's predictions align with actual outcomes.

Probabilistic Forecast (Shaded Region): Surrounding the point forecast, the shaded orange area represents the model's probabilistic forecast, capturing the range within which future sales are expected to fall with a given confidence level. This range reflects forecast uncertainty, with a wider shaded area indicating greater variability and less certainty in predictions.

By incorporating both point and probabilistic forecasts, the Chronos Large model not only provides a central estimate of expected sales but also accounts for uncertainty, which is crucial for inventory management, risk assessment, and strategic planning in retail. The shaded probabilistic forecasts help retailers understand potential fluctuations, such as peaks during promotions or drops in off-season periods, aiding in more informed decision-making.

4.4.1.4 Summary

MASE values across all versions remain above 1, indicating that the model's point forecasting ability is close to a naive benchmark. This suggests that Zero-Shot settings may not capture the M5 dataset's granular patterns, such as promotional spikes and regional variability, and that fine-tuning could enhance these scores.

CRPS consistently improves with model size, showing that larger versions of Chronos are better equipped to handle probabilistic forecasts, a critical advantage in retail demand forecasting. This makes Chronos Large and Chronos Base particularly suitable for managing demand uncertainty and forecasting range-bound predictions.

In conclusion, the Chronos Large and Base models are the most capable for probabilistic forecasting on the M5 dataset, though all model sizes could benefit from further fine-tuning to improve point forecast accuracy.

4.4.2 Analysis of TimeGPT Performance (Zero-Shot and Fine-Tuning Evaluation)

This analysis focuses on two key metrics—MASE (Mean Absolute Scaled Error) and CRPS (Continuous Ranked Probability Score)—to evaluate TimeGPT's performance in both Zero-Shot and Fine-Tuning settings. These metrics assess TimeGPT's ability to handle both point forecasts (MASE) and probabilistic forecasts (CRPS), which are critical for forecasting retail demand with accuracy and confidence.

Table 2 - Performance of the TimeGPT model under different evaluation settings (Zero-Shot and Fine-Tuning), measured by MASE and CRPS.

Model	Evaluation Setting	MASE	CRPS
TimeGPT's	**Zero-Shot**	1.0842	0.6080
	Fine-Tuning	1.0163	0.5794

4.4.2.1 MASE (Mean Absolute Scaled Error) Analysis

- Zero-Shot Evaluation (MASE: 1.0842)

In the Zero-Shot setting, TimeGPT achieves a MASE of 1.0842, which is slightly above 1. A MASE value above 1 indicates that the model's point forecasting accuracy is close to a naive forecast (e.g., assuming that tomorrow's sales will be the same as today's).

This result suggests that while TimeGPT can capture general sales patterns, it struggles to precisely capture finer details in the data, such as seasonal peaks or promotional effects without fine-tuning. This limitation affects its utility in retail settings, where demand fluctuations can be sudden and often tied to external events like holiday sales or special promotions.

- Fine-Tuning Evaluation (MASE: 1.0163)

After Fine-Tuning, the MASE improves to 1.0163, bringing the model's performance closer to an ideal benchmark. This improvement shows that TimeGPT is better able to capture the nuances of the data when specifically trained on it, such as seasonal demand variations, promotional spikes, and regional differences.

A MASE value close to 1 in the Fine-Tuning setting means that TimeGPT's point forecasts are now more accurate and actionable for retail applications, where precise demand predictions are crucial for managing inventory levels and supply chain planning. Fine-tuning enables TimeGPT to adapt to patterns specific to the dataset, thereby improving its ability to predict demand more accurately during high-demand periods (e.g., holiday seasons).

4.4.2.2 CRPS (Continuous Ranked Probability Score) Analysis

- Zero-Shot Evaluation (CRPS: 0.6080)

In the Zero-Shot evaluation, TimeGPT's CRPS is 0.6080, indicating moderate performance in terms of probabilistic forecasting. CRPS measures how well the model can predict the range of possible outcomes (i.e., upper and lower bounds of demand), which is critical in retail forecasting where uncertainty is high, especially during promotions or unexpected events.

A higher CRPS value here indicates that TimeGPT has some difficulty capturing the full range of demand fluctuations without fine-tuning, which limits its effectiveness in managing inventory risks. For instance, in cases where demand spikes unpredictably due to a sudden sale event or regional promotion, the model may under- or over-estimate the extent of demand variability.

- Fine-Tuning Evaluation (CRPS: 0.5794)

After fine-tuning, TimeGPT's CRPS improves to 0.5794, reflecting its enhanced ability to manage uncertainty in sales forecasts. This reduction in CRPS means that the model is now better at predicting the spread of possible sales outcomes, which is particularly important in retail environments where accurately estimating upper and lower demand bounds can help in avoiding stockouts or minimizing overstock.

The lower CRPS score after fine-tuning suggests that TimeGPT is now better able to account for the probabilistic elements in sales data, such as uncertain demand spikes during promotions or sales volatility during holiday seasons. This improvement makes the model more practical for real-world retail applications where managing demand uncertainty is essential for efficient inventory planning and cost-effective supply chain management.

4.4.2.3 TimeGPT Model: Probabilistic and Point Forecasting of Retail Sales

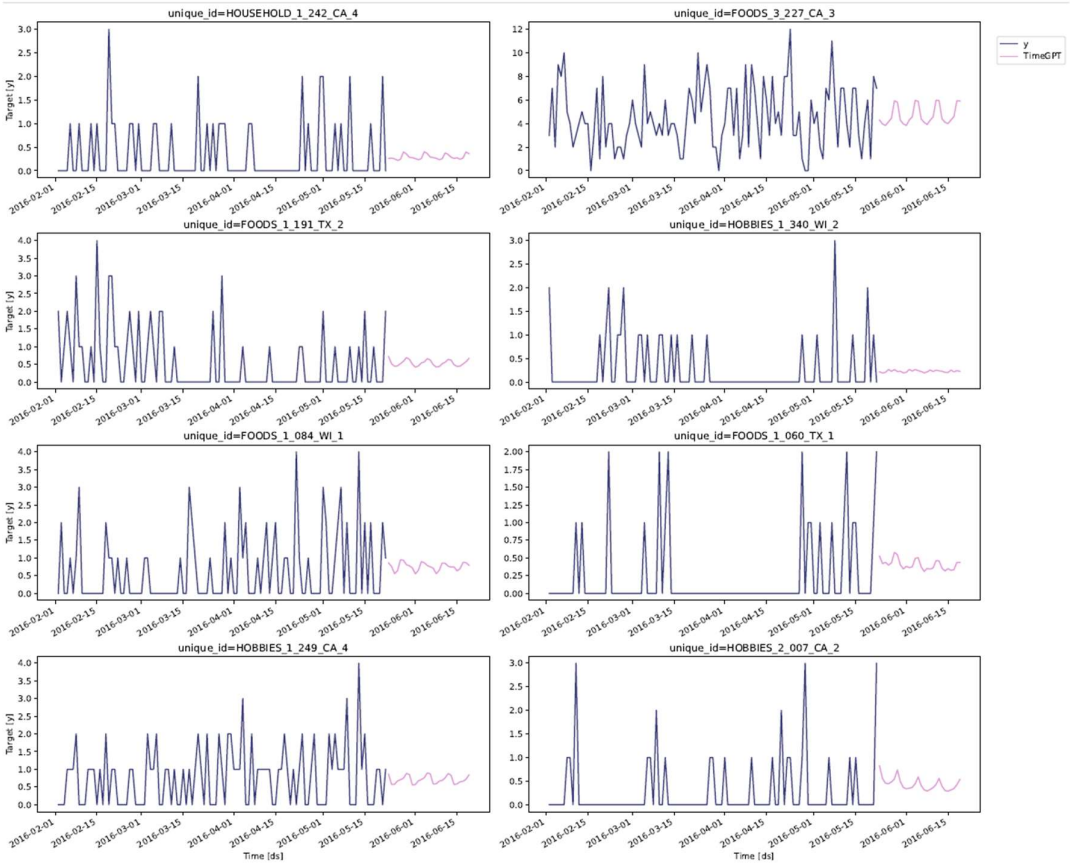


Figure 10 - Point forecasts of retail sales using the TimeGPT model across various product-region combinations.

Figure 10 illustrates the point forecasts generated by the TimeGPT model for various retail products, shown alongside actual sales data over a period. Each subplot represents a unique product-region combination across categories like Household, Foods, and Hobbies.

Observed Sales (Blue Line): The historical sales data are represented by the blue line, covering the period from early February to mid-June 2016.

Point Forecast (Pink Line): The pink line represents the TimeGPT model's point forecast, indicating the model's best estimate for future sales at each time step.

