JOHANNES KEPLER
UNIVERSITY LINZ

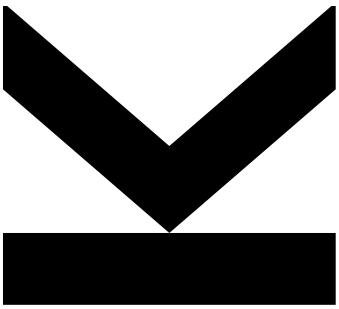
Submitted by
B.Sc. Luca Della Mura

Submitted at
Institute for Business Informatics - Data & Knowledge Engineering

Supervisor
Assoz.-Prof. Mag. Dr. Christoph Schütz

May 2026

A Comparative Study of Foundation Models and Classical Methods for Retail Time Series Forecasting



Master Thesis

to obtain the academic degree of

Master of Science

in the Master's Program

Business Informatics

JOHANNES KEPLER
UNIVERSITY LINZ
Altenbergerstraße 69
4040 Linz, Österreich
www.jku.at

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Professor Dr. Christoph Schütz, for his invaluable guidance, insightful feedback, and continuous support throughout the course of this thesis.

I am also grateful to my colleagues at SCCH for the stimulating discussions and collaborative spirit. Furthermore, I would like to thank my family and friends for their unwavering encouragement and patience during this time.

My sincere thanks also go to the team at PAS Prüfungsservice and to Margit Brandl for their valuable behind-the-scenes support in handling the administrative aspects of this thesis, without which none of this would have been possible.

The research presented in this master's thesis was partly funded by the Federal Ministry for Innovation, Mobility and Infrastructure (BMIMI), the Federal Ministry for Economy, Energy and Tourism (BMWET), and the State of Upper Austria within the framework of the SCCH competence center INTEGRATE (FFG grant no. 892418), as part of the COMET – Competence Centers for Excellent Technologies Programme, managed by the Austrian Research Promotion Agency (FFG).

Declaration of use of Generative AI

In the course of preparing this master's thesis, Generative AI was used as a supportive tool during the writing process. In particular, it assisted with linguistic revision, paraphrasing sentences, improving readability, and stylistically smoothing individual sections of the text.

All academic content, which includes the implementation, analysis, evaluation, interpretation of results, and conclusions, was developed independently by the author. AI did not contribute to the development of scientific ideas, arguments, or findings.

All generated suggestions were critically reviewed by the author and, where necessary, manually revised to ensure factual correctness, accuracy, and compliance with academic standards.

Abstract

Accurate daily demand forecasting is a critical operational challenge for bakeries, where perishable goods must be produced in appropriate quantities to minimize waste while meeting customer demand. This thesis empirically evaluates the performance of time-series foundation models (TSFMs), specifically TinyTimeMixer (TTM) and Moirai MoE, against classical forecasting approaches including naive baselines, Exponential Smoothing (ETS), Prophet, and XGBoost.

The evaluation uses real-world sales data from the organic bakery Brotsüchtig in Linz, Austria. Transaction-level point-of-sale data were aggregated to daily demand per stock keeping unit (SKU) and location. The empirical analysis focuses on the flagship branch and a portfolio of 48 frequently sold products, covering the period from January 2022 to June 2025. Model performance is assessed using a rolling-window backtesting setup with a one-day forecast horizon. Forecast accuracy is primarily evaluated using Weighted Absolute Percentage Error (WAPE), supplemented by seasonal Mean Absolute Scaled Error (sMASE) and mean error to assess competitiveness and systematic bias.

Results show that Moirai MoE achieves the lowest aggregated WAPE of 0.24, marginally outperforming the four-week median heuristic (0.25), ETS (0.26), and Prophet (0.27), while TTM and XGBoost perform substantially worse. Moirai MoE's advantage is most pronounced for difficult-to-forecast products and is statistically significant, although the practical margin over strong classical baselines remains small. Foundation models further exhibit a narrower distribution of per-product forecast errors compared to classical approaches, suggesting more consistent performance across the product portfolio.

The inclusion of weather and holiday covariates in Prophet does not improve overall forecast accuracy, with statistically significant but small negative median effects observed for both feature groups. Finetuning TTM on the bakery dataset substantially improves its performance, reducing median WAPE by 0.10 and benefiting 95.4% of products, although the model does not close the performance gap to the strongest classical approaches. Finetuning is also associated with an increase in systematic over-forecasting bias.

These findings suggest that demand in small-scale bakery retail is largely dominated by stable seasonal structure and that increasing model complexity yields diminishing returns under real-world conditions. Simple and interpretable forecasting methods remain competitive and may therefore represent the most practical choice for operational deployment in resource-constrained retail environments.

Kurzfassung

Eine präzise tägliche Nachfrageprognose stellt für Bäckereien eine zentrale operative Herausforderung dar. Da Backwaren in der Regel nur am Produktionstag verkauft werden können, müssen Produktionsmengen möglichst genau geplant werden, um sowohl Lebensmittelverschwendung als auch Fehlmengen zu vermeiden.

Diese Arbeit untersucht empirisch die Leistungsfähigkeit von Time-Series Foundation Models (TSFMs), konkret TinyTimeMixer (TTM) und Moirai MoE, im Vergleich zu klassischen Prognoseverfahren wie naiven Basismodellen, ETS, Prophet und XGBoost.

Die Analyse basiert auf realen Verkaufsdaten der Bio-Bäckerei Brotsüchtig in Linz, Österreich. Transaktionsbasierte Kassendaten wurden zu täglichen Absatzwerten pro Produkt und Standort aggregiert. Die empirische Untersuchung konzentriert sich auf die Hauptfiliale und ein Portfolio von 48 häufig verkauften Produkten im Zeitraum von Januar 2022 bis Juni 2025. Die Modelle werden mithilfe eines Rolling-Window-Backtesting-Ansatzes mit einem Prognosehorizont von einem Tag bewertet. Als zentrale Kennzahl dient der Weighted Absolute Percentage Error (WAPE), ergänzt durch den seasonal Mean Absolute Scaled Error (sMASE) sowie den mittleren Fehler zur Analyse von Prognosegenauigkeit und systematischen Verzerrungen.

Die Ergebnisse zeigen, dass Moirai MoE mit einem aggregierten WAPE von 0,24 die beste Prognoseleistung erzielt und damit die 4-Wochen-Median-Heuristik (0,25), ETS (0,26) und Prophet (0,27) knapp übertrifft. TTM und XGBoost schneiden hingegen deutlich schlechter ab. Der Vorteil von Moirai MoE zeigt sich insbesondere bei schwer prognostizierbaren Produkten und ist statistisch signifikant, auch wenn der praktische Vorsprung gegenüber starken klassischen Basismodellen gering bleibt. Darüber hinaus weisen Foundation Models eine geringere Streuung der produktspezifischen Prognosefehler auf als klassische Verfahren, was auf eine konsistentere Leistung über das gesamte Produktportfolio hinweist.

Die Einbeziehung von Wetter- und Feiertagsvariablen in das Prophet-Modell führt insgesamt zu keiner Verbesserung der Prognosegenauigkeit; für beide Variablengruppen zeigen sich statistisch signifikante, jedoch operativ nur geringe negative Median-Effekte. Finetuning von TTM auf den vorliegenden Datensatz verbessert dessen Prognoseleistung deutlich: Der mediane WAPE reduziert sich um 0,10, und 95,4 % der Produkte profitieren von dieser Anpassung. Dennoch bleibt ein Leistungsabstand zu den besten klassischen Verfahren bestehen. Gleichzeitig nimmt mit dem Finetuning die Tendenz zu systematischen Überprognosen zu.

Die Ergebnisse deuten darauf hin, dass die Nachfrage im kleinbetrieblichen Bäckereieinzelhandel stark durch stabile saisonale Muster geprägt ist und dass steigende Modellkomplexität unter realen Bedingungen nur begrenzte zusätzliche Verbesserungen bringt. Einfache und gut interpretierbare Prognosemethoden bleiben daher wettbewerbsfähig und stellen in ressourcenbeschränkten Einzelhandelsumgebungen möglicherweise die praktikabelste Lösung für den operativen Einsatz dar.

Contents

List of Figures

List of Tables

List of Abbreviations

CPU Central Processing Unit

DSR Design Science Research

ETS Exponential Smoothing

GPU Graphics Processing Unit

IQR Interquartile Range

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

ML Machine Learning

ME Mean Error

MoE Mixture of Experts

SCCH Software Competence Centre Hagenberg

sMASE Seasonal Mean Absolute Scaled Error

SKU Stock Keeping Unit

TSFM Time Series Foundation Model

TTM TinyTimeMixer

WAPE Weighted Absolute Percentage Error

1

Introduction

Accurate demand forecasting is a fundamental operational requirement in retail, particularly for categories involving perishable goods. Unlike durable products, perishables impose a hard constraint on inventory management: unsold stock cannot be carried forward, meaning that forecasting errors translate directly into economic and environmental costs (Arunraj & Ahrens, 2015; Fildes et al., 2022). Underestimating demand leads to stockouts and lost revenue, while overestimating results in waste of raw materials, labour, and energy (Arunraj & Ahrens, 2015). This two-sided cost structure makes short-term demand forecasting a problem of considerable practical consequence across a wide range of retail contexts.

Classical approaches to retail demand forecasting, which include heuristics, statistical methods, and machine learning approaches, have been extensively studied and deployed in real-life settings. Their performance, however, is highly sensitive to the characteristics of the underlying data: they require sufficient historical observations to estimate parameters reliably and struggle in settings where demand is sparse, noisy, or intermittent (Fildes et al., 2022). These conditions are particularly common at the product level (Fildes et al., 2022).

The recent development of time-series foundation models (TSFMs) represents a potentially significant methodological advance. By pre-training on large and diverse corpora of time series data, these models promise strong zero-shot and few-shot forecasting performance across a wide range of domains (Dayama et al., 2024; Garza et al., 2023; Liu et al., 2024). This property may be especially valuable in settings where local training data is scarce or noisy, as the learned representations may allow signal to be extracted that classical estimators would otherwise miss. Benchmark evaluations of TSFMs typically involve large, well-structured datasets under controlled conditions (Li et al., 2025), and whether such performance transfers to noisier, smaller real-world datasets remains an open empirical question.

In particular, it is unclear whether the benchmark performance of TSFMs transfers to noisy and intermittent demand data typical of perishable goods retail. Furthermore, benchmark evaluations typically focus on global accuracy metrics, while domain-specific aspects such as error stability, the contribution of exogenous covariates, and the effects of finetuning on small real-world datasets remain insufficiently explored. A model that leads on aggregate metrics may still exhibit operationally relevant behaviours, such as systematic bias on certain days, which would remain undetected without closer analysis.

To investigate this question in a realistic setting, this thesis examines the forecasting problem in the context of daily demand for perishable bakery products. Daily sales series of perishable food products in retail can be volatile, skewed, affected by external demand factors, and difficult to forecast accurately (Arunraj & Ahrens, 2015). Forecasting demand for perishable bakery goods shares several characteristics with broader retail perishables forecasting, including short shelf life, demand uncertainty, and the operational consequences of over- and underproduction, making it a suitable and practically relevant application case for the problem described above (Fries & Ludwig, 2024). Moreover, bakery-specific data are likely underrepresented or entirely absent from the pre-training of most TSFMs, which raises further questions about the relevance of learned temporal representations to this domain (Dayama et al., 2024).

This thesis was conducted in cooperation with the Software Competence Centre Hagenberg (SCCH) and with the support of Brotsüchtig, an organic bakery operating four branches in and around Linz, Austria. A project report in this context indicates that the introduction of BI dashboard support can reduce production waste by up to 20% and cut the time required for production planning by half ('Digitale Lebensmittelrettung', 2024). However, this process remains manually driven, and previous attempts to establish automated forecasting encountered difficulties due to the inherently noisy, low-signal nature of the sales data, further motivating empirical investigation.

To address this gap, the thesis investigates TSFM performance through three distinct lenses, giving rise to the following research questions:

RQ1: How do pre-trained time-series foundation models (TTM, Moirai MoE) compare with classical statistical and machine learning approaches (naive baselines, XGBoost, ETS, Prophet) for daily demand forecasting?

This question forms the empirical core of the thesis. It evaluates the strengths and weaknesses of TSFMs relative to established forecasting approaches, conducted in the context of next-day demand forecasting on a real-world bakery sales dataset. Since demand may be influenced by external factors, the contribution of operationally relevant covariates is examined separately in RQ2.

RQ2: To what extent do exogenous covariates, such as calendar effects and weather variables, improve forecast quality in daily demand forecasting?

This question examines whether commonly assumed external drivers of demand contribute meaningfully to forecast quality under the conditions of the case dataset under study.

RQ3: How does finetuning affect the forecast quality of pre-trained time-series foundation models in short-term demand forecasting?

This question addresses the adaptability of TSFMs to a specific application context, supports a fairer comparison with classical approaches trained directly on the case dataset, and examines whether cross-series information can be exploited to improve item-level forecasts.

The contribution of this thesis lies in providing empirical evidence on the comparative performance of time-series foundation models and classical forecasting approaches in a real-world bakery retail context. This thesis further examines how exogenous covariates and finetuning influence forecast quality, providing insight into the conditions under which foundation models can be meaningfully

adapted to small-scale, domain-specific retail settings. These empirical findings are intended to inform model selection and system design within a subsequent design and implementation cycle.