This plot allows for a straightforward comparison between observed sales and the model's predictions, providing insights into TimeGPT's ability to capture demand trends and patterns without accounting for prediction uncertainty.

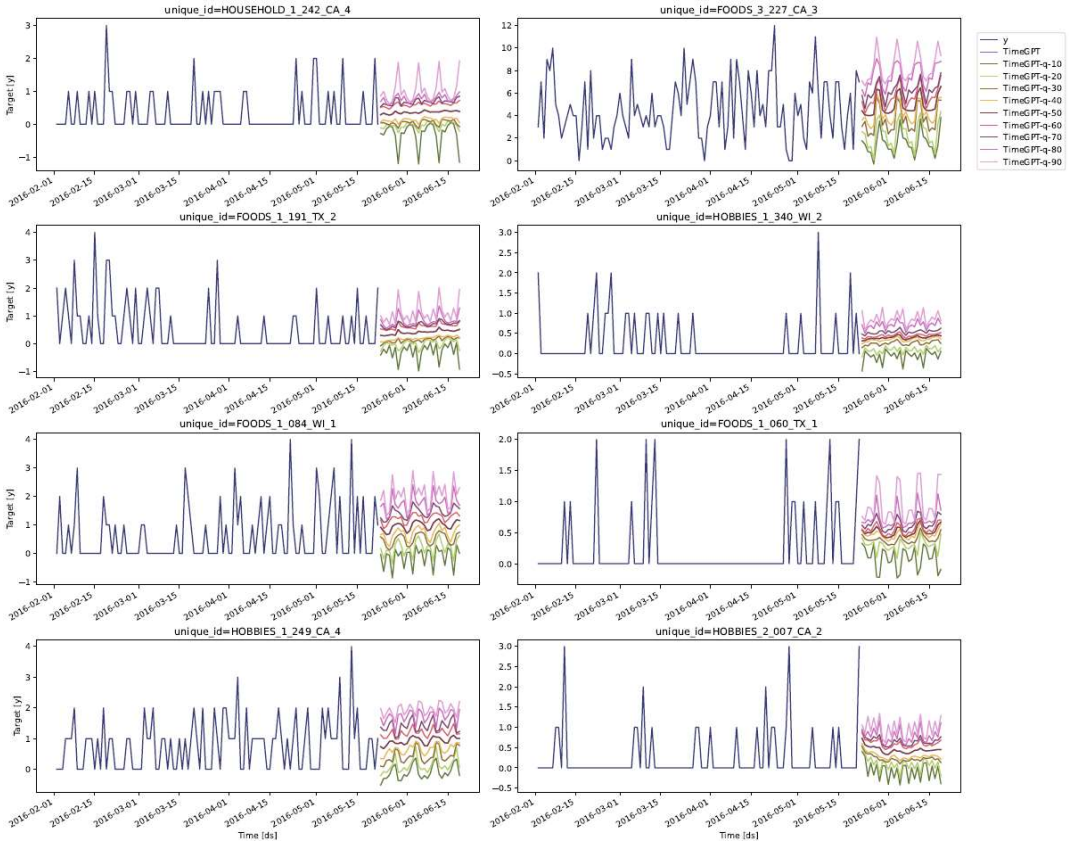


Figure 11 - Probabilistic forecasts of retail sales using the TimeGPT model, illustrating multiple quantiles for future sales projections across various product-region combinations.

Figure 11 shows probabilistic forecasts by the TimeGPT model for the same set of retail products. Here, different quantiles are visualized, highlighting the range of potential future sales outcomes and the model's confidence levels.

Observed Sales (Blue Line): The actual sales values, depicted in blue, serve as a baseline to evaluate forecast performance.

Quantile Forecasts (Colored Lines): Each colored line represents a specific quantile (e.g., 0.1, 0.2, ..., 0.9), capturing the forecasted sales range with different levels of confidence. Lower quantiles represent conservative estimates, while higher quantiles cover more optimistic forecasts.

This probabilistic forecast plot is valuable for managing demand uncertainty, providing retailers with a range of possible sales outcomes that help in preparing for demand surges or troughs. This is particularly helpful during peak seasons, promotional periods, and other high-variability events, supporting more robust inventory and supply chain planning.

4.4.2.4 Summary

Fine-Tuning significantly enhances TimeGPT's performance, making it more adept at both point and probabilistic forecasting. This improvement positions TimeGPT as a valuable tool for retail forecasting, where managing uncertainty and accurately predicting sales volumes are essential for operational efficiency and cost management.

4.4.3 Analysis of Lag-Llama Performance (Zero-Shot and Fine-Tuning Evaluation)

This analysis focuses on two key metrics—MASE (Mean Absolute Scaled Error) and CRPS (Continuous Ranked Probability Score)—to evaluate the Lag-Llama model's performance in both Zero-Shot and Fine-Tuning settings. These metrics provide insights into the model's ability to deliver accurate point forecasts (MASE) and handle probabilistic forecasts (CRPS), which are essential for managing demand uncertainty in retail sales forecasting.

Table 3 - Performance of the Lag-Llama model under different evaluation settings (Zero-Shot and Fine-Tuning), measured by MASE and CRPS.

Model	Evaluation Setting	MASE	CRPS
Lag-Llama	**Zero-Shot**	1.0326	0.5955
	Fine-Tuning	1.0028	0.5715

4.4.3.1 MASE (Mean Absolute Scaled Error) Analysis

- Zero-Shot Evaluation (MASE: 1.0326)

In the Zero-Shot setting, where Lag-Llama is applied directly to the M5 dataset without specific tuning, the model achieves a MASE of 1.0326. A MASE value just above 1 indicates that Lag-Llama performs slightly better than a naive forecast (e.g., assuming tomorrow's sales will be similar to today's).

This result implies that Lag-Llama captures some general patterns in the dataset, such as overall sales trends, but struggles to account for more complex sales fluctuations like seasonal peaks and promotional spikes. These aspects are particularly challenging in retail forecasting, where demand is heavily influenced by external events and promotional periods.

- Fine-Tuning Evaluation (MASE: 1.0028)

After Fine-Tuning, the MASE drops to 1.0028, bringing the model's performance very close to the naive benchmark, with a significant improvement in point forecast accuracy. This reduced MASE value reflects Lag-Llama's enhanced capability to capture seasonal patterns and promotional effects when trained on specific aspects of the dataset.

With this improvement, Lag-Llama becomes better equipped to predict demand surges during holidays or sales events, providing more precise point forecasts. This is especially valuable for retail applications, where accurate sales forecasts enable retailers to optimize inventory levels and plan promotions effectively.

4.4.3.2 CRPS (Continuous Ranked Probability Score) Analysis

- Zero-Shot Evaluation (CRPS: 0.5955)

In the Zero-Shot evaluation, Lag-Llama's CRPS is 0.5955, indicating moderate performance in probabilistic forecasting. CRPS assesses how well the model can predict the range of possible outcomes, which is critical in retail for managing uncertainty, particularly during high-variance periods like promotions or seasonal peaks.

A CRPS of 0.5955 suggests that, without fine-tuning, Lag-Llama has a basic ability to estimate demand variability but may struggle to capture the full scope of uncertainties in sales outcomes, such as unexpected demand spikes or low-demand periods. This limits its effectiveness in retail planning where accurately predicting upper and lower bounds for sales is essential to avoid stockouts or excess inventory.

- Fine-Tuning Evaluation (CRPS: 0.5715)

After fine-tuning, the CRPS improves to 0.5715, showing that Lag-Llama is now more capable of handling probabilistic forecasts and managing uncertainty in sales predictions. A lower CRPS score indicates that the model can better predict sales ranges, making it more useful for inventory management and risk mitigation.

This improvement in CRPS means that Lag-Llama is better able to provide probabilistic forecasts for high-demand events and promotional sales periods, allowing retailers to prepare more accurately for demand fluctuations. A reliable CRPS helps retailers anticipate stock requirements and manage inventory levels more efficiently, minimizing both stockouts and overstocking during volatile periods.

4.4.3.3 Fine-Tuned Lag-Llama Model: Probabilistic and Point Forecasting of Retail Sales

Figure 12 displays the forecasted vs. observed sales generated by the Lag-Llama model after fine-tuning, focusing on retail product categories over time. The Lag-Llama model has been fine-tuned on a specific dataset to better capture complex sales patterns, such as seasonality, demand spikes, and variability due to promotions.

Observed Sales (Blue Line): The blue line in each subplot represents the actual sales over time, spanning from early March to mid-June 2016. This historical data serves as the baseline to assess the model's forecast accuracy.

Point Forecast (Green Line): The central green line within each shaded area represents the model's point forecast—its best estimate of expected future sales. Comparing this line to the actual observed sales provides insights into the model's accuracy after fine-tuning.