The overarching design objective of this research is to improve demand forecasting for perishable bakery products. To investigate this objective and address the research gap identified above, the study adopts the Design Science Research (DSR) framework following Wieringa (2014).

Design Science aims to develop artefacts that provide practical utility within a socio-technical context (Hevner & Chatterjee, 2010; Wieringa, 2014). In the context of this thesis, the artefacts under investigation are different classes of time-series forecasting models. The broader design goal is to support improved operational decision making in retail environments where demand is volatile and products are highly perishable.

Within Wieringa's framework, this thesis contributes to the *empirical cycle* of Design Science. Rather than focusing on the implementation of a specific forecasting system, the work investigates knowledge questions concerning the behaviour and performance of alternative forecasting approaches. In particular, the research examines comparative performance, sensitivity to external covariates, and the effects of model adaptation through finetuning. The resulting empirical evidence is intended to inform future design decisions regarding forecasting systems in similar operational contexts.

Methodologically, the study adopts a validation research design based on a single-case setting, performing technology-oriented experiments as defined by Wohlin et al. (2024) to compare alternative forecasting implementations on the same real-world dataset. The empirical analysis is conducted using real sales data from a multi-branch organic bakery. This case context provides a realistic environment in which the behaviour of forecasting models can be examined under conditions typical for small-scale retail demand: noisy observations, sparse time series, and strong day-to-day variability.

Following Wieringa (2014), the results of the empirical investigation contribute to answering knowledge questions about forecasting artefacts and their behaviour in practice. These insights are subsequently interpreted in light of the broader design objective of reducing operational inefficiencies such as overproduction and product waste through improved demand forecasting.

This thesis is organised in six chapters. The present chapter motivates the research problem, introduces the case context, defines the research questions, and positions the thesis scientifically. Chapter 2 provides the theoretical background necessary to contextualise the empirical work, covering the fundamentals of time series forecasting in retail settings and introducing both classical and foundation model approaches. Chapter 3 details the methodology. The dataset and its preprocessing are described, and the model selection, implementation, and evaluation strategy are justified and documented. Chapter 4 presents the results of the empirical analysis across all three research questions. Chapter 5 discusses these results, contrasts them with findings within the broader literature, and reflects on their implications for the design objective. Chapter 6 concludes the thesis by summarising the key findings, acknowledging limitations, and proposing directions for future work.

2

Background

Chapter 2 explains the necessary theoretical background and current state of the art. First, foundational terms and concepts for time series prediction in general, as well as retail demand forecasting specifically are introduced. Then different approaches for the prediction task are discussed, beginning with classical statistical approaches, machine-learning based forecasters, and culminating with foundation models. Lastly, the study is put within the context of the current scientific state of the art by examining related work, as well as positioned within its methodological context.

2.1 Foundations of Time Series and Demand Forecasting in Retail

The following section gives an overview across the necessary basics regarding the task of time series forecasting within the context of retail demand prediction. A time series is a collection of data points indexed in a sequential, temporal order (Box et al., 2016, Chapter 1). This dependence on a temporal sequence is a defining characteristic of time series data and makes the data points partly temporally dependent (Box et al., 2016, Chapter 1). A time series can have regular and irregularly spaced time intervals (Hyndman & Athanasopoulos, 2021, Chapter 1). Within the context of this study, such equally spaced time series is informally referred to as a "full" time series. A time series can generally be decomposed into trend, seasonality, cycles and a remainder (Hyndman & Athanasopoulos, 2021, Chapter 3). Trend and cycle are often combined and declared as trend (Hyndman & Athanasopoulos, 2021, Chapter 3). The trend captures the direction of a time series (Box et al., 2016, Chapter 4), while the seasonality denotes any cyclical pattern, with known and fixed periodicity (Box et al., 2016, Chapter 9). There can be more than one seasonality for one time series (Hyndman & Athanasopoulos, 2021, Chapter 3). Time series with no discernible trend are characterized as stationary (Box et al., 2016, Chapter 1). All structure not accounted for by these components is defined as the remainder (Hyndman & Athanasopoulos, 2021, Chapter 3). A time series is called multivariate as opposed to univariate, if it contains multiple time series with relevant relationships among themselves (Box et al., 2016, Chapter 14). These relationships are characterized as cross-series effects within the scope of this study. Data points, which are relevant to the forecasting outcome and external to a time series are called exogenous (Lima et al., 2025). Exogenous variables are one reason of multivariate time series (Box et al., 2016, Chapter 14). Orthogonally to the multivariate classification, if the objective is the forecast of more than one entity, the time series has a panel structure. Other often used features within time series prediction

are lag-features, which represent time-shifted values of a time series (Hyndman & Athanasopoulos, 2021). The prediction task can be further shaped by the length of the horizon h, which denotes the number of periods the forecaster predicts at each step (Hyndman & Athanasopoulos, 2021).

Within retail context, demand forecasting refers to the prediction of the quantity a product will sell within a given timeframe (Fildes et al., 2022). The domain context of retail demand has several characteristic challenges: Intermittent demand refers to sales patterns with a substantial share of zero demand periods, which violate the structural assumption of some forecasting approaches (Fildes et al., 2022). This also has implications on forecast evaluation and metric choice (Hyndman & Koehler, 2006). Intermittent demand also leads to demand distribution being non-normal and skewed (Fildes et al., 2022). Additionally, retail demand is often influenced by exogenous factors (Rose & Dolega, 2022).

2.2 Classical Approaches

In the context of this study, forecasting approaches can be broadly clustered into classical approaches, comprising baseline forecasters, statistical methods, as well as machine-learning models, and the recent development of time series foundation models. All approaches make distinct structural assumptions about the data and the modeled influences of the variables within the dataset. This is known as inductive bias and can help or hinder model performance, as well as generalization to unknown domains (Battaglia et al., 2018).

2.2.1 Baselines and Heuristics

To establish the forecasting quality advantage of sophisticated models, simple baselines and heuristics are commonly chosen as reference (Makridakis et al., 2020). These often prove surprisingly difficult to beat in forecast accuracy, especially within intermittent demand regimes (Fildes et al., 2022). Commonly used baselines are naive last observation forecast, which predicts the most recently observed value; the seasonal naive forecast, which predicts the value observed at the same point in the previous season; and rolling window statistics (Hyndman & Athanasopoulos, 2021, Chapter 5).

2.2.2 ETS

Exponential Smoothing State Space models, commonly referred to as ETS models, are a family of statistical forecasting methods that represent a time series as a combination of error, trend, and seasonal components (Hyndman & Athanasopoulos, 2021, Chapter 8). These components can be modeled as either additive or multiplicative (Hyndman & Athanasopoulos, 2021, Chapter 8). The trend component can be additionally dampened (Hyndman & Athanasopoulos, 2021, Chapter 8). The core characteristic of the forecasting process for ETS models is the recursive exponential weighing, which through a smoothing parameter controlling the rate of decay, disproportionally considers recent data points over more distant ones (Hyndman & Athanasopoulos, 2021, Chapter 8). Thus, ETS makes few assumptions about the stationarity of the dataset, as well as data distributions (Hyndman & Athanasopoulos, 2021, Chapter 8). This makes it a common canonical choice in forecasting competitions (Makridakis et al., 2020). A smooth evolution of the data points is however

required, which makes it less effective with intermittent demand (Hyndman & Athanasopoulos, 2021, Chapter 8).

2.2.3 XGBoost

XGBoost is a tree-based, machine-learning based approach to time series prediction (Chen & Guestrin, 2016). It step-by-step constructs an ensemble of decision trees by continuously correcting the residual error of each of the previous decision trees (Chen & Guestrin, 2016). XGBoost models are applied to time series prediction task by interpreting the forecasting problem as a regression task, with lag-features encoding temporal dependencies (Zhang et al., 2021). XGBoost is widely used and has demonstrated strong performance across multiple domains, as well as in incorporating complex non-linear relationships between features (Chen & Guestrin, 2016).

2.2.4 Prophet

Prophet is a forecasting approach developed by the company Meta based on additive decomposition of time series into trend, a Fourier-series representation of yearly and weekly seasonality and holiday effects (Taylor & Letham, 2018). This is optionally augmented by additional, linearly additive, regressors (Taylor & Letham, 2018). It is robust to missing data, shifting demand regimes, as well as being designed to be particularly well-suited for retail demand prediction tasks (Taylor & Letham, 2018). One key limitation to Prophet is that it assumes that trend, seasonality, holiday effects, and additional regressors contribute independently and additively to the final prediction. This may limit its ability to capture complex non-linear interactions (Hyndman & Athanasopoulos, 2021, Chapter 12)

2.3 Foundation Models

A foundation model is commonly defined as a large, pretrained model which can be used across diverse domains through zero-shot use or finetuning (Schneider et al., 2024). The most widely known example of foundation models are within the domain of language processing, with the transformer-based ChatGPT, Claude, Gemini as well as various other competing models (Schneider et al., 2024).

Within the scope of this study Time Series Foundation Model (TSFM)s are considered foundation models, which are predominantly trained on time series data, with the goal of building a time series forecaster. This excludes forecasting by LLMs. Zero-shot forecasting refers to the capability of a foundation model of producing forecasts across diverse domains and datasets, without explicit prior training on the data (Dayama et al., 2024). This is made possible by the inductive bias the model learned during its pre-training (Lovering et al., 2021). Church et al. (2021) defines finetuning as the adaptation of a foundation model to a specific domain or downstream task, by modifying some or all of its pre-trained model parameter.

2.3.1 Tiny Time Mixer

TinyTimeMixer is a compact TSFM developed by IBM (Dayama et al., 2024). It is trained primarily on selected datasets from known repositories such as Monash and LibCity in domains,

like weather, traffic, electricity, solar/wind, web traffic, births, bitcoin, taxi/traffic sensors (Dayama et al., 2024).

Its architecture is based on a lightweight mixer design based on multi-layer perceptron-architecture that captures temporal dependencies and cross-channel interactions with exogenous variables without the computational overhead of full transformer architectures, making it compatible with Central Processing Unit (CPU) inference environments (Dayama et al., 2024). Finetuning enables TinyTimeMixer (TTM) to learn cross-series effects producing forecasts for panel structured data (Dayama et al., 2024).

2.3.2 Moirai MoE

Moirai MoE is a transformer-based TSFM, developed by Salesforce, building on the previous Moirai model by including a sparse Mixture of Experts (MoE) architecture (Liu et al., 2024). These experts, selectively activate different components of the model depending on the input, thereby improving the handling of heterogeneous time series and data regimes (Liu et al., 2024). Moirai MoE is trained on a vast and diverse repository of datasets including both real and synthetic data, from various domains such as energy, buildings, climate, transport, cloud ops, finance, healthcare, sales, web traffic, air quality and epidemiology, giving it a strong cross-domain applicability (Woo et al., 2024). It supports the use of exogenous information, as well as both zero-shot and finetuned forecasting (Liu et al., 2024).

2.4 Related Work

This master thesis lies in the intersection of three areas of interest which structure the following section: foundation model Forecasting, Retail Demand Forecasting and Forecasting of Perishable Goods specifically.

Foundation Model Forecasting

Drawing inspiration from the success of foundation models in Natural Language Processing, applications of this model class to problems of time series forecasting have been garnering academic interest (Schneider et al., 2024). Multiple models claim competitive forecasting performance in comparison to classical approaches to time series prediction (Das et al., 2024; Dayama et al., 2024; Garza et al., 2023; Liu et al., 2024) Surveys confirm strong performance on benchmark datasets (Liang et al., 2024; Miller et al., 2024). However, these empirical results are contested by skepticism about highly complex models outperforming much simpler forecasting methods. Zeng et al. (2023) shows that simple one-layer linear model outperforms a range of sophisticated Transformer-based forecasters on long-horizon benchmarks.

Finetuning TSFMs on specific domains has been associated with inconsistent impact on forecasting performance, as demonstrated in Li et al. (2025). Additionally, some concerns about overfitting and disappointing real-life performance have been raised (Qiao et al., 2025).

Retail Demand Forecasting

Retail demand forecasting is a well-established research field and has consistently been a topic of extensive academic interest. Fildes et al. (2022) reviews both the current research landscape, as well as practical considerations within the field in a comprehensive way. Characteristic for the domain of retail demand are structural challenges, such as intermittent demand, short training data and complex influence of sales by exogenous factors, such as weather, holidays and promotions (Fildes et al., 2022). A large share of zero sales periods complicate both model selection, as well as metric choice (Hyndman & Koehler, 2006). Calendar-based covariates have been identified within the literature to be effective at improving forecasting quality, as Huber and Stuckenschmidt (2020) shows. Weather has similarly been identified as a significant driver of retail sales volumes, with machine learning models successfully taking advantage of the link between temperature, precipitation, and wind on daily demand (Chan & Wahab, 2024; Rose & Dolega, 2022).

Despite the success of complex machine learning approaches, simple statistical models maintained relevance in large time series forecasting competition, such as M4 and M5 (Makridakis et al., 2020, 2022). This skepticism toward complexity is echoed by Elsayed et al. (2021), who showed that gradient boosted tree model in some cases outperformed a range of deep learning approaches across nine standard datasets.