Probabilistic Forecast (Green Shaded Area): Surrounding each point forecast, the green shaded area illustrates the probabilistic forecast, which provides a confidence interval or range within which future sales are expected to fall. A wider shaded area suggests higher uncertainty in predictions, often influenced by volatile demand periods such as promotions or seasonal events.

The fine-tuning process allows the Lag-Llama model to better adapt to the specific characteristics of retail sales data, improving its ability to predict both typical sales levels and potential variability. The model's probabilistic forecasting capability, visualized by the shaded uncertainty region, is particularly useful for retailers in planning around demand spikes or seasonal fluctuations, aiding in inventory management and operational planning.

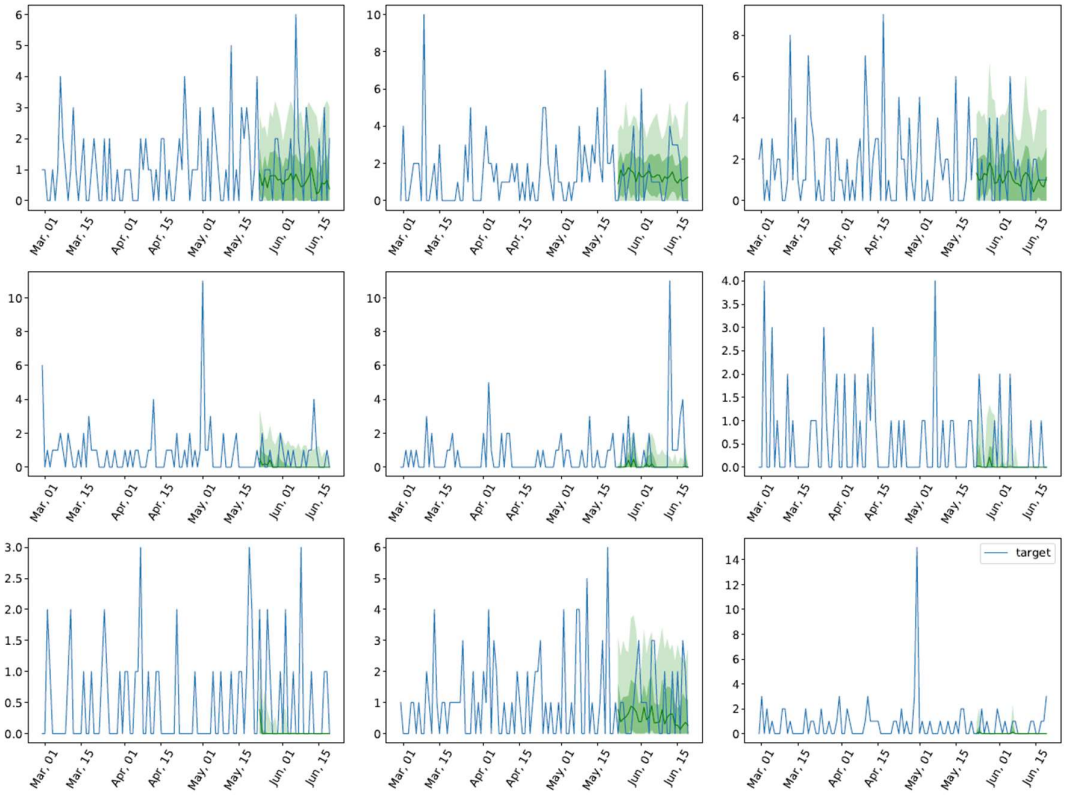


Figure 12 - Forecasted versus observed sales for retail product categories over time, generated by the fine-tuned Lag-Llama model.

4.4.3.4 Summary

Fine-Tuning greatly improves Lag-Llama's performance, making it better at both point and probabilistic forecasting. This enhanced accuracy and ability to capture uncertainty make Lag-Llama a valuable tool for retail forecasting, helping retailers to efficiently manage stock levels, prepare for demand surges, and reduce operational risks associated with inventory mismanagement.

4.4.4 Analysis of Uni2TS Performance (Zero-Shot and Fine-Tuning Evaluation)

This analysis evaluates the Uni2TS model across different model sizes—Small, Base, and Large—using MASE (Mean Absolute Scaled Error) and CRPS (Continuous Ranked Probability Score) as key performance metrics. These metrics allow us to assess each model's point forecast accuracy (MASE) and probabilistic forecasting ability (CRPS), both of which are crucial in retail forecasting where uncertainty and variability are common.

Table 4 - Performance of the UniTS model across different evaluation settings (Zero-Shot and Fine-Tuning) and model sizes (Small, Base, Large), measured by MASE and CRPS.

Model	Evaluation Setting	Model Size	MASE	CRPS
Uni2TS	Zero-Shot	**Small**	1.0321	0.5938
		Base	1.0451	0.5973
		Large	1.0350	0.5901
	Fine-Tuning	**Small**	1.4742	0.5677
		Base	1.4818	0.5660
		Large	1.4571	0.5593

4.4.4.1 Zero-Shot Evaluation

- MASE (Mean Absolute Scaled Error) Analysis

Small Model: With a MASE of 1.0321, the Uni2TS Small model performs slightly above the naive benchmark, suggesting it can capture general sales trends but may struggle with more complex patterns, such as seasonal spikes or promotions. This result shows moderate point forecasting ability in the Zero-Shot setting.

Base Model: The MASE of 1.0451 for the Base model is slightly higher than the Small model, indicating a minor decrease in point forecast accuracy. This may suggest that, without fine-tuning, the Base model size does not offer substantial gains in capturing complex retail sales patterns compared to the smaller model.

Large Model: With a MASE of 1.0350, the Large model performs similarly to the Small model. This indicates that, in Zero-Shot mode, increasing the model size does not necessarily enhance point forecast accuracy for Uni2TS, as the model continues to perform close to the naive benchmark.

- CRPS (Continuous Ranked Probability Score) Analysis

Small Model: The CRPS of 0.5938 for the Small model indicates moderate capability in probabilistic forecasting, with limited effectiveness in handling uncertainty. This is essential in retail, where capturing the range of demand fluctuations (e.g., during promotions) is critical.

Base Model: The CRPS for the Base model is 0.5973, showing a slight degradation in probabilistic forecasting compared to the Small model. This suggests that the Base model, without fine-tuning, may face challenges in handling demand uncertainties effectively.

Large Model: The CRPS improves slightly to 0.5901 in the Large model, indicating a marginally better ability to manage uncertainty in demand. This small improvement suggests that the larger model size helps the model capture a broader range of possible outcomes, which is useful in retail forecasting.

4.4.4.2 Fine-Tuning Evaluation

- MASE (Mean Absolute Scaled Error) Analysis

Small Model: After fine-tuning, the MASE jumps significantly to 1.4742 for the Small model, indicating a notable decline in point forecasting accuracy. This increase suggests potential overfitting, where the model has become too specialized to the specific training data and is unable to generalize effectively.

Base Model: The MASE rises to 1.4818 for the Base model, marking the highest MASE among the three sizes. This shows that the Base model, when fine-tuned, struggles the most with point forecast accuracy, likely due to overfitting or a lack of adaptability to the full data distribution.

Large Model: The MASE for the Large model is 1.4571, slightly better than the other two sizes but still much higher than the Zero-Shot performance. This indicates that fine-tuning does not improve point forecast accuracy and may instead introduce overfitting issues, especially in the smaller model sizes.

- CRPS (Continuous Ranked Probability Score) Analysis

Small Model: The CRPS decreases to 0.5677 for the Small model, showing improved probabilistic forecasting after fine-tuning. This suggests that, despite worsened point forecasting, the model becomes better at handling uncertainties and predicting sales ranges.

Base Model: With a CRPS of 0.5660, the Base model further improves in probabilistic forecasting, becoming more adept at predicting the range of demand fluctuations. This is valuable in retail for managing inventory during unpredictable periods.

Large Model: The CRPS for the Large model is 0.5593, the best among the three sizes. This improvement in CRPS indicates that the fine-tuned Large model excels at handling demand uncertainty, making it more effective for scenarios that require probabilistic forecasting rather than precise point predictions.

4.4.4.3 Point and Probabilistic Forecasting for Retail Sales Using Uni2TS-Large Model

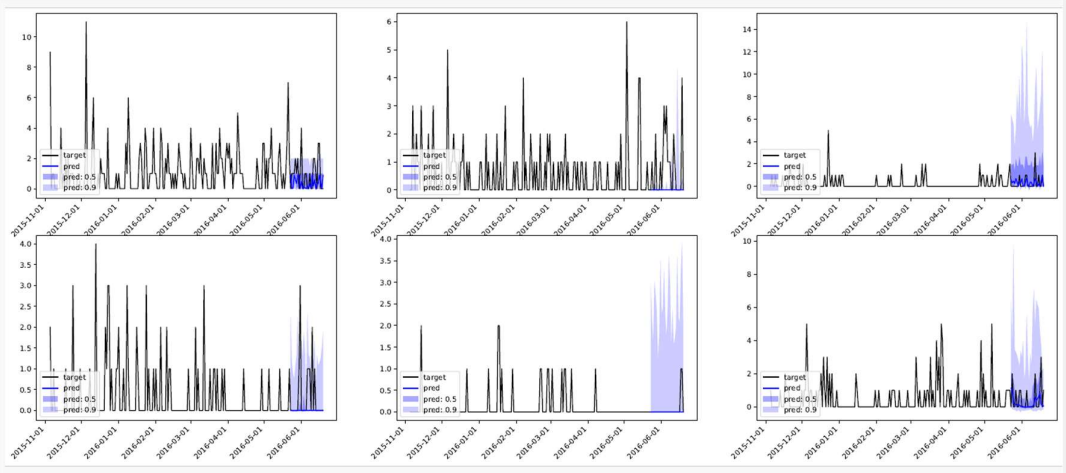


Figure 13 - Point and probabilistic forecasts of retail sales using the Uni2TS-Large model across various product-region combinations.

Figure 13 showcases point and probabilistic forecasts generated by the Uni2TS-Large model for several retail product categories over time. Each subplot represents a unique product and region combination, displaying both observed sales data and the Uni2TS-Large model's predictions.

Observed Sales (Black Line): Historical sales data for each product-region pair, providing a reference for evaluating forecast accuracy.

Point Forecast (Blue Line): The central prediction by the Uni2TS-Large model, representing the expected sales value at each time step.

Prediction Intervals (Shaded Blue Areas):

Dark Blue (0.5 Quantile): Represents the median forecast, giving a mid-point estimation.

Light Blue (0.9 Quantile): Wider shaded areas indicate the model's higher quantile forecasts, showing the range where 90% of expected outcomes are predicted to fall.

This combination of point forecasts and prediction intervals allows for both precise estimates and a probabilistic range, providing valuable insights into demand uncertainty. This is essential for managing inventory and preparing for sales variations due to seasonal trends, promotions, or regional demand fluctuations.

4.4.4.4 Summary

The Uni2TS model shows mixed performance across Small, Base, and Large sizes. In the Zero-Shot setting, it performs only slightly better than a naive forecast in point forecasting, with minimal gains from increasing model size. However, fine-tuning leads to a significant drop in point forecasting accuracy (MASE), likely due to overfitting, while probabilistic forecasting (CRPS) improves notably.

For retail applications, the fine-tuned Large model is the most effective for handling demand uncertainties and predicting a range of possible outcomes, which is essential for inventory management and planning during high-variability periods. However, Uni2TS's point forecasting abilities remain limited, especially after fine-tuning, suggesting that it may be better suited for scenarios where capturing uncertainty is more important than exact point forecasts. Further refinement is needed to enhance its generalization and balance between point and probabilistic forecasting.

4.4.5 Analysis of TimesFM Performance (Zero-Shot and Fine-Tuning Evaluation)

The TimesFM model has been evaluated across Zero-Shot and Fine-Tuning settings using two key metrics: MASE (Mean Absolute Scaled Error) for point forecasting accuracy and CRPS (Continuous Ranked Probability Score) for probabilistic forecasting accuracy. These metrics provide a comprehensive view of TimesFM's ability to generate accurate sales predictions and capture demand uncertainty, which are essential in retail forecasting for inventory management and strategic planning.

Table 5 - Performance of the TimesFM model under different evaluation settings (Zero-Shot and Fine-Tuning), measured by MASE and CRPS.

Model	Evaluation Setting	MASE	CRPS
TimesFM	**Zero-Shot**	0.9971	0.5656
	Fine-Tuning	0.9667	0.5519

4.4.5.1 Zero-Shot Evaluation

- Point Forecast Accuracy (MASE)

MASE: The TimesFM model achieves a MASE of 0.9971 in the Zero-Shot setting, which is below the critical threshold of 1. This indicates that the model performs better than a naive benchmark (e.g., assuming tomorrow's sales will be the same as today's) without any prior fine-tuning on specific data. A MASE below 1 suggests that TimesFM can capture essential patterns and trends in retail data, such as seasonality and basic demand fluctuations, making it a reliable choice even in a Zero-Shot application. Achieving a MASE below 1 in Zero-Shot mode shows that TimesFM has strong generalization capabilities, allowing it to handle typical retail sales trends without specialized training. This is beneficial for real-world applications where models might need to operate on new data without extensive training time, providing retailers with reliable forecasts from the outset.

- Probabilistic Forecasting Accuracy (CRPS)

CRPS: With a CRPS of 0.5656, TimesFM demonstrates moderate effectiveness in probabilistic forecasting in the Zero-Shot scenario. This score indicates that TimesFM can predict not only a single sales forecast but also the range of possible sales outcomes with reasonable accuracy. This ability is critical in retail, where anticipating demand variability (e.g., during promotions or seasonal peaks) can help prevent stockouts or excess inventory. The CRPS score of 0.5656 suggests that TimesFM is relatively capable of handling uncertainty in sales data, even without fine-tuning. This makes the model suitable for retail forecasting applications that require a basic level of probabilistic forecasting to understand demand ranges, which is useful for inventory and risk management.

4.4.5.2 Fine-Tuning Evaluation

- Point Forecast Accuracy (MASE)

MASE: After fine-tuning, the MASE decreases to 0.9667, reflecting an improvement in point forecast accuracy. This reduction from 0.9971 indicates that fine-tuning allows TimesFM to better adapt to specific patterns and nuances in retail sales data, such as regional demand variations, seasonality, and promotional effects. The improvement, though modest, suggests that fine-tuning makes TimesFM more precise in forecasting

daily sales values. The MASE improvement after fine-tuning means that TimesFM becomes even more reliable for generating exact sales forecasts, which is essential for effective inventory management in retail. The ability to fine-tune the model for greater accuracy makes it a valuable tool for retailers who need dependable forecasts to meet demand fluctuations accurately.

- Probabilistic Forecasting Accuracy (CRPS)

CRPS: The CRPS value further improves to 0.5519 after fine-tuning, showing that TimesFM has enhanced its ability to provide reliable probabilistic forecasts. A lower CRPS score means that TimesFM becomes more proficient at managing uncertainties in sales predictions, helping retailers better understand the range of potential sales outcomes and plan for high-variability periods. The reduction in CRPS to 0.5519 reflects TimesFM's improved ability to capture uncertainty, which is particularly useful during promotional events or seasonal peaks when sales patterns can be volatile. The model's enhanced probabilistic forecasting allows retailers to make informed decisions around stock levels, minimizing the risk of overstocking or stockouts.