A few researchers, such as Chowdhury and Rozony (2025) and Ribeiro (2025), compare TSFMs and alternative forecasting approaches with promising results in the domains of e-commerce retail and Portuguese bakery sweets, respectively. These studies are currently available as preprints rather than peer-reviewed publications, but they nevertheless indicate growing academic interest in applying TSFMs to real-world demand forecasting problems.

W. Yang et al. (2025) represents a recent attempt to apply foundation models to real-world sales forecasting through hierarchical and architectural ensembling on the M5 benchmark, but focus on large-scale retail environments without exogenous covariates rather than small-scale, single-location bakery data. Gu et al. (2025) demonstrate that foundation models show promising performance in the energy domain.

Forecasting of Perishable Goods and Bakery Goods

Perishable goods have the attribute of only being able to be sold for a short amount of time. Thus, any demand miscalculation leads directly to stock-outs or waste. This waste problem has been identified as meaningful and addressed by Lebersorger and Schneider (2014) and Pietrangeli et al. (2023). Several proposals have been made to estimate perishable goods demand and thus reduce waste with the help of machine learning. In the bakery domain specifically, Hübner et al. (2024) conducted a life-cycle assessment of a machine learning forecasting service and found that a reduction in bakery returns of approximately 30% was associated with environmental benefits substantially exceeding the computational cost of the system. Prior work done by Wikamulia et al. (2023) applied XGBoost combined with k-means clustering to bakery demand forecasting and reported competitive accuracy, though their study was limited to a single product category. Taken together, this strand of literature confirms both the practical relevance of accurate bakery demand forecasting and the relative scarcity of rigorous empirical comparisons across model classes in this domain.

To the best of the author's knowledge, no prior study has conducted a controlled empirical comparison of time series foundation models against classical forecasting approaches using real-world bakery sales data, nor has the effect of finetuning and cross-series panel learning on TSFM performance been evaluated in this context. This thesis aims to address that gap.

3

Experimental Setup and Implementation

Chapter 3 describes the empirical evaluation setup, detailing the design decisions made, as well as the implementation details needed to replicate this technology-oriented comparison of forecasting implementations. As established in Chapter 1, the work is situated within the empirical cycle of the DSR framework, investigating effect, sensitivity, and trade-off questions about different forecasting model classes in a single-case setting.

First, the origins and composition of the data used for the empirical implementation comparisons are described, together with the pre-processing and feature engineering applied to the data. Then the technology-oriented experimental set-up and how the implementation comparisons were conducted are elaborated upon. Subsequently, the model selection is justified and their implementation detailed. Lastly, the metric selection is motivated and the analytical approach toward answering the knowledge questions is described.

3.1 Datasets and Pre-processing

This section presents the data used in the empirical evaluation. First, the origin and pseudonymisation of the raw data are outlined, followed by a description of the dataset structure and content. Subsequently, the pre-processing steps and feature engineering applied to prepare the data for modelling are detailed, including the construction of exogenous variables, calendar features, and lag-based predictors.

3.1.1 Data Source and Pseudonymisation

The data provided by the organic bakery Brotsüchtig comprise sales records from four locations in and around Linz. In total, the dataset contains 356 distinct product types and includes 2,306,527 recorded transactions covering the period from January 3, 2022 to June 30, 2025.

The data originate from the point-of-sale system used by the bakery and supplied by the company Ready2Order. The historical sales data are available as monthly exports and are organised in a directory structure segmented by branch and fiscal year. Before further processing, the data were pseudonymised. Sales volume was scaled by a constant factor and the date and the values in the column `Artikel_Bezeichnung` were replaced with an ordered index in order to protect sensitive

operational data. The data are not publicly available[1].

3.1.2 Dataset Structure and Content

The dataset follows a transaction-based structure. Each row represents a single line item within a transaction at a specific point in time. A line item may contain one or multiple units of the same product type.

The export contains a total of 79 columns including information on invoices, products, discounts, taxes, customers, suppliers, and organisational metadata. A complete list of the included columns is presented in Appendix A.

3.1.3 Aggregation and Filtering

As demonstrated by the column overview, which can be found in Appendix A, each item is assigned to one of seven higher-level product categories. Within the scope of this study, the analysis was restricted to items belonging to category 7, which comprises sold food products. Other categories, such as vouchers or non-sales-relevant entries, are excluded from all subsequent analyses.

Since sales behaviour differs between locations, the transactions were grouped by location. As the forecasting task requires predicting the daily demand of a single product for the following day, the data were additionally grouped by product, location, and date, then aggregated to the daily level.

Columns were included in the forecasting task if the information they contained was known at forecasting time and was not itself generated by the sale. This constraint ensures that the models only use information that would be available in a real forecasting scenario, preventing information leakage from future observations. The resulting dataset contains the columns `location`, `date`, `product`, and `artikel_menge`.

For the purposes of this analysis, a complete time series with regularly spaced daily intervals was constructed. Expected zero-sales days, such as holidays and Sundays, were not excluded from the training dataset. This approach was chosen to ensure a fair comparison between models, as some methods such as ETS, and shift-based naive baselines, are disadvantaged when applied to irregular time series containing gaps caused by zero-sales days (Hyndman & Athanasopoulos, 2021, Chapter 13).

The training data exhibit clear seasonal patterns at multiple levels. Figure 3.1 shows a pronounced yearly pattern in total sales volume. Figure 3.2 reveals strong weekly seasonality, particularly on weekends. As shown in Figure 3.3, location 1 records the highest sales volume across all locations.

3.1.4 Exogenous Variables

In addition to historical sales data, exogenous data are incorporated into the analysis.

[1] Researchers seeking access to the dataset for replication purposes may contact the data owner directly at office@scch.at.

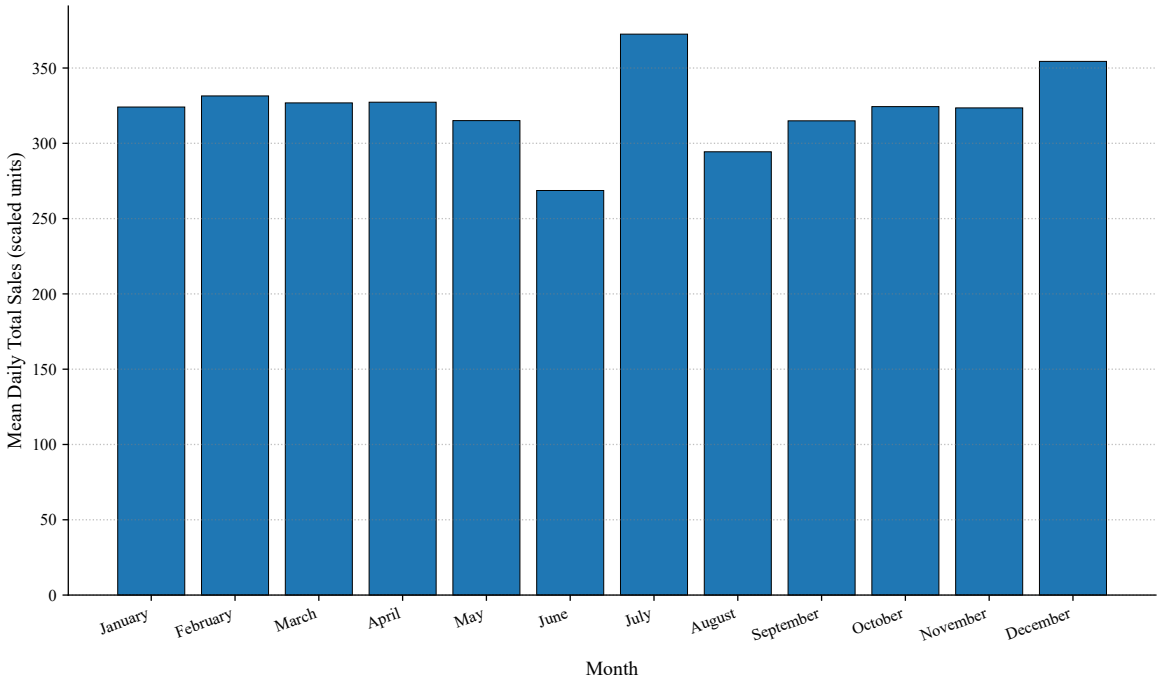


Figure 3.1: *Mean daily sales volume for every month. Average over two years.*

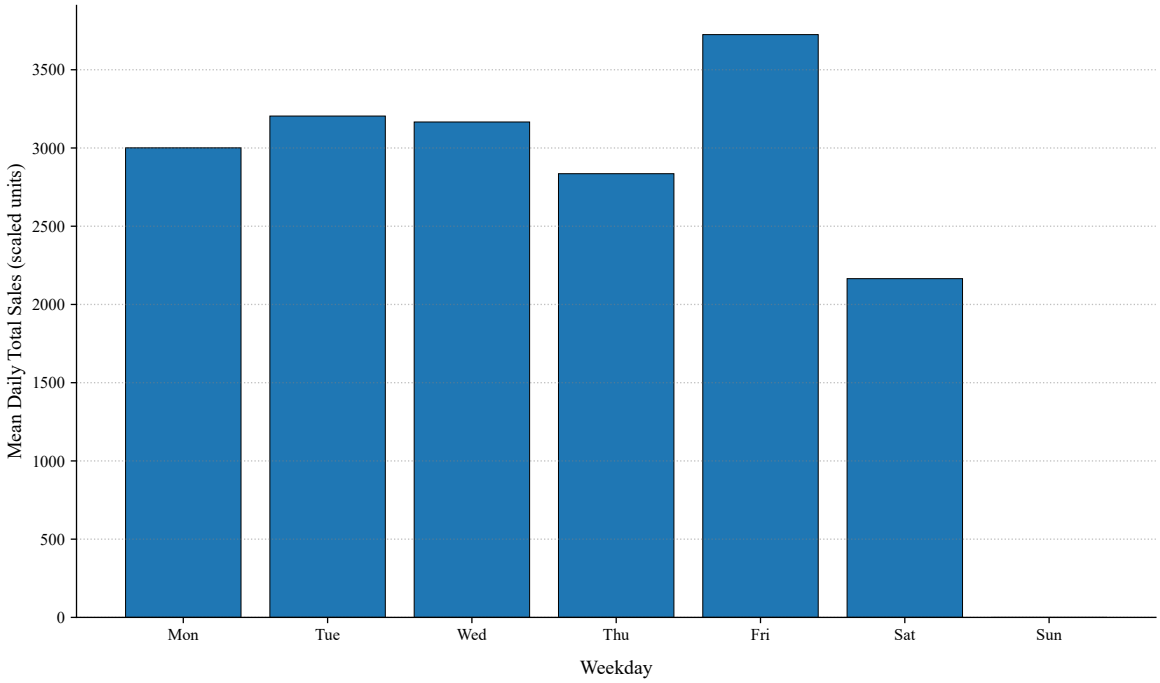


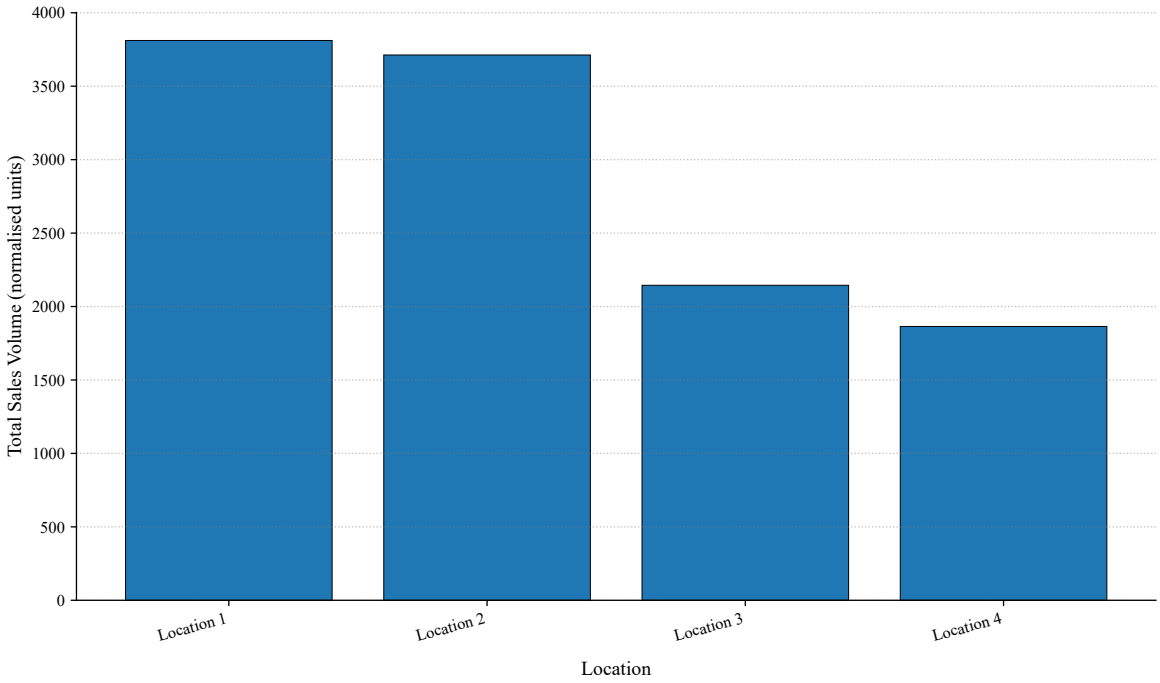
Figure 3.2: *Mean daily sales volume for every weekday. Average over two years.*

Weather

The first kind of exogenous data concern weather. These data are obtained through the Open-Meteo weather API and temporally aligned with the sales data ('Historical Forecast API | Open-Meteo.Com', 2026).