4.4.5.3 Fine-Tuned TimesFM Model: Probabilistic and Point Forecasting of Retail Sales

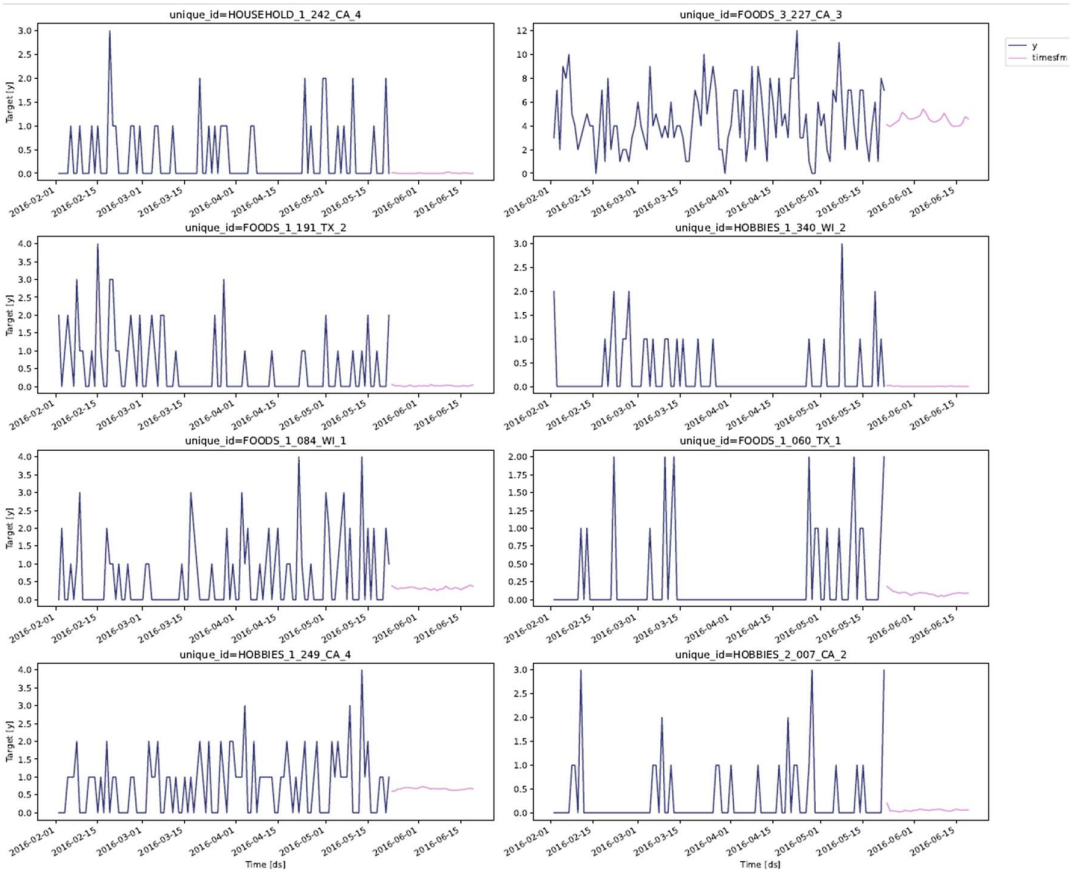


Figure 14 - Observed versus point forecasts for retail products over time, generated by the TimesFM model across various categories and regions.

Figure 14 displays the observed vs. point forecasts generated by the TimesFM model for various retail products over a given period. Each subplot corresponds to a specific product and region, covering categories such as Household, Foods, and Hobbies.

Observed Sales (Blue Line): The historical sales data, shown as a continuous blue line, represents actual recorded sales from early March to mid-June 2016.

Point Forecast (Pink Line): The pink line at the end of each subplot represents the TimesFM model's point forecast, which is its best estimate of future sales for each time step.

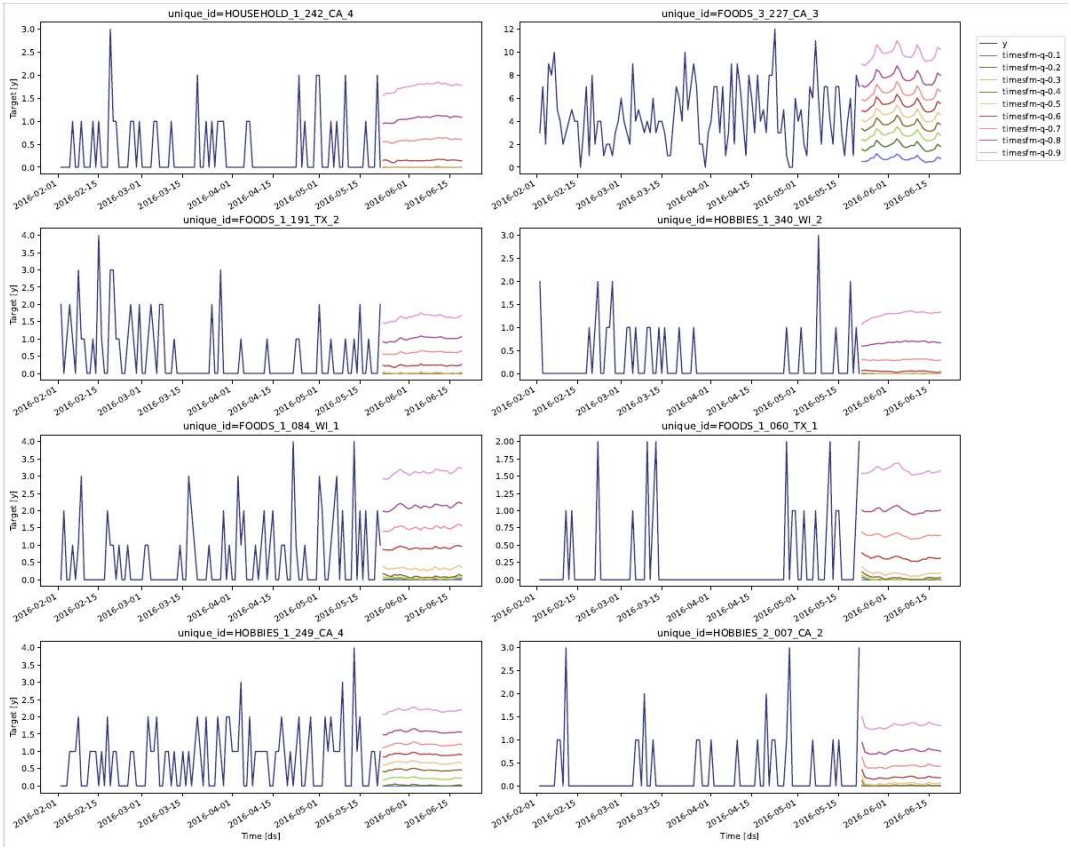


Figure 15 - Probabilistic forecasts of retail sales using the TimesFM model, showing confidence intervals for future sales across various products and regions.

Figure 15 demonstrates probabilistic forecasts by the TimesFM model for the same set of retail products, visualizing the range of possible future sales outcomes at different confidence levels.

Observed Sales (Blue Line): As in the first plot, the blue line represents actual historical sales data, providing a reference for model performance.

Quantile Forecasts (Colored Lines): Each colored line represents a different quantile forecast (e.g., 0.1, 0.2, …, 0.9), ranging from lower (0.1 quantile) to upper sales bounds (0.9 quantile). Higher quantiles represent higher confidence levels in forecasting.

By visualizing multiple quantiles, this plot captures the uncertainty in future sales predictions, with higher quantiles representing possible upper bounds for demand. This probabilistic approach is particularly useful for inventory management and planning, allowing retailers to prepare for a range of demand scenarios, from conservative to more optimistic estimates.

This visualization allows for direct comparison between historical sales and the model's single-point predictions, providing insights into TimesFM's ability to capture trends and seasonality in retail sales.

4.4.5.4 Summary

The TimesFM model stands out as a highly effective tool for retail demand forecasting, with strong performance in both Zero-Shot and Fine-Tuning settings. In the Zero-Shot scenario, TimesFM already achieves a MASE below 1 and a competitive CRPS, indicating that it can handle typical sales patterns and demand uncertainty without needing extensive training. This makes it ideal for applications requiring immediate deployment and reliable generalization.

After fine-tuning, TimesFM further enhances its point forecast accuracy and probabilistic forecasting capability. The model's ability to adapt to specific data nuances allows it to capture demand spikes and sales fluctuations more accurately, making it even more suitable for real-world retail applications. The consistent improvement in MASE and CRPS scores shows that TimesFM is effective not only in generating precise sales forecasts but also in predicting a range of possible outcomes, essential for inventory planning and risk management.

In summary, TimesFM is a versatile and robust model for retail forecasting. Its strong performance in Zero-Shot mode and its further improvements after fine-tuning make it valuable for retailers looking to optimize inventory levels, prepare for seasonal peaks, and manage promotional demand with greater confidence.

4.5 Detailed Analysis of Results: Model Comparison (Zero-Shot and Fine-Tuning)

Time series forecasting requires models capable of capturing complex temporal patterns, such as trends and seasonality, while also dealing with inherent uncertainties. The models were tested under two scenarios: Zero-Shot (where models are applied directly without additional training on the test data) and Fine-Tuning (where models are adjusted to the specific dataset). The metrics used—MASE (Mean Absolute Scaled Error) and CRPS (Continuous Ranked Probability Score)—are widely accepted in the forecasting community, measuring, respectively, the accuracy of point forecasts and the quality of estimated probabilistic distributions.

4.5.1 Zero-Shot Performance

In the Zero-Shot scenario, where models are tested without any specific training on the dataset, the capacity for generalization emerges as the key factor. Models that perform well here show great potential for real-world applications where specific training data might be limited, or there is no time to perform fine-tuning.

4.5.1.1 Chronos Family (Tiny, Mini, Small, Base, Large)

The Chronos family follows a pattern where increasing the model size (and thus its capacity to learn more complex representations) leads to a gradual reduction in error, though not always in a linear fashion:

Chronos-Tiny: With a MASE of 1.0920 and CRPS of 0.5973, this is the smallest model in the Chronos family, implying that its ability to capture patterns is limited, especially in terms of MASE, which is more sensitive to point forecast errors. This performance is expected for a model of this size, where computational efficiency is often favored at the expense of slight precision losses.

Chronos-Mini: Performance improves slightly with a MASE of 1.1040 and CRPS of 0.5858, reflecting the model's better ability to capture complex patterns compared to Chronos-Tiny. Curiously, the MASE increases slightly, but the decrease in CRPS suggests that Chronos-Mini is better at capturing uncertainty in the forecasts.

Chronos-Small: With a MASE of 1.0940 and CRPS of 0.5817, this model approaches the higher end of the Chronos range, showing a reasonable balance between complexity and accuracy. The CRPS, which measures the quality of probabilistic forecasting, indicates that this model is improving its estimation of variability.

Chronos-Base and **Chronos-Large**: As models increase in size, we expected a substantial improvement in accuracy. Chronos-Base presents a MASE of 1.1072 and CRPS of 0.5726, while Chronos-Large offers MASE of 1.1098 and CRPS of 0.5706. These models, though competitive, do not outperform some of the other models presented in your results. Curiously, there's a stagnation in performance gains, suggesting that increasing model size beyond a certain point might not yield proportional improvements without more refined fine-tuning.

Interpretation:

The performance of the Chronos models follows an expected trend, where larger models capture more temporal characteristics. However, the difference in performance between Chronos-Small and Chronos-Large is not as significant as anticipated. This may suggest that, without specific adjustment for the dataset, the larger models' generalization ability isn't as superior as their parameter size would suggest. Moreover, a potential saturation in performance gains could be related to the Chronos architecture, which without additional adjustments, struggles to capture more specific nuances in the data.

4.5.1.2 TimeGPT and Lag-Llama

These two models exhibit distinct behaviors, with Lag-Llama standing out.

TimeGPT: With a MASE of 1.0842 and CRPS of 0.6080, TimeGPT performs reasonably well, but the higher CRPS indicates a slight limitation in estimating uncertainty. Compared to Chronos models, TimeGPT seems to be more balanced, achieving better MASE performance, though it doesn’t stand out as the best in either metric.

Lag-Llama: This model is impressive, with a MASE of 1.0326 and CRPS of 0.5955, making it one of the best in the Zero-Shot scenario. Lag-Llama consistently outperforms almost all Chronos models in both MASE and CRPS. The MASE performance suggests that this model has excellent generalization ability, capturing general temporal series patterns well, while maintaining a competitive CRPS. These results position Lag-Llama as a strong choice for scenarios where generalization is critical in unseen forecasting tasks.

Interpretation:

Lag-Llama has remarkable performance in Zero-Shot, suggesting that its architecture is particularly well-suited for forecasting tasks where the data is unknown. Its ability to maintain a low MASE, coupled with a competitive CRPS, reflects the model’s ability to accurately predict both central values and variability of time series.

4.5.1.3 Uni2TS (Small, Base, Large)

The Uni2TS models present interesting results, as they don’t follow the expected pattern of improvement with increasing model size.

Uni2TS-Small: With a MASE of 1.0321 and CRPS of 0.5938, this model is very close to Lag-Llama in terms of MASE, suggesting strong forecasting capacity without additional adjustments. However, the slightly higher CRPS indicates that its probabilistic forecasts aren't as well-calibrated.

Uni2TS-Base and **Uni2TS-Large:** While these models show slight improvements in CRPS, the MASE gains are marginal, with Uni2TS-Base recording a MASE of 1.0451 and Large with 1.0350. These results suggest that Uni2TS may have limitations in scalability in Zero-Shot, meaning that increasing model complexity does not necessarily result in significant improvements, possibly due to overfitting during initial training phases.

Interpretation:

Uni2TS-Small is notably competitive in Zero-Shot, suggesting that this model, even with a smaller architecture, can effectively capture temporal patterns. However, the performance of the larger models does not improve significantly, indicating that Uni2TS may be more efficient with smaller capacity models in Zero-Shot scenarios, where simplicity may be an advantage.

4.5.1.4 TimesFM

TimesFM stands out as the best-performing model in Zero-Shot. With a MASE of 0.9971 and CRPS of 0.5656, it is the only model achieving a MASE below 1, which indicates a highly precise forecasting ability, even without specific training on the test data. Additionally, its lower CRPS suggests that it also captures uncertainty in the data well, making it a highly versatile model.

Interpretation:

TimesFM's ability to generalize well in Zero-Shot is remarkable. This result suggests that the model is particularly efficient in capturing general time series patterns and predicting reliable distributions, making it a robust choice for a wide range of time series forecasting tasks.

4.5.2 Fine-Tuning Performance

In the Fine-Tuning scenario, models are adjusted with specific dataset data, allowing us to assess their ability to adapt and refine for improved performance.

4.5.2.1 TimeGPT and Lag-Llama

After fine-tuning, the results for TimeGPT and Lag-Llama improve significantly:

TimeGPT: Performance improves with a MASE of 1.0163 and CRPS of 0.5794, showing that fine-tuning brings considerable accuracy gains. However, this model is still outperformed by TimesFM.

Lag-Llama: It continues to show strong performance after fine-tuning, with a MASE of 1.0028 and CRPS of 0.5715. Fine-tuning improves Lag-Llama significantly, bringing it closer to the best results, making it a viable choice when fine-tuning is allowed.

Interpretation:

Both models clearly benefit from fine-tuning, with significant improvements in both MASE and CRPS. However, despite these improvements, TimesFM maintains its lead, suggesting that even with specific data adjustments, these models cannot quite reach the top performance.

4.5.2.2 Uni2TS (Small, Base, Large)

The fine-tuning results for the Uni2TS models reveal an interesting trend:

Uni2TS-Small, which had promising performance in Zero-Shot, now registers a very high MASE of 1.4742, indicating that fine-tuning may have led to severe overfitting or sub-optimal adjustment to the specific data. The CRPS of 0.5677 is competitive, suggesting that the model still captures uncertainty well but fails to predict the central value accurately.

Uni2TS-Base and **Large** show similar patterns, with MASEs of 1.4818 and 1.4571, respectively, and CRPS of 0.5660 and 0.5593. These results suggest that Uni2TS struggles with fine-tuning, possibly due to an architecture that does not adapt well to new data after initial training.

Interpretation:

The fine-tuning results for Uni2TS are concerning, especially as they indicate that the model is not effectively adjusting to the test data. The high MASE suggests that the fine-tuning process is not successful in training the model to capture new patterns in the data, resulting in inaccurate forecasts.

4.5.2.3 TimesFM

TimesFM continues to demonstrate its versatility and robustness with fine-tuning, further improving its performance, with a MASE of 0.9667 and CRPS of 0.5519. These values are significantly better than the other models, confirming the superiority of this model in both scenarios.

Interpretation:

Not only is TimesFM the best model in the Zero-Shot scenario, but it also dominates in fine-tuning, demonstrating an impressive ability to adapt to specific dataset characteristics without sacrificing accuracy or probabilistic distribution quality.

4.5.3 General Comparison and Final Considerations

When comparing the results, it becomes clear that different model architectures and forecasting strategies lead to substantial variations in performance, particularly when considering the MASE and CRPS metrics. These metrics reflect the accuracy of future value predictions (MASE) and the quality of probabilistic forecasts (CRPS), both crucial for reliable time series predictions with uncertainty.

4.5.3.1 Zero-Shot: The Importance of Generalization

In the Zero-Shot scenario, where models are applied directly to new data without specific training, the ability to generalize is the most critical factor. Models that perform well here have great potential for practical applications in environments where specific training data is limited, or where the time needed for fine-tuning is not available.

TimesFM clearly stands out, with the lowest MASE (0.9971) and CRPS (0.5656). This suggests an incredible ability to capture general time series patterns and provide reliable forecasts without needing adjustments. This performance makes it a robust choice in practical scenarios, such as short-term forecasts in business environments or automated forecasting systems, where specific data collection and training may be limited.

Lag-Llama is also impressive, especially in MASE (1.0326), coming close to TimesFM, though falling slightly short in CRPS (0.5955). This suggests that Lag-Llama can capture central time series trends well but does not handle uncertainty as effectively as TimesFM. This balance between forecast accuracy and uncertainty capture makes Lag-Llama a solid choice, particularly in contexts where data variability is less critical.

While the Chronos models show a trend of improvement with increasing model size (from Tiny to Large), the improvement is not linear. In fact, models such as Chronos-Large (MASE 1.1098, CRPS 0.5706) do not outperform smaller competitors like Lag-Llama or Uni2TS-Small. This raises questions about the efficacy of the Chronos architecture in Zero-Shot scenarios, where simplicity may indeed be beneficial. Chronos-Small and Chronos-Mini, for example, perform solidly, but there is a saturation in performance gains as the model grows, indicating that increasing complexity does not always yield proportional improvements.