In contrast to a real deployment scenario, these weather data obtained are not a next-day forecast of the weather. They are the historical ground-truth and therefore do not share the inherent uncertainty of a weather forecast, which would be relevant in real-life deployment. Thus, the

Figure 3.3: *Mean sales per location over 2 years*



ability to exploit the effect of weather on sales found in this analysis is biased to be overoptimistic compared to the external truth. Using historical data rather than historical forecast simplifies the analysis process by isolating the feature usefulness and avoiding confounding from forecast model errors. As the research goal focuses on comparing forecasting approaches rather than identifying externally valid drivers of bakery sales, perfect knowledge of exogenous covariates is assumed.

The following weather variables are used in the analysis: `weathercode`, `temperature_2m_max`, `temperature_2m_min`, and `precipitation_hours`. The selection of these variables is based on consultations with stakeholders and prior experience with weather-sensitive sales patterns at the bakery.

Calendar and Holiday Effects

Further covariates are grouped into three categories: lag variables, holiday-related features, and so-called spike days. As discussed in Chapter 2.4, Huber and Stuckenschmidt (2020) demonstrates the impact of holidays, school vacations, and other special calendar days on forecast accuracy and these features were therefore added during the feature engineering process.

Following stakeholder consultation, additional calendar-related effects were incorporated: School vacation periods and "bridge days", which are working days falling between a public holiday and a weekend, as practitioners reported these periods to be associated with systematically higher sales volumes.

3.1.5 Feature Engineering and Imputation

Lagged demand effects are incorporated into the models via explicit lag variables. The feature `sales_4weeksmedian` is used to approximate the manual forecasting heuristics currently employed by staff by modelling medium-term demand level.

Furthermore, all exogenous variables are augmented with temporal lags of 1, 4, 7, and 28 days, a standard approach to capture delayed effects in time-series forecasting (Hyndman & Athanasopoulos, 2021).

The sales data are highly concentrated, with 10% of the days accounting for nearly half of the total sales volume. This can be seen in Figure 3.4, where the x-axis shows the share of days and the y-axis shows the share of sales volume generated by those days. The grey dashed line indicates a uniform distribution for reference. The curve lying substantially above this line indicates that the majority of sales volume is concentrated within relatively few days.

Only four days account for more than a tenth of the annual sales volume, highlighting the presence of extreme demand peaks. Such concentration suggests a highly uneven demand distribution and poses forecasting challenges, as rare high-demand days are inherently difficult for models to learn.

Specifically, models may have difficulty extracting the patterns of high-demand days given their rarity. To address this, the feature `P_Spike` was introduced to make high-demand days more visible to the model. To construct this feature, an XGBoost classifier was trained to estimate the probability of a day belonging to the top 10% of sales days. The threshold was defined per group based solely on training data. The classifier used only information available at prediction time, ensuring no temporal leakage. The XGBoost classifier was chosen due to its state-of-the-art performance on tabular data, particularly for tasks with highly non-linear feature interactions, which fits the task of predicting high sale days (Chen & Guestrin, 2016).

The predicted probability was then used as an additional feature in the primary forecasting model. To prevent information leakage, feature and threshold limit construction strictly excluded future information and relied only on predictors available at prediction time. The complete feature specification and implementation details are provided in Appendix B.

Additionally, this feature may decrease under-forecasting caused by missed high-sale-events and thus impact bias direction and magnitude.

Non-exogenous columns containing missing values were imputed with zero. The point-of-sale system records every completed transaction; therefore, missing entries are reasonably interpreted as structural zero-demand days rather than measurement errors. Based on stakeholder feedback, supply shortages or stocking failures were not reported and are therefore considered unlikely to explain missing sales records.

This imputation ensures a continuous time series representation and enables the evaluation of models under intermittent demand conditions.

Exogenous variables did not contain missing values and therefore required no imputation.

The final dataset thus contains 61200 rows and 45 columns. The full list of all columns can be found in Appendix C.

3.2 Experimental Design and Backtesting Protocol

The target variable for the forecasting task is the daily quantity sold in units, represented by the column *artikel_menge* for each location–SKU combination.

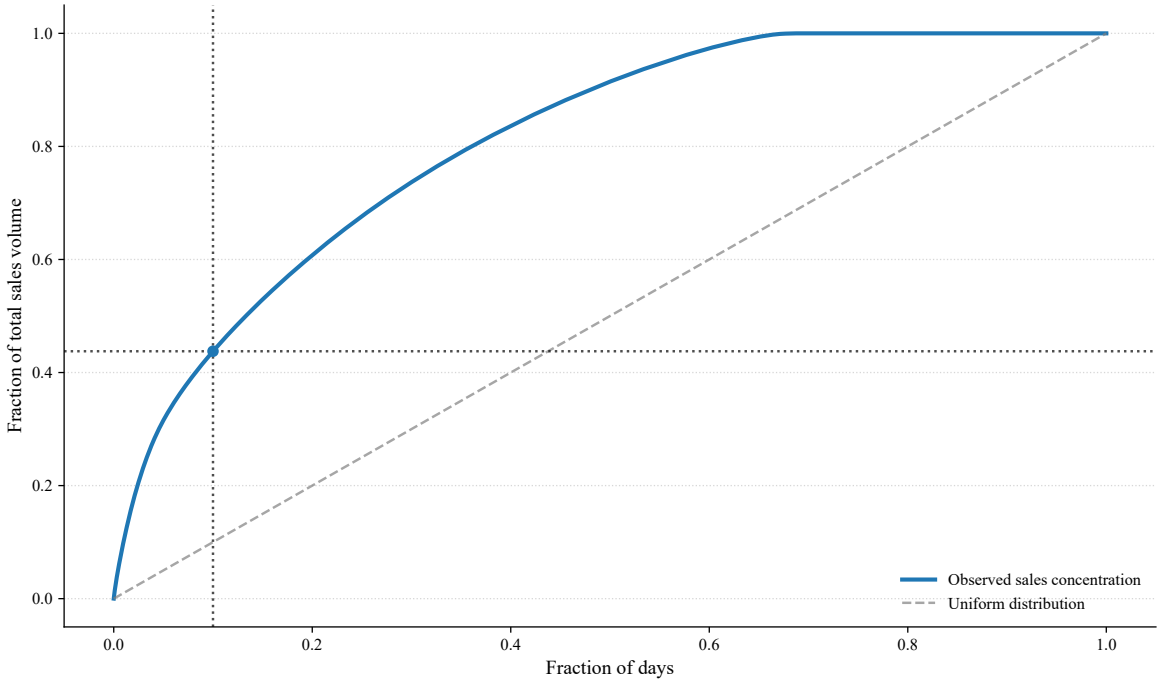


Figure 3.4: *Share of days contributing to share of sales volume relative to a uniform distribution*

To estimate model performance under conditions approximating real-world deployment, a rolling-window backtesting protocol is employed. This approach preserves the temporal order of observations and ensures forecasts are generated using only information available prior to the prediction horizon, thereby mimicking real-world deployment conditions.

Allocating data for training versus evaluation involves a fundamental trade-off: models need sufficient data to learn relevant patterns, but also require a large and diverse test set to enable systematic comparison. Because RQ2 concerns how seasonal holidays affect forecast quality, the training data must span at least two years to capture two instances of each annual holiday and seasonal cycle. For the same reason, the test period must cover a minimum of one year so that each holiday and season is represented at least once. The data are therefore partitioned so that the first 912 days form the initial training period, with the remaining data reserved for testing. To ensure the stability of the lag features, a 28-day buffer is applied: the first 28 days following the training period are excluded from the test set, allowing the lag features to be populated with sufficient historical observations before evaluation begins. Evaluation is conducted using a fixed-length sliding window of $L = 912$. At each cutoff c_i a new instance of each model is fitted on the current window to generate a one-day-ahead forecast ($h = 1$). After each prediction, the window advances by one day, incorporating the observed ground-truth sales quantity into the training data for the next iteration. This ensures models are updated with actual values rather than recursive predictions. A horizon of one day was chosen to best reflect the current daily forecasting process, while also simplifying evaluation and avoiding the error propagation inherent in multi-step forecasting. Holding the window length constant ensures a fair comparison across cutoffs and keeps the computational cost of training at later cutoffs manageable. The number of cut-offs processed within the scope of this empirical comparison is thus one year including the lag buffer.

To address the research questions, both panel-level and single-series analyses are required. Panel

modelling evaluates cross-series learning effects, while single-series analysis assesses performance at the individual SKU–location level. Panel-level analysis is relevant for RQ3.

For RQ2, an additive feature comparison based on predefined feature groups is employed to isolate the contribution of different feature clusters. Three feature groups are defined: BASE, WEATHER, and HOLIDAY, whose composition is elaborated upon in Table 3.1.

Table 3.1: Overview of feature groups and engineered variables

Category	Feature Group	Included Variables
Autoregressive	BASE	`sales_yesterday`, `sales_yesterweek`, `sales_avg4weeks`
Calendar & Context	HOLIDAY	`is_holiday`, `is_day_before_holiday`, `school_holiday`, `start_school_holiday`, `fenstertag`, `days_until_christmas`, `is_december_week_4`, `is_fenstertag_friday`, `is_payday_week`, `holiday_proximity_7d`, `P_spike`
Exogenous	WEATHER	`weathercode (wmo code)`, `temperature_2m_max (°C)`, `temperature_2m_min (°C)`, `precipitation_hours (h)`

The BASE-only model is thus compared against BASE + WEATHER and BASE + HOLIDAY. RQ3 similarly compares the finetuned model against the zero-shot model.

3.2.1 Scope and Limitations

Due to the large number of cutoffs and the high computational demands of the analysis, the scope of the study had to be limited in two respects: the number of locations and the number of products.

Of the four available bakery locations, only the flagship location with the highest overall sales volume was selected for analysis, ensuring sufficient sales density for model evaluation. Of more than 300 distinct products sold by the bakery, a subset of 48 products was selected for demand forecasting, accounting for 75.6% of all recorded transactions. Selection was based on sales frequency, as measured by the number of transactions per product, rather than total sales volume. Sales frequency was chosen as the selection criterion to align with the per-product evaluation design of this study. Since forecast accuracy is assessed individually for each product and aggregated using equal-weight summary statistics, such as the median, selecting by total sales volume would implicitly over-represent high-volume items in the evaluated sample. Frequency-based selection mitigates this bias by favouring regularly selling products without disproportionately weighting economically dominant SKUs.

To limit the computational budget, the models selected for RQ2 and RQ3 are subject to the following criteria:

- The effect of exogenous covariates on model performance, as required by RQ2, was examined only for models whose forecasting error was at or below that of the *Naive Last Week* baseline. This threshold restricts the analysis to models with competitive baseline performance, preventing the high error variance of weaker models from masking the impact of the covariates.

Additionally, models that fail to meet the standards of this baseline are reasonably unlikely to have learned enough signal from the data to be practically interesting for further covariate analysis.

- Finetuning effects were tested only on the smaller TTM variant. As the most compact foundation model included in this study, this restriction keeps the investigation of finetuning within a feasible timeframe and computational budget. The findings are considered indicative of the effect of finetuning on foundation model performance within this domain more broadly.

3.3 Model Selection

To enable a systematic comparison of TSFMs with alternative forecasting approaches, model selection follows the principle of covering representative approaches and paradigms from across the methodological spectrum of time-series forecasting.

Accordingly, this analysis examines a classical statistical model, a general machine learning approach, a model specifically designed for business forecasting, and two time-series foundation models.

These forecasting approaches are complemented by the inclusion of baseline techniques. Despite often exhibiting strong predictive performance, complex forecasting models frequently struggle to demonstrate a clear qualitative advantage over simple heuristics and baseline methods as seen in forecasting benchmark competitions like M4 (Makridakis et al., 2020). The absence of such baselines can lead to an overestimation of model performance as noted by Beck et al. (2025).

The baseline forecasts are derived from the demand observed from the previous day, the same day in the prior week, the same day in the prior year, and the four-week weekday median. Seasonality-based baselines were prioritised, as the data exhibit a strong weekly and yearly pattern, with sales volume often strongly influenced by both the day of the week and the time of the year, as demonstrated in Figures 3.2 and 3.1.

ETS was chosen as a representative of the statistical approach to time series forecasting, as the model is well-established in demand and inventory planning (Hyndman & Athanasopoulos, 2021). This is further demonstrated by its inclusion as a canonical choice within widely cited forecasting competitions such as the M4 (Makridakis et al., 2020). In contrast to other classical statistical models, most famously ARIMA and its variants, it is more robust for heterogeneous data, less rigid in its assumption of stationarity and fixed seasonal patterns (Hyndman & Athanasopoulos, 2021). The simplicity of the algorithm also reduces the risk of misconfiguration, strengthening the external validity of the empirical findings.

XGBoost is among the most widely used machine learning models with competitive results across diverse domains and has been considered a popular choice for prediction tasks in both research and practical areas (Chen & Guestrin, 2016; Kaggle, 2021). XGBoost has been applied successfully to retail demand forecasting in prior work as demonstrated in Chapter 2.4 (Zhang et al., 2021). Its strength in capturing complex non-linear feature interactions makes it well-suited to test the relative contribution of exogenous factors to forecast performance. Should these characteristics dominate over seasonality and autoregressive effects within this case context, XGBoost would be expected to perform particularly well.

Prophet was chosen as it was specifically developed for business forecasting contexts, with native support for seasonality and holiday effects (Taylor & Letham, 2018). Its performance in related retail demand forecasting work, as reviewed in Chapter 2.4, further supports its inclusion. Additionally, Prophet's additive decomposition in weekly and yearly cycles with consideration of public holidays, may fit well with the characteristics of the data used, as shown in 3.1. This additive structure contrasts with XGBoost's strength in capturing non-linear patterns, making the two models complementary.

Specific criteria were defined for the selection of time-series foundation models, including the ability to incorporate exogenous covariates and support cross-series learning. There was no access to a Graphics Processing Unit (GPU) environment, so the models had to be compatible with a CPU environment.

Based on these considerations, the TTM model developed by IBM was chosen as the primary foundation model (Dayama et al., 2024). Moirai Mixture of Experts (MoE), developed by Salesforce (Liu et al., 2024) was included as an additional reference model to probe whether zero-shot foundation models without task-specific adaptation already outperform classical baselines, and whether a transformer-based architecture shares converging forecasting behaviour with the multi-layer-perceptron-based architecture of TTM. Convergence or divergence in their forecasting behaviour would allow a preliminary assessment of whether the empirical findings hold across a structurally different foundation model.

ChronosX was not chosen despite support for exogenous variables, as it is primarily an adaptation mechanism requiring additional adapter training and design choices. To keep the comparison focused on out-of-the-box pretrained forecasters, only models that can be applied directly in a zero-shot setting, are considered (Arango et al., 2025).

3.4 Implementation Details

This section documents key implementation details of the models chosen for the study. This ensures that results are transparent and reproducible.

3.4.1 Software Environment

All analysis was conducted using Python 3.12.12 in a Google Colab environment. The processing unit type used was the default CPU offered by Google Colab. All the implementation comparisons were run without paid compute resources.

For reproducibility, random seeds were set to 42. The key libraries used were pandas as introduced by McKinney (2010) and numpy shown in Harris et al. (2020); a full list of libraries used as well as their corresponding versions can be found in the online appendix attached to this study.

3.4.2 Model Configurations

This section describes the model configurations and implementation details for each forecaster included in the study. Models that have the capability of handling panel data as well as single-series data were implemented separately for each setting.

Baseline Models

The baseline models do not require training and cannot consider exogenous variables. Therefore, predictions are generated for one time series per product per location. Any missing sales days are imputed to be zero. There are no explicitly modeled fallbacks for structural zero sales days, such as Sundays or holidays. While this decision may increase the error rate of the baselines, it preserves their purpose as quality gates for the other models. The baselines implemented were `Naive Yesterday`, `Naive Last Week` and `4-Weeks-Median`. `Naive Yesterday` was defined to be the last value in the series per product and location. `Naive Last Week` similarly takes the value of the series shifted by seven days, and does not correct for zero sales days. The `Four-Week Median` and `Four-Week Mean` heuristics take the median and mean respectively of the last four same weekdays. Missing lags, caused for instance by history shorter than four weeks, are ignored.

ETS

The exponential smoothing forecaster was implemented using the statsmodels library (Seabold & Perktold, 2010). The minimum valid length was set to 14 days. A weekly seasonality was assumed by setting `seasonal_periods = 7`. The horizon is set to one day. Four candidates are defined: additive trend with weekly seasonality, additive trend without seasonality, no trend with weekly seasonality, and no trend without seasonality. The first candidate to successfully fit is chosen as the forecasting approach. This fallback strategy ensures that for every day, ETS produces a valid forecast. This fallback is appropriate for noisy, potentially sparse time series where ETS may otherwise fail to converge. The trade-off is that the first successfully fitting candidate is selected, which may not be the globally optimal configuration. This is accepted as a reasonable compromise, as the empirical focus on foundation model performance consistent with the research goal necessitates consistent forecast coverage over optimal ETS configuration.

Prophet

Prophet was initialized using the `prophet` Python library (Taylor & Letham, 2018). Weekly and yearly seasonal effects were set to true with `weekly_seasonality=True` and `yearly_seasonality=True`. Daily seasonality was disabled (`daily_seasonality=False`). Holidays were modeled through the built-in holiday constructor. Austrian federal holidays were incorporated by setting `country=AT`. Additionally, Good Friday, Christmas Eve and New Year's Eve were defined as custom holidays and added to the model. Both holiday and seasonality effects were assigned the default prior scale `seasonality_prior_scale=10` and `holidays_prior_scale=10`. The additional regressors defined in Section 3.1.4 were included as exogenous inputs to the model.

XGBoost

The XGBoost model was configured with a fixed set of moderately regularised hyperparameters, with a small learning rate, medium tree depth, and subsampling of observations and features. No automated hyperparameter tuning was performed in order to maintain computational feasibility and ensure a fair comparison across models. The full hyperparameter list can be found in Appendix D.3.

Tiny Time Mixer

TTM was accessed via the sktime library (Löning et al., 2019). Configurations used for zero-shot evaluation were left unchanged and can be found in Appendix D.2. As TTM natively produces multi-step predictions at h=69 by default, only the first step was retained to match the one-day-ahead forecasting horizon of this study.

Moirai MoE

Moirai was accessed via sktime, loading a pre-trained model from Hugging Face (Löning et al., 2019). The specific model chosen was `Moirai MoE R2 small`. Extending Moirai MoE to panel and exogenous configurations was outside the computational scope of this study, as it would have required an additional set of experimental conditions equivalent in scale to those conducted for other models. Moirai MoE was therefore evaluated in its zero-shot configuration only.

3.4.3 Finetuning Protocol

For RQ3, TTM was finetuned using Optuna for hyperparameter search. The search space was iteratively refined based on intermediate results within the finetuning process. The final configuration used a learning rate of 4.39×10^{-4}, a dropout rate of 0.164, and trained for 19 epochs. Finetuning was conducted on panel data with cross-series learning enabled across all products.

3.5 Evaluation Metrics and Analytical Approach

Weighted Absolute Percentage Error (WAPE) and sMASE serve as error metrics for assessing the performance of the various forecasting methods. Tests were conducted to determine statistical significance.

Answering RQ1 requires more than a single aggregate metric, as the question concerns how model performance varies across products and relates to product characteristics such as sales volume and forecast difficulty. This necessitates per-product level analysis and a detailed error analysis across different product segments. The method used to achieve this is distributional analysis using box plots and Interquartile Range (IQR). The analysis is additionally stratified. Model performance across strata is compared through pairwise differences in error metrics between models to ensure the results of the analysis are robust to metric artifacts. RQ2 and RQ3 are framed as additive feature comparison studies and WAPE difference with differing capabilities and feature sets is used to quantify the marginal contribution of the added features.

3.5.1 Error Metrics

Due to the strong heterogeneity in sales volumes across SKUs, error metrics were computed per product and summarized across the portfolio using the median, which is robust to outlier products. To enable comparability across products with different demand scales, evaluation relies on unitless, scale-free error measures.

The Weighted Absolute Percentage Error (WAPE) as introduced by Kolassa and Schütz (2007) serves as the primary evaluation metric.

$$\text{WAPE} = \frac{\sum_{t=1}^{T} |y_t - \hat{y}_t|}{\sum_{t=1}^{T} y_t} \tag{3.1}$$

where T is the total number of observations. Unlike the Mean Absolute Percentage Error (MAPE) as defined by Hyndman and Koehler (2006), which is undefined when $y_t = 0$,

the Weighted Absolute Percentage Error (WAPE) accumulates absolute errors in the numerator and remains well-defined on zero-sales days. It is interpretable as the share of total demand that is misforecast, and its volume-weighting naturally aligns with operational objectives such as waste reduction, as it penalizes errors on high-volume products more heavily, reducing absolute waste quantities.

However, because WAPE's denominator scales with total sales volume, higher-volume products mechanically yield smaller values even when relative forecast errors are proportionally similar. To ensure that patterns observed in stratified analyses reflect genuine differences in forecasting quality rather than metric scaling behaviour, the seasonal Mean Absolute Scaled Error (sMASE) as introduced by Hyndman and Koehler (2006) is used as a secondary metric:

$$\text{sMASE} = \frac{\text{MAE}}{\text{MAE}_{\text{naive}}} = \frac{\frac{1}{T}\sum_{t=1}^{T} |y_t - \hat{y}_t|}{\frac{1}{T-s}\sum_{t=s+1}^{T} |y_t - y_{t-s}|} \tag{3.2}$$

where $s = 7$ reflects the weekly demand cycle. A value below 1.0 indicates the model outperforms the seasonal naive baseline; A value above 1.0 indicates the converse. Because Seasonal Mean Absolute Scaled Error (sMASE) normalizes by the in-sample naive error, it is invariant to the sales scale and provides a consistent basis for comparing model competitiveness across strata of differing volume or difficulty. It is not used as the primary metric because it is less directly operationalizable than the more intuitive percentage-error interpretation of WAPE, and because WAPE's volume-weighting is itself desirable from a business perspective.

To address the research questions, a global comparison of error metrics alone is insufficient. A more detailed analysis of the forecast bias and direction is therefore required. To address waste-related questions, an absolute quantity measure is needed alongside percentage-based metrics. For this reason, WAPE is complemented by the Mean Error (ME).

$$\text{ME} = \frac{-1}{T}\sum_{t=1}^{T} (\hat{y}_t - y_t) \tag{3.3}$$

Zero-sales days known in advance, such as Sundays and public holidays, were neither excluded from the test data nor omitted from the calculation of error metrics. An argument for excluding these days would be that forecasts are operationally irrelevant when closures are already known, and their inclusion may therefore increase measured forecast errors compared to real-world deployment scenarios. However, since the primary objective of this research is to evaluate model adaptability and the ability to exploit calendar-related information, such days provide an important test case, providing a more robust empirical basis for model comparison, which is central to RQ1 and RQ2.

Including them reveals whether a model can correctly identify structural demand patterns and learn calendar-based constraints. Their inclusion provides a more robust empirical basis for model comparison.

To balance these considerations, the analysis primarily reports results including Sundays and closure days, while additionally examining in a limited comparison within RQ1 whether conclusions change when these days are excluded.

3.5.2 Statistical Testing and Hypotheses

In addition to the p-value, effect size is measured by the median WAPE difference between models, complemented by win rate, loss rate, and tie rate across products. A tie is defined as a WAPE difference below 0.001 between two models for a given product. Win and loss rate refers to the share of products where model A has a better or worse performance respectively and the difference in WAPE is larger than a tie.

The hypotheses tested for each research question are presented in Table 3.2:

Table 3.2: Research hypotheses by research question

RQ	Hypothesis	Statement
RQ1	H_0	The leading model and the second-best model have equal median WAPE across products.
	H_1	The leading model has lower median WAPE than the second-best model.
RQ2	H_{A0}	The WAPE difference between the base model and the model with weather covariates is zero.
	H_{A1}	Weather covariates have a non-zero effect on WAPE.
	H_{B0}	The WAPE difference between the base model and the model with holiday covariates is zero.
	H_{B1}	Holiday covariates have a non-zero effect on WAPE.
RQ3	H_0	The WAPE of the zero-shot and finetuned model is equal.
	H_1	The WAPE of the zero-shot and finetuned model is not equal.

Statistically significant effects identified in the aggregated analysis for RQ2 and RQ3 are further analyzed across volume strata to assess heterogeneous effects across product segments.

To assess whether the performance advantage of the leading model over the second-best model is statistically significant rather than due to chance, the Wilcoxon signed-rank test (Wilcoxon, 1945) was applied. This test is appropriate because the unit of analysis is the individual product, resulting in paired per-product error metrics, and because the distribution of these metrics cannot be assumed to be normal

4

Results

This chapter reports the empirical findings of the forecasting models in the context of next-day demand prediction using the bakery sales dataset described in Chapter 3. Results are structured according to the research questions in Chapter 1.

4.1 Comparative Forecasting Performance Across Models

Models are compared across the full product portfolio using multiple performance dimensions, including aggregate WAPE, sMASE, and bias, as well as error distributions across products stratified by forecast difficulty and sales volume.

Table 4.1: Performance comparison of forecasting models. Models are sorted by ascending WAPE (lower is better).

Model Name	WAPE
Moirai MoE	0.24
4 Weeks Median	0.25
4 Weeks Mean	0.26
ETS	0.26
Prophet	0.27
Naive Last Week	0.29
TTM	0.48
Naive Last Year	0.52
Naive Yesterday	0.62
XGBoost	0.65

Table 4.1 summarizes forecasting performance measured by WAPE. The results show a clear separation between a group of well-performing models and a set of substantially weaker approaches. Several models cluster at the top of the ranking, with only marginal differences in average accuracy.

Moirai MoE achieves the lowest WAPE at 0.24, followed closely by the 4 Weeks Median heuristic with a WAPE of 0.25. The absolute difference between the best and second-best model is 0.01

WAPE points, equivalent to one percentage point. Classical seasonal approaches are competitive: ETS and 4 Weeks Mean attain a WAPE of 0.26, and Prophet achieves a WAPE of 0.27, corresponding to absolute differences of one to two percentage points relative to the best-performing model.

A pronounced performance gap emerges beyond this group. While the Naive Last Week baseline yields a WAPE of 0.29, forecast error increases sharply for more complex learning-based models. TTM records a WAPE of 0.48, comparable to the Naive Last Year baseline at 0.52. XGBoost and Naive Yesterday perform worst overall with a WAPE of 0.65 and 0.62 respectively.

These results indicate that while a time-series foundation model (Moirai MoE) achieves the best average forecasting accuracy, its advantage over strong classical baselines is small. TTM and XGBoost perform substantially worse, with WAPE values more than double that of the leading model.

Table 4.2: Performance comparison of forecasting models with Sundays excluded. Models are sorted by WAPE ascending.

Model Name	**WAPE**
Moirai MoE	0.23
ETS	0.23
Prophet	0.23
4 Weeks Median	0.24
4 Weeks Mean	0.25
Naive Last Week	0.29
TTM	0.35
Naive Last Year	0.37
XGBoost	0.42
Naive Yesterday	0.50

To determine whether the aggregated performance lead of Moirai MoE over the second best model found in Table 4.1 is statistically significant, a Wilcoxon signed-rank test was used comparing per-product WAPE values across all products.

Moirai MoE is found to have a win rate of 72% against the 4 Weeks Median heuristic, which produces a better forecast in 28% of cases. With a test statistic of 167 and a one-sided p-value of $p < 0.001$ (5.78×10^{-5}) the test is statistically significant. This indicates that the results are consistent across the product portfolio and are not an artefact of aggregation.

To assess the robustness of these results with regard to structural zero demand days, such as Sundays, the error metrics are recalculated in Table 4.2 with Sundays excluded. Notably, WAPE decreased for all observed models as a result of this recomputation. The performance gap between the top performing models and the rest persisted, and the composition of the top-performing group remains unchanged from Table 4.1 with similar forecasting performances. There are some substantial changes for models worse than the Naive Last Week baseline: TTM improves 0.13

absolute percentage points, from 0.48 to 0.35 with the exclusion of Sundays. XGBoost improves even more substantially to a WAPE of 0.42 with a 0.23 absolute difference in percentage points. In contrast, top performing models only improve by an absolute percentage point difference of 0.01 for Moirai MoE and the 4 Weeks Median and Mean heuristic, 0.03 absolute percentage point difference for ETS and a 0.04 absolute percentage point difference in WAPE for the Prophet model. The cluster of top performing models shares an identical median WAPE over all observed products of 0.23. This shows that model performance is influenced by structural patterns, such as zero sales on Sunday. The effect is uneven over all models and stronger the weaker a model is.

Table 4.3: Systematic Bias (Mean Error) of forecasting models. Models are sorted by proximity to zero (least biased first).

Model Name	**Mean Error (ME)**
Naive Last Week	0.00
Naive Yesterday	0.00
4 Weeks Mean	-0.01
ETS	-0.05
Prophet	-0.08
4 Weeks Median	-0.09
TTM	-0.12
Moirai MoE	-0.16
XGBoost	-0.77
Naive Last Year	-4.11

Note: Negative values indicate a systematic tendency to over-forecast (predicted > actual).

Table 4.3 reports the systematic bias of all forecasting models measured by ME, where negative values indicate a tendency to over-forecast demand. All evaluated models exhibit negative or near-zero ME values, indicating that none of the approaches systematically under-forecast sales.

The bias results reveal three distinct groupings across models. The smallest biases are observed for the baseline methods Naive Last Week and Naive Yesterday, both exhibiting ME values close to zero, with the 4 Weeks Mean heuristic close behind at ME = -0.01. These models can be considered approximately neutral in terms of systematic bias.

A second group of models shows moderate to strong over-forecasting behaviour. ETS records an ME of -0.05, while Prophet and the 4 Weeks Median heuristic exhibit ME values of -0.08 and -0.09, respectively. Notably, the median-based heuristic displays a substantially larger bias than its mean-based counterpart, despite similar forecasting accuracy measured by WAPE.

The foundation models exhibit stronger systematic bias within the second cluster. TTM records an ME of -0.12, while Moirai MoE shows an ME of -0.16, indicating a stronger tendency to over-forecast relative to classical baselines.

Finally, two models stand out due to extreme bias. XGBoost exhibits a markedly larger ME of

-0.77, and the Naive Last Year baseline shows the strongest over-forecasting behavior with an ME of -4.11, representing a qualitative increase in bias compared to all other approaches.

Table 4.4: Systematic Bias (Mean Error) of forecasting models with Sundays excluded. Models are sorted in descending order by ME value.

Model Name	**Mean Error (ME)**
TTM	1.34
Naive Yesterday	1.29
XGBoost	1.27
Naive Last Week	0.00
4 Weeks Mean	-0.01
4 Weeks Median	-0.10
ETS	-0.14
Prophet	-0.17
Moirai MoE	-0.25
Naive Last Year	-2.71

Note: Negative values indicate a systematic tendency to over-forecast (predicted > actual). Positive values indicate a tendency to under-forecast.

To assess how sensitive these results are to structural zero demand days, such as Sundays, Table 4.4 excludes those days from the computation of the bias. Three main clusters of positive bias, neutral bias, and negative bias, can be identified from the resulting bias Table. The baseline approaches Naive Last Week with a ME of 0.00 and 4 Weeks Mean with a slight over-forecast of -0.01 are closest to being neutral in terms of bias among all observed models, both of which are unchanged relative to the results of Table 4.3. All other forecasters, except Naive Last Year, increase in bias with the exclusion of Sundays from the error calculation.

This increase is smallest in absolute terms for the 4 Weeks Median heuristic, increasing its ME from -0.09 to -0.10. Its divergence from the 4 Weeks Mean therefore remains unchanged. The relative rank of the models within the group of negative ME remains identical to the findings in Table 4.3. The bias of the ETS model worsens from -0.05 to -0.14. Prophet's bias increases by a similar absolute magnitude from -0.08 to -0.17. Moirai MoE similarly worsens by 0.09, reaching a ME of -0.25. Contrary to the trend, naive Last Year's bias decreases in magnitude from -4.11 to -2.71. In absolute terms Naive Last Year nonetheless remains the most biased model by a substantial margin.

Three models switch from over-forecasting to under-forecasting the actual demand. Among these, Naive Yesterday shows the largest increase in bias, with a ME of 1.29. The bias of TTM worsens in absolute terms from 0.12 to 1.34 ME. XGBoost shifts from an ME of -0.77 to 1.27, an absolute change of 2.04.

The exclusion of Sundays in summary leads to a heterogeneous increase in bias among the models as well as a change in bias direction for TTM, XGBoost and Naive Yesterday.

Taken together, the WAPE results in Tables 4.1 and 4.2 and the bias results in Tables 4.3 and 4.4 reveal that similar aggregate accuracy can mask fundamentally different bias profiles across models

Beyond aggregate WAPE and bias, model performance is analysed as a function of forecasting difficulty. Forecasting difficulty is defined in this context as the minimum WAPE of a Stock Keeping Unit (SKU) across all models. Figure 4.1 reports the median per-product WAPE difference between the leading model, Moirai MoE, and the best-performing alternative model within each difficulty quantile. Products are grouped into six forecast difficulty quantiles along the x-axis, with positive values indicating lower WAPE for Moirai MoE.

The largest median WAPE advantage for Moirai MoE is observed in the hardest difficulty quantile, where the median difference reaches 0.12 absolute difference. In the subsequent quantiles, the median advantage decreases substantially and remains below 0.02. In the second-easiest quantile, the median WAPE difference becomes negative, indicating that a competing model achieves lower median error. This pattern persists in the easiest quantile, where the competing model's median advantage increases to approximately 0.02 in absolute terms.

This demonstrates that model performance as measured by WAPE of the foundation model Moirai MoE systematically differs from alternative approaches and changes across difficulty quantiles.

To test whether this phenomenon is robust to metric change, the same analysis is repeated with sMASE instead of WAPE in Figure 4.2. Similar to Figure 4.1 the advantage of Moirai MoE over other models also shows itself measured by sMASE. Additionally, this advantage also in general decreases for easier quantiles. Contrary to the previous plot, Q3 shows a small negative advantage. Q4 demonstrates an advantage for both metrics. However, in the case of fig 4.2 the size of this advantage is contrary to the general trend of decreasing advantage and is the largest advantage by absolute value.

This demonstrates that the advantage of Moirai MoE within harder quantiles is robust to a change in metrics.

Figure 4.3 shows the distribution of per-product WAPE across the product portfolio for each forecasting model using box plots. Foundation models are highlighted in blue, while the Prophet model is highlighted in orange.

Substantial differences across model classes are observed in the spread of forecasting performance. Models in the first group exhibit a wide dispersion of WAPE values across products, indicating high variability in per-product accuracy. Although ETS and Prophet show similar median WAPE values, Prophet attains lower WAPE values for a subset of products, with lower-tail values below 0.20. At the same time, Prophet exhibits a pronounced upper tail, with WAPE values exceeding 0.50 for some products, corresponding to substantially higher forecast error than observed for other models in this group.

In contrast, the foundation models display a comparatively narrow WAPE distribution across products. Their interquartile ranges are small, and extreme values are limited relative to classical approaches. Among all models, Moirai MoE exhibits the lowest upper-quartile WAPE, with the 75th percentile at approximately 0.40 and only marginally lower error at the 25th percentile.

Across all models, the relative ordering of median WAPE remains consistent across the quantiles shown in the box plots.

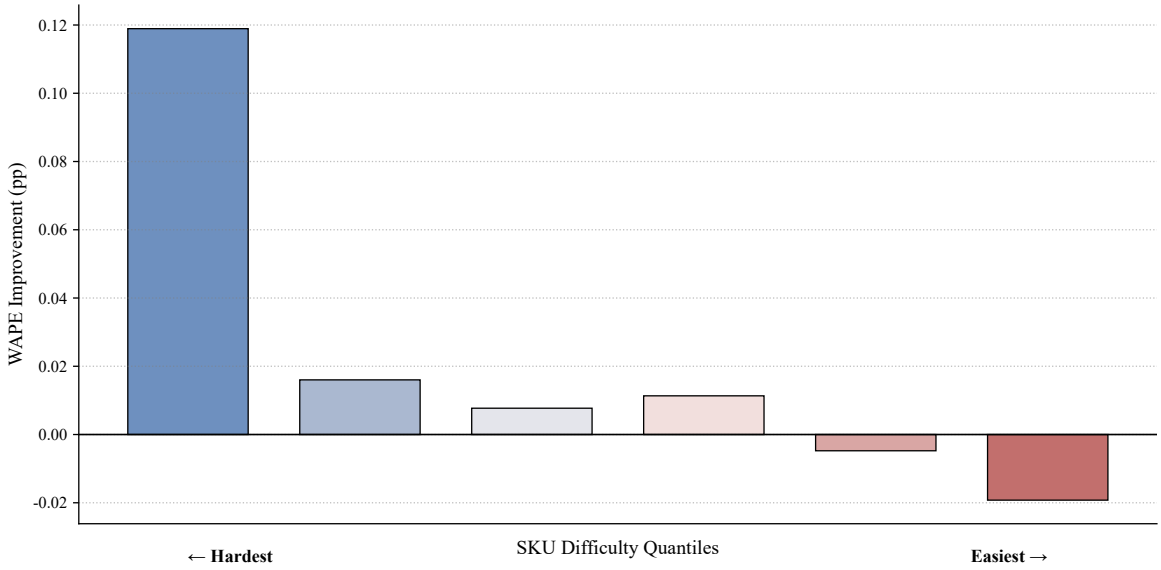


Figure 4.1: *WAPE advantage of Moirai MoE over the best competing model across SKU difficulty quantiles. Positive values indicate improved accuracy.*

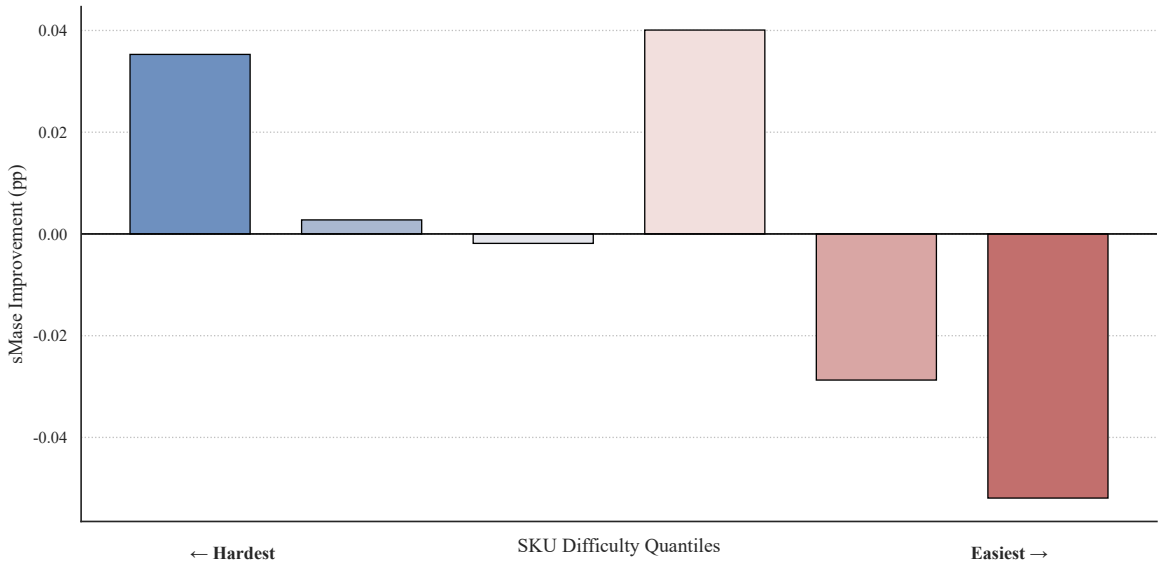


Figure 4.2: *sMASE advantage of Moirai MoE over the best competing model across SKU difficulty quantiles. Positive values indicate improved accuracy.*

Within the context of the research question this demonstrates that there is a spread of model performance across the product portfolio and the full error distribution systematically differs across models beyond their aggregate median error rate.

The IQR visible in the previous Figure 4.3 is explicitly plotted in Figure 4.4 against the median WAPE of each model across all products to better show the tradeoff between accuracy and stability. TTM is highlighted in blue. In general, a negative trend can be observed between IQR and model accuracy.

The foundation models together with the Naive Yesterday heuristic, and XGBoost demonstrate the lowest performance spread as measured by IQR. With a value lower than 0.06 absolute difference in WAPE, TTM has the lowest IQR of all observed models. Within this group both foundation models have a substantially lower relative WAPE compared to the other forecasting approaches.

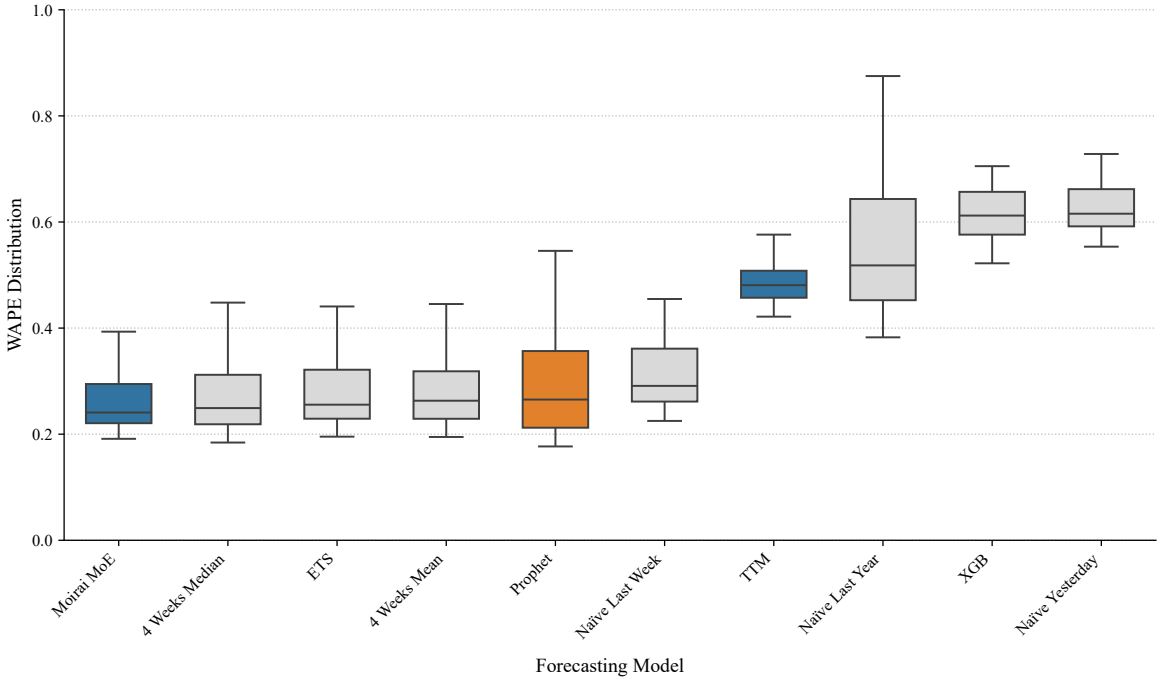


Figure 4.3: *Distribution of WAPE scores per product over all 48 products per model. Foundation models and Prophet highlighted*

Moirai MoE especially, is the closest model to the origin of the plot, with an IQR of under 0.08 and a WAPE of under 0.25.

Another lens of analysis is to look at the relationship between sales volume and forecasting quality of all model classes. Figure 4.5 shows the performance with regard to WAPE of the models binned into 5 quantiles according to sales volume. Moirai MoE is highlighted in blue and Prophet is highlighted in orange.

All models improve their error rate with increasing sales volume of a product. This change in WAPE is most prominent with Prophet. This model is the worst performing forecaster in quantile with the lowest volume and ties for first place together with the 4 Weeks Median in the quantile with the highest volume. The majority of improvement can be observed going from the first quantile to the second. This phenomenon generally holds for all forecasting approaches. For Moirai MoE this decrease is the smallest in absolute terms. With the exception of XGBoost, whose improvement from quantile one to quantile two is slower than the rest of the models and Prophet for the first two quantiles , the relative change in WAPE between the quantiles does not differentiate itself between the model classes.

The relative ranking of each model follows the aggregate performance observed in 4.1. Outside of Prophet, the clustering of model performances remains stable.

Within the context of RQ1 this demonstrates how WAPE is influenced by different strata, such as sales volume and that this influence is homogenous between the forecasting approaches.

To assess whether the volume-dependent improvement observed in Figure 4.5 reflects genuine forecasting gains or is partly a metric artefact driven by WAPE's sensitivity to low actual values, Figure 4.6 reports the median sMASE across the same volume quantiles. As sMASE normalizes forecast error against the in-sample naive baseline, it controls for the scale effect introduced by

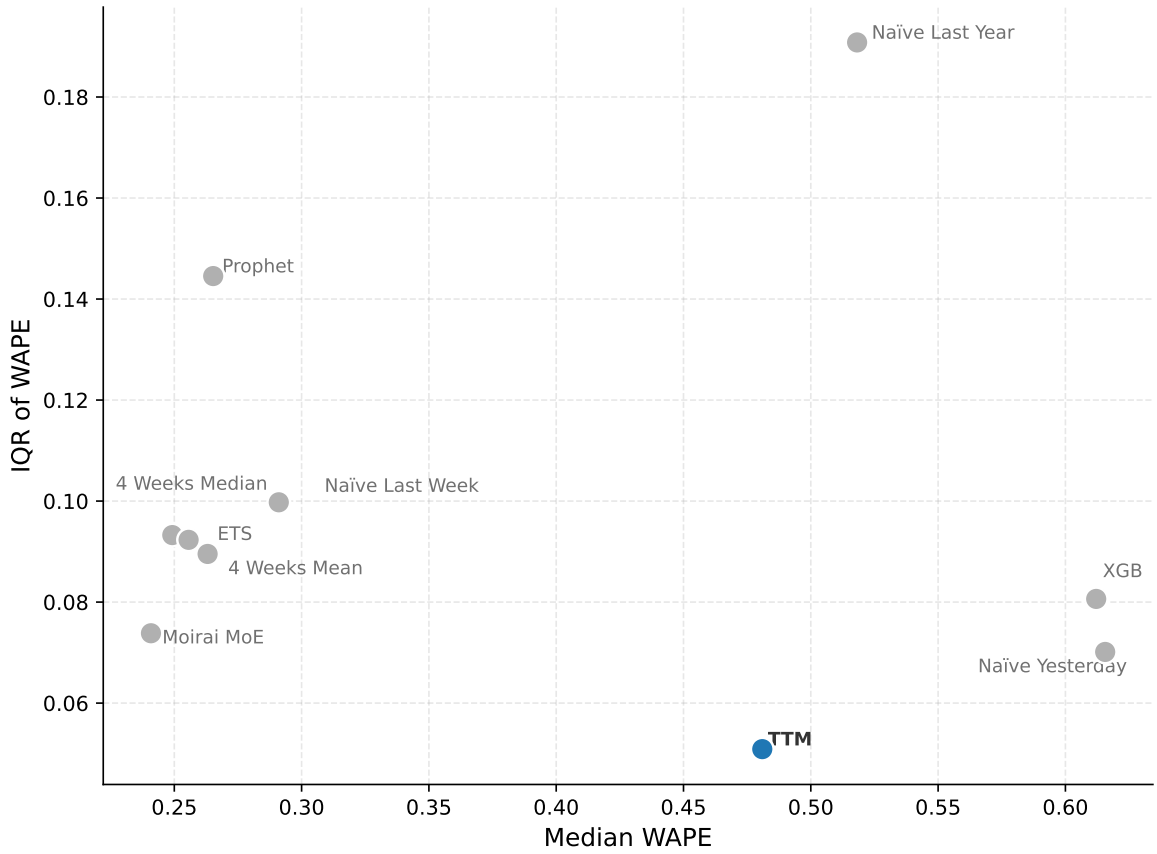


Figure 4.4: *IQR per WAPE score–TTM highlighted*

varying sales volumes and provides a complementary perspective on relative model competitiveness. The sMASE results reveal additional insight. compared to the WAPE volume trend. The cluster of top-performing models — Moirai MoE, Prophet, 4 Weeks Median, 4 Weeks Mean, and ETS — maintains a stable sMASE of approximately 0.75 to 1.00 across all five volume quantiles, indicating that their advantage over the naive baseline is consistent regardless of product volume. The relative model performance trajectories among the models in this cluster remain robust even after the metric change coinciding with the observations made from Figure 4.5 In contrast, the weaker models — Naive Yesterday, XGBoost, naive Last Year, TTM, and Naive Last Week — show a flat or increasing sMASE trend with rising volume, meaning their performance relative to the naive benchmark does not improve and in some cases deteriorates at higher volumes.

Within the context of RQ1 this demonstrates that the competitive advantage of the top-performing models over the naive baseline is robust to changes in product sales volume and that this phenomenon is consistent even when a different metric is chosen.

RQ1 asks how foundation models compare to classical approaches in the domain of next-day demand prediction on a real-life bakery dataset. The foundation model Moirai MoE achieves the lowest aggregated WAPE of 0.24. Followed by 4 Weeks Median, ETS and Prophet, who trail the top performance of Moirai MoE by 0.03 difference in WAPE at most. In contrast, the other observed foundation model TTM together with XGBoost demonstrate a substantially higher error rate.

The forecasting quality is impacted by the inclusion of structural zero demand days, which decreases

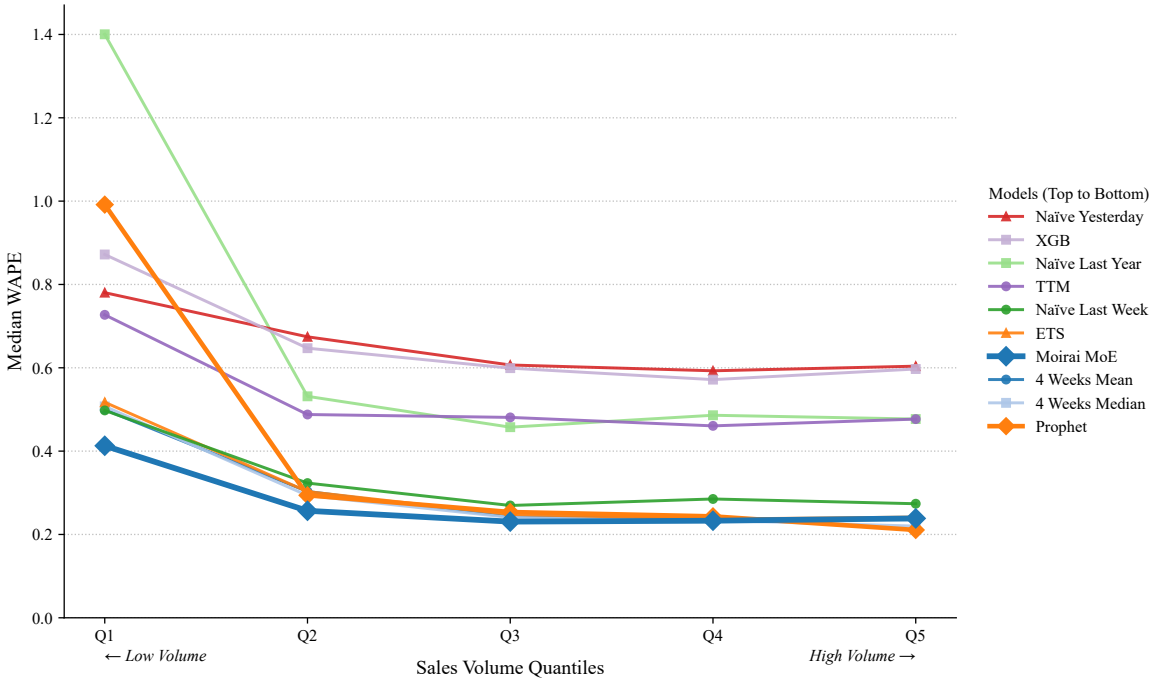


Figure 4.5: *Relationship of WAPE to Sales Volume over Sales Volume quantiles. Smaller values indicate improved accuracy. Moirai MoE and Prophet highlighted.*

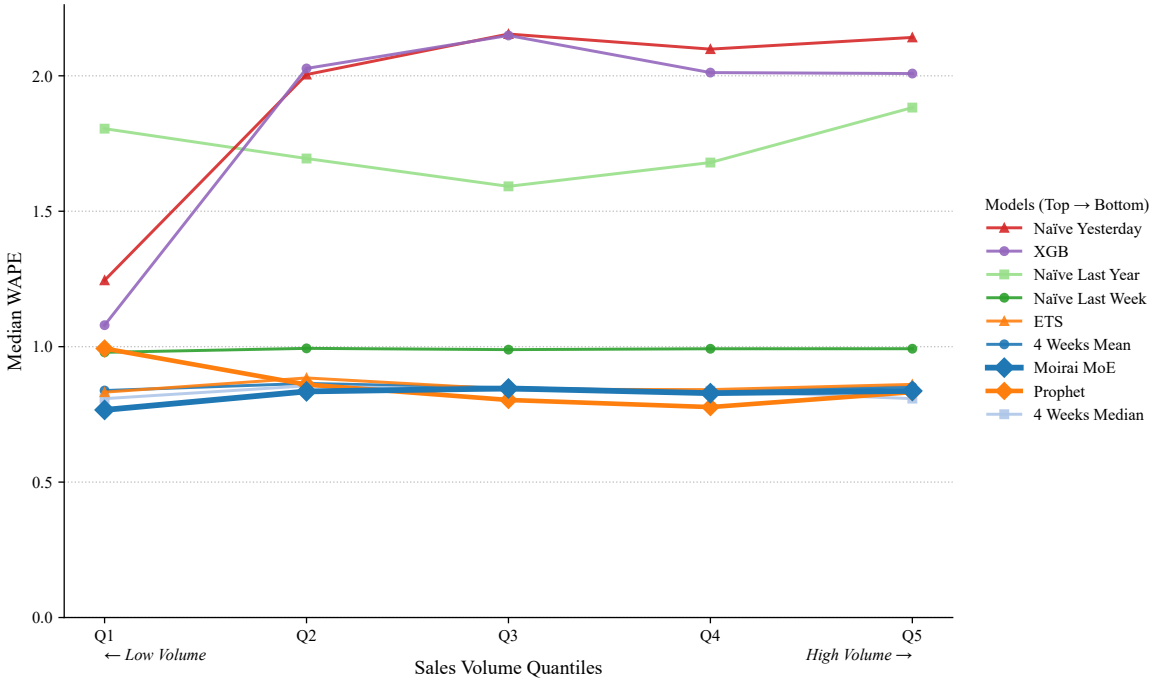


Figure 4.6: *Relationship of sMASE to Sales Volume over Sales Volume quantiles. Smaller values indicate improved accuracy. Moirai MoE and Prophet highlighted.*

model performance among all models an effect which is especially pronounced among the worse performing models.

Distributional analyses revealed that models with similar aggregate accuracy displayed different per-product error spreads and bias profiles. While Moirai MoE and TTM exhibited a systematic tendency to over-forecast relative to neutral baselines, they showed a comparatively narrow IQR of WAPE across SKUs, whereas Prophet combined low errors for some products with a pronounced

upper tail of large errors.

Stratified evaluation indicated that Moirai MoE's relative advantage increased for higher forecast-difficulty products, while performance differences between methods diminished for easier items. All models demonstrated lower WAPE for higher-volume products, with largely unchanged relative rankings across volume quantiles.

4.2 Impact of Exogenous Variables on Prophet Forecast Accuracy

This section evaluates the impact of exogenous variables—specifically weather and calendar effects—on the forecasting performance of the Prophet model across products. The variables are grouped into the feature sets BASE, WEATHER, and HOLIDAY, as defined in Chapter 3.5, and assessed using an ablation study.

Figure 4.7 illustrates the change in WAPE for each product relative to the base model after adding either weather or holiday features. Positive values on the y-axis indicate a deterioration in performance and are shown in red, whereas negative values indicate an improvement in WAPE and are shown in green. A box plot is overlaid to summarize the distribution of performance changes.

Table 4.5 reports the mean and median WAPE differences as well as the proportion of products whose forecasts improved after including the respective feature group.

Both covariate groups lead, on average, to a deterioration in forecast accuracy compared to the baseline model. The HOLIDAY feature group exhibits a larger dispersion in performance changes, indicating greater variability in its effect across products. Consequently, both the largest improvements and the largest degradations are observed when holiday features are included.

In both cases, the median change in WAPE remains below 0.02, suggesting that the typical effect size is small. Approximately one third of the products benefit from incorporating weather information, whereas only about one quarter show improvement when holiday effects are added. For the WEATHER feature group, the mean change in WAPE exceeds the median in absolute magnitude, indicating a skewed distribution in which a subset of products experiences comparatively large degradations.

Statistical testing confirms that the observed differences are unlikely to be due to random variation. Wilcoxon signed-rank tests yield p-values of 0.004433 for WEATHER and 0.001299 for HOLIDAY, leading to rejection of the null hypothesis of equal predictive performance relative to the baseline model.

While the previous analysis evaluates the isolated contribution of each feature group, the following comparison assesses their combined effect when all exogenous variables are included simultaneously. Figure 4.8 demonstrates the value add of all exogenous factors by plotting the WAPE per product of Prophet including only the features in the BASE feature group on the x-axis and the per-product accuracy of the model including all features of both weather and holiday information. A vertical dashed line is included to indicate model performance not changing after the inclusion of the covariates. The area which denotes forecasting improvement is colored green, whereas the area of the plot indicating a deterioration in accuracy is colored red. Within the scope of the following

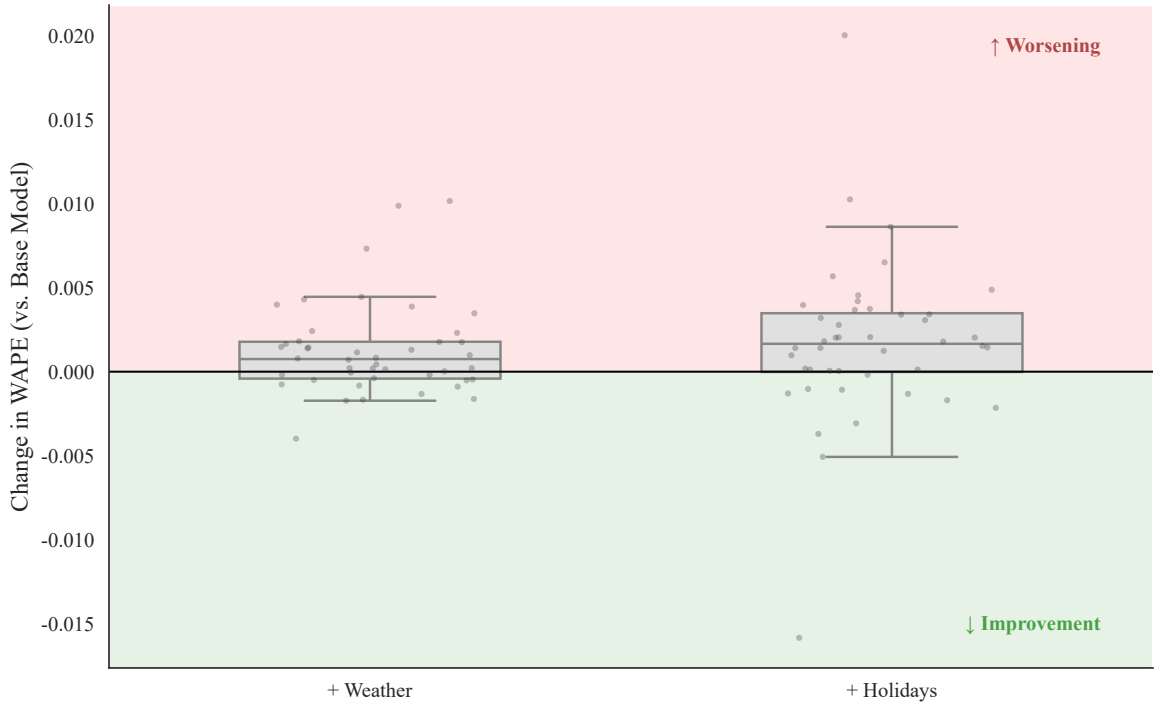


Figure 4.7: *Distribution of WAPE with and without covariates*

analysis the accuracy refers to the WAPE of the base model, as opposed to the WAPE of the full model.

It can be observed, that no SKU falls within the green area. The impact of the covariates is heterogeneous across the product portfolio. Most products cluster near the origin of the plot, where the error rate is below 0.5 WAPE, with some products having an unchanged WAPE after the inclusion of the exogenous information. With increasing WAPE the negative impact grows larger. Beyond 0.5 WAPE no SKU can be found on the dashed line. The largest change in performance is found between a WAPE of 1.0 and 1.2 relative to base model accuracy. The exception to the large deterioration in model quality for difficult to forecast SKUs are the two Products with the largest WAPE, which decrease only slightly in error rate compared to the base model.

Subsequently, the impact of weather and holiday covariates can be summarized as overall negative. The median change in error rate as well as the spread of error rate change is larger for holiday features. For a third of SKUs in the case of weather and a fourth in the case of holidays a moderate increase in forecasting quality can be observed. Looking at the inclusion of exogenous features in general, no improvement of forecast accuracy can be observed. The negative change in WAPE increases with forecasting difficulty up to a point.

4.3 Effect of Finetuning on TTM Forecast Performance

The third section investigates the change in forecasting performance between the foundation model TTM in zero-shot and after finetuning the model on the retail data.

Figure 4.9 compares the distributions of WAPE values across products for the TTM model in zero-shot mode and after finetuning. The finetuned model is shown in blue, and the mean WAPE is indicated by a black star.

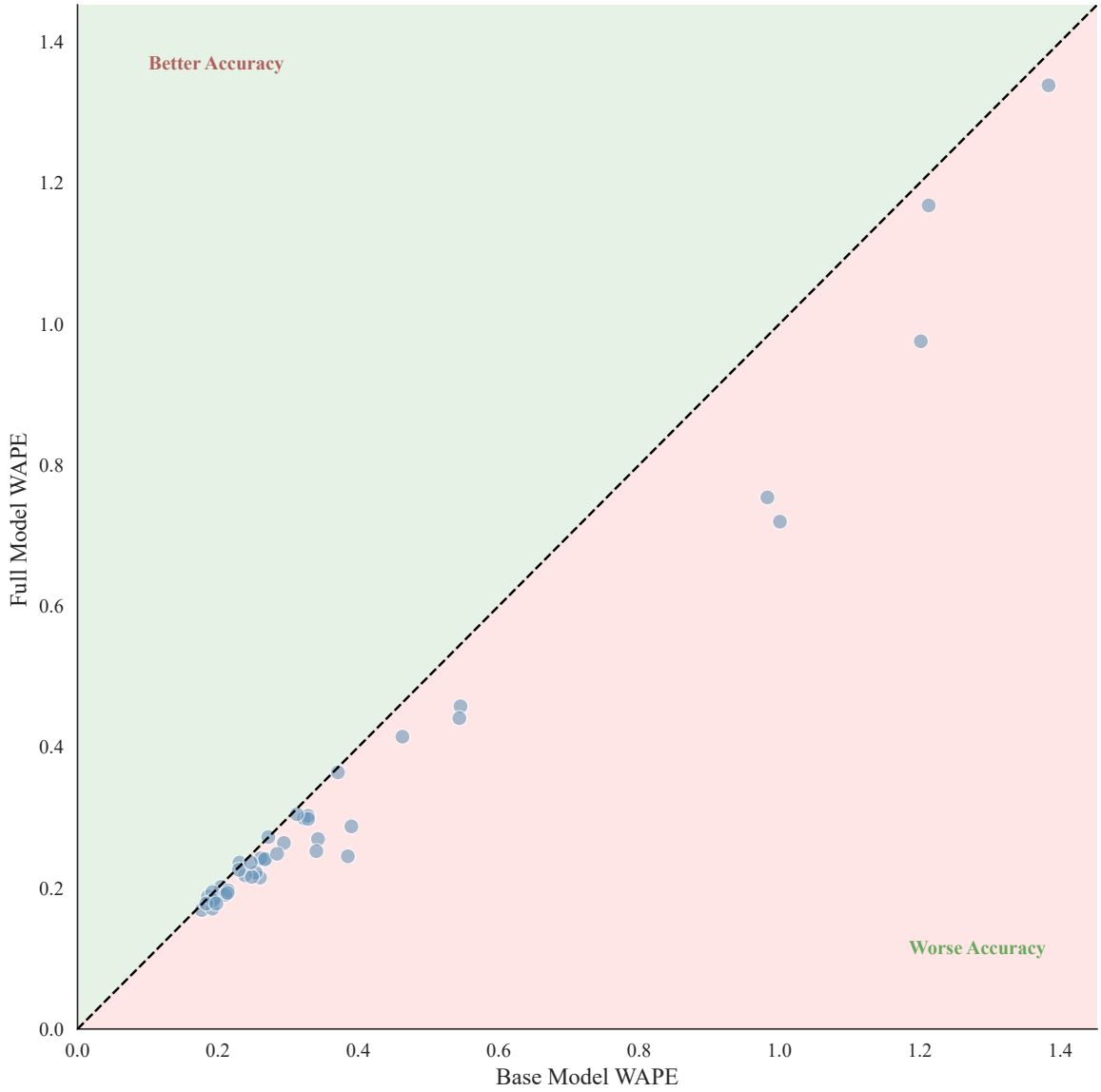


Figure 4.8: *Distribution of WAPE with both covariates combined compared