The Uni2TS-Small model offers an excellent balance between simplicity and performance with a MASE of 1.0321, close to Lag-Llama. However, the larger models (Uni2TS-Base and Large) do not improve significantly, suggesting that Uni2TS may be more efficient with smaller complexity models in Zero-Shot. This behavior suggests that Uni2TS may suffer from overfitting in its initial training stages, resulting in a lack of adaptation to general time series behavior.

Reflection on Zero-Shot:

Zero-Shot performance is particularly relevant for environments with dynamic or heterogeneous datasets, where training models for each scenario is impractical. TimesFM proves to be the most robust model, standing out for its ability to generalize without compromising accuracy. Lag-Llama offers a strong alternative, especially for forecasts that do not require a high focus on uncertainty. Meanwhile, the performance of Chronos and Uni2TS suggests that their efficiency may be limited in scenarios without dedicated training, and the larger models in these families may suffer from overfitting or inefficiency when used in Zero-Shot.

4.5.3.2 Fine-Tuning: When Personalization Pays Off

Fine-Tuning allows models to adjust their parameters based on specific problem data, which typically improves performance on seen datasets. However, the results reveal that not all models respond equally well to this process.

TimesFM, once again, remains the standout performer, with the lowest MASE (0.9667) and CRPS (0.5519). Its ability to adapt to specific data, combined with its already strong Zero-Shot performance, shows that it is an extremely versatile model. This makes TimesFM an obvious choice for cases where training data is available and personalization

is allowed, such as demand forecasting in supply chains or inventory management, where historical data can be leveraged for fine-tuning.

Although it improves with fine-tuning (MASE 1.0028, CRPS 0.5715), Lag-Llama fails to surpass TimesFM. This result is significant because it indicates that, despite being a strong competitor in Zero-Shot, the model does not have the same ability to adapt to specific data during fine-tuning, limiting its potential in environments that require intensive customization. However, its overall robustness suggests that it is still an excellent choice for models that require less retraining and more robustness.

TimeGPT model improves significantly with fine-tuning (MASE 1.0163, CRPS 0.5794), but remains behind the leaders. TimeGPT seems to benefit from fine-tuning, but not as expressively as would be desirable for scenarios that require highly personalized predictions. Nevertheless, TimeGPT presents itself as a viable option when seeking a balance between simplicity and adaptability.

The Uni2TS series suffers a noticeable degradation in fine-tuning. With alarmingly high MASEs (1.4742 for Small and 1.4818 for Base), it becomes clear that the fine-tuning process is generating overfitting or suboptimal adjustments during training. This failure may be due to the Uni2TS model architecture, which seems to struggle in adjusting appropriately to specific dataset patterns. Although CRPS remains competitive, suggesting the model still captures uncertainty well, the high MASE raises serious concerns about the applicability of Uni2TS in contexts that require fine-tuning.

Reflection on Fine-Tuning:

In the Fine-Tuning scenario, the adaptability of models to dataset-specific characteristics is crucial. TimesFM dominates this scenario, showing that its architecture can capture both general and specific data patterns with the same degree of accuracy. Lag-Llama and TimeGPT improve but fail to reach TimesFM's level, indicating that while useful, they may not be the best choice for applications that demand high personalization. Uni2TS results suggest clear limitations, especially in datasets that require fine-tuning, raising questions about its ability to generalize after initial training.

4.5.4 Conclusion

From the detailed analysis of the results, several important conclusions emerge about the effectiveness of different models in time series forecasting scenarios. It's crucial to

highlight the strengths and weaknesses of each approach, which can help guide practical decisions for various applications.

4.5.4.1 Clear Dominance of TimesFM

TimesFM stands out as the most robust and versatile model in both Zero-Shot and Fine-Tuning scenarios. Its ability to generalize well to new data without requiring fine-tuning, coupled with its superior performance when adjusted to specific datasets, makes it the most reliable and safe choice. This model is ideal for a wide range of applications, from short-term forecasting in financial markets to automated recommendation systems handling temporal data.

4.5.4.2 Lag-Llama: Strong in Generalization, Less Effective in Fine-Tuning

Lag-Llama has proven to be a formidable contender, particularly in the Zero-Shot scenario, where it almost rivaled TimesFM in terms of MASE. However, its performance in Fine-Tuning, while good, was not sufficient to place it at the top. This model is highly recommended for scenarios where strong generalization is needed, and uncertainty is not the primary focus.

4.5.4.3 Chronos: Potential Yet Unrealized

The Chronos series, while showing a trend of improvement with increasing model size, was unable to compete with the leaders, such as TimesFM and Lag-Llama. This may reflect limitations in the architecture or training process. For specific applications, Chronos-Small and Mini may still be good options, but the results suggest that significant optimizations are needed, particularly in terms of fine-tuning capacity and efficient generalization.

4.5.4.4 Uni2TS: Problems in Fine-Tuning

The behavior of Uni2TS raises serious concerns. The initial reasonable performance in Zero-Shot of Uni2TS-Small suggests that the model has potential, but the fine-tuning performance (with very high MASEs) reveals structural problems in the adjustment process, making it a risky choice for applications that depend on fine-tuning.

4.5.4.5 Final Recommendations

For high-precision forecasting in dynamic and varied time series data, TimesFM is the clear choice in both Zero-Shot and Fine-Tuning scenarios.

Lag-Llama is a strong alternative in situations where robust generalization is needed without much reliance on fine-tuning.

Chronos and Uni2TS require additional optimizations, with Chronos being more suitable for scenarios with limited computational resources and Uni2TS needing substantial improvements to perform adequately in fine-tuning.

5 Conclusion

This chapter summarizes the key findings from applying five forecasting models to the M5 dataset, which focuses on retail sales data. It discusses the objectives achieved through the analysis, the limitations encountered, and possible directions for future work.

Accomplished Goals

The main goal of this study was to evaluate the performance of five foundation models—TimesFM, Lag-Llama, Chronos, TimeGPT, and Uni2TS—in predicting retail sales using the M5 dataset. These models were tested under both Zero-Shot and Fine-Tuning scenarios to assess their ability to generalize and adapt to complex retail data.

The following key objectives were successfully met:

G1 – Comprehensive Comparison of Models: The analysis provided valuable insights into the relative performance of the models. TimesFM consistently outperformed the other models in terms of both accuracy and generalization, especially in Zero-Shot scenarios, where minimal training was needed. Its ability to deliver accurate forecasts without extensive fine-tuning highlights its practical utility in real-world retail applications. Lag-Llama also performed strongly, particularly in Zero-Shot scenarios, making it a viable option for rapid retail forecasting when data availability is limited or immediate results are needed.

G2 – Assessment of Model Adaptability and Scalability: The study showed how the models adapted to the complex and variable nature of the retail sales data in the M5 dataset. Chronos and TimeGPT demonstrated solid performance in Fine-Tuning scenarios, where the models were allowed to train more extensively on the dataset, thus improving their ability to handle nuances in sales patterns. Uni2TS, while effective in certain cases, showed limitations in generalizing across diverse product categories, especially when faced with imbalanced data. However, the insights from this model helped to identify key factors that need to be considered in future model development for retail forecasting.

G3 – Analysis of Retail-Specific Challenges: By applying these models to the M5 dataset, the study explored the challenges inherent to retail forecasting, such as the presence of high variability in sales patterns across products and stores. The results showed that models like TimesFM and Lag-Llama could effectively manage these complexities,

maintaining stable performance across different data distributions. This is particularly relevant for retailers who deal with fluctuating demand and require models that can generalize well to new or unseen data.

Overall, the analysis successfully demonstrated the practical applicability of these forecasting models for retail sales prediction, particularly TimesFM, which emerged as the most reliable and efficient model for achieving high accuracy with minimal processing overhead.

Limitations and Future Work

Despite the promising results achieved in this study, several limitations were identified that highlight potential areas for improvement in future research. These include methodological enhancements, expansion of the scope, and real-world applicability that could strengthen the robustness and impact of the findings.

Lack of Cross-Fold Validation: The current study used a single train-test split to evaluate model performance. While this approach provided initial insights, incorporating cross-fold validation in future research would offer a more robust evaluation of how well these models generalize across different subsets of the data. This is particularly important for retail datasets like M5, where sales can exhibit high variability due to seasonal and regional factors. Implementing such techniques would enhance the reliability and consistency of model performance assessments.

Exploring Multivariate Approaches: Although this research has predominantly focused on univariate time series, future investigations could explore multivariate approaches, leveraging the ability of models to capture interrelationships between variables such as prices, promotions, and external factors. Multivariate analysis can provide even more robust and detailed forecasts, enabling a broader understanding of demand patterns and their interdependencies in the retail sector. By incorporating additional dimensions, these models could identify complex relationships that contribute to more accurate and actionable insights.

Exploring Additional Models and Approaches: While this study focused on five key models—Chronos, TimeGPT, Lag-Llama, TimesFM, and Uni2TS—future research could extend the analysis by incorporating additional time series forecasting models, such as Prophet or DeepAR. These models could potentially provide further improvements in

accuracy or efficiency when applied to retail data. Moreover, ensemble learning techniques could be explored to combine the strengths of different models, delivering more robust and reliable forecasting performance for diverse retail scenarios.

Scaling to More Complex Retail Networks: The M5 dataset offers a detailed examination of retail sales for a specific subset of products and locations. However, to understand the scalability of these models, future studies should consider applying them to larger and more complex datasets. This includes retail operations with broader product ranges, varying store formats, and global retail networks. Such research would help determine the adaptability of the models to diverse retail environments and challenges.

Real-World Implementation: While this study was conducted in a simulated environment, certain operational challenges, such as data integration, latency, and real-time forecasting, were not fully addressed. Future research should prioritize pilot implementations within real-world retail businesses to assess the feasibility of deploying these models at scale. These studies could provide valuable insights into practical challenges, such as handling large-scale data pipelines and ensuring real-time decision-making capabilities.

Broader Retail Applications Beyond Sales Forecasting: The primary focus of this study was on sales forecasting, but future investigations could explore the application of these models to other critical areas of retail operations, such as inventory management, demand planning, and promotion effectiveness. Expanding the use cases would offer a more comprehensive understanding of how these models can support various aspects of retail decision-making, ultimately driving operational efficiency and profitability.

These directions for future research aim to build on the findings of this study, leveraging its strengths while addressing its limitations. By exploring these avenues, future work could enhance the practical utility, scalability, and versatility of transfer learning models in retail forecasting and operations.

Final Considerations

This study provided a detailed examination of five foundation models applied to the M5 retail dataset. The results demonstrated the models' strengths in terms of accuracy, generalization, and efficiency, with TimesFM consistently emerging as the top performer. Lag-Llama also showed strong potential for scenarios requiring quick and accurate forecasts with minimal training.

The findings of this analysis underscore the importance of selecting models that balance forecasting performance with operational feasibility in retail settings. Models like TimesFM provide a strong foundation for improving retail decision-making, particularly in forecasting sales trends and optimizing stock management.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., ... & Wang, Y. (2024). Chronos: Learning the Language of Time Series. *Proceedings of the 40th International Conference on Machine Learning (ICML 2024)*, 500-515.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *Proceedings of the International Conference on Learning Representations (ICLR 2015)*.

Box, G. E. P., & Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C, Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. (2020). *Language Models are Few-Shot Learners*. *arXiv preprint arXiv:2005.14165*.

Caruana, R. (1997). *Multitask Learning*. Machine Learning, 28, 41–75.

Chen, M., Xu, Z., Zeng, A., & Xu, Q. (2023). *FrAug: Frequency Domain Augmentation for Time Series Forecasting*. *arXiv preprint arXiv:2301.10096*.

Covington, P., Adams, J., Sargin, E. (2016). *Deep Neural Networks for YouTube Recommendations*. RecSys '16: Proceedings of the 10th ACM Conference on Recommender Systems. 191-198.

Das, A., Li, X., Kumar, R., Banerjee, A., & Rao, S. (2024). *TimesFM: A decoder-only foundation model for time-series forecasting*. Proceedings of NeurIPS 2024.

Devlin, J., Chang, M-W., Lee, K., Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Dooley, R., Oreshkin, B., & Martin, R. (2023). ForecastPFN: Zero-shot Time Series Forecasting. *Advances in Neural Information Processing Systems (NeurIPS 2023)*.

Garza, A., Challu, C., Mergenthaler-Canseco, M. (2023). *TimeGPT-1*. *arXiv preprint arXiv:2310.03589*

Gasthaus, J.; Benidis, K.; Wang, Y.; Rangapuram, S.S.; Salinas, D.; Flunkert, V.; Januschowski, T. (2019). Probabilistic Forecasting with Spline Quantile Function RNNs. In Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, Naha, Okinawa, Japan; PMLR, 89, 1901–1910.

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Gneiting, T., & Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359-378.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice (2nd ed.). OTexts.

Hyndman, R.J., & Koehler, A.B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679-688.

Jeong, C. S. (2023). Fine-Tuning and Utilization Methods of Domain-Specific LLMs. SAMSUNG SDS, arXiv preprint.

Krizhevsky, A., Sutskever, I., Hinton, G. (2012). *ImageNet Classification with Deep Convolutional Neural Networks*. Advances in Neural Information Processing Systems 25.

Lalor, J. P., Wu, H., & Yu, H. (2017). Improving Machine Learning Ability with Fine-Tuning. University of Massachusetts, arXiv preprint.

Lim B., Arık S.O., Loeff N., Pfister T. (2021). *Temporal fusion transformers for interpretable multi-horizon time series forecasting*. International Journal of Forecasting, 37(4), 1748-1764.

Luong, T., Pham, H., Manning, C. (2015). Effective Approaches to Attention-based Neural Machine Translation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 1412-1421.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1), 54–74.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and Applications* (3rd ed.). John Wiley & Sons.

Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. *International Conference on Learning Representations (ICLR)*.

Oliveira, J.M.; Ramos, P. (2024). Evaluating the Effectiveness of Time Series Transformers for Demand Forecasting in Retail. Mathematics, 12, 2728. https://doi.org/10.3390/math12172728.

Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2020). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *International Conference on Learning Representations (ICLR)*.

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359.

Press, O., & Wolf, L. (2016). Using the output embedding to improve language models. *arXiv preprint arXiv:1608.05859*.

Rabanser, S., Shchur, O., & Günnemann, S. (2020). Quantifying data characteristics for time series forecasting. *Proceedings of the 26th ACM SIGKDD Conference*, 944-953.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1-67.

Raj, M., VM, K., Warrier, H., & Gupta, Y. (2024). Fine-Tuning Large Language Models for Enterprise: Practical Guidelines and Recommendations. HCLTech, arXiv preprint.

Ramos, P., Santos, N., & Rebelo, R. (2015). Performance of state space and ARIMA models for consumer retail sales forecasting. *Robotics and Computer-Integrated Manufacturing*, 34, 151–163.

Rasul, K., Ashok, A., Williams, A. R., Khorasani, A., Adamopoulos, G., Bhagwatkar, R., Biloš, M., Ghonia, H., Hassen, N., & Schneider, A. (2023). Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting. *NeurIPS 2023*.

Salinas, D., Bohlke-Schneider, M., Callot, L., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3), 1181-1191.

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3), 1181-1191.

Shafer, G., Vovk, V. (2008). *A Tutorial on Conformal Prediction*. Journal of Machine Learning Research 9, 371-421.

Shchur, O.; Turkmen, C.; Erickson, N.; Shen, H.; Shirkov, A.; Hu, T.; Wang, Y. (2023). AutoGluon-TimeSeries: AutoML for Probabilistic Time Series Forecasting. In Proceedings of the International Conference on Automated Machine Learning, Potsdam/Berlin, Germany.

Student. (1908). The probable error of a mean. *Biometrika*, 6(1), 1-25.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., & Liu, Y. (2024). Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568, 127063.

Su, J., Lu, Y., Pan, S., Wen, B., & Liu, Y. (2021). RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv preprint arXiv:2104.09864*.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 3104-3112.

Taylor, S. J., Letham, B. (2018). *Forecasting at Scale*. The American Statistician, 72(1), 37–45.

Tian, K., Mitchell, E., Yao, H., Manning, C. D., & Finn, C. (2023). Fine-Tuning Language Models for Factuality. Stanford University, arXiv preprint.

Touvron, H., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Torgo, L., & Gama, J. (1997). Regression by classification. *Proceedings of the Portuguese Conference on Artificial Intelligence (EPIA)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, 5998-6008.

Woo, G., Liu, C., Kumar, A., & Sahoo, D. (2023). CoST: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. *International Conference on Learning Representations (ICLR)*.

Woo, G., Liu, C., Zhang, J., Kumar, A., & Sahoo, D. (2024). MOIRAI: Unified training of universal time series forecasting transformers. *NeurIPS 2024*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhang, B., & Sennrich, R. (2019). Root Mean Square Layer Normalization. *Advances in Neural Information Processing Systems*, 32.

Zhang, Y., Sun, X., Ren, X., & Tan, X. (2023). Adapting Language Models with Direct Preference Optimization for Factual Generation. Transactions of the Association for Computational Linguistics, 11, 78-94.

Zhou, Z., Siddiquee, M., Tajbakhsh, N., Liang, J. (2018). *UNet++: A Nested U-Net Architecture for Medical Image Segmentation*. *arXiv preprint arXiv:1807.10165.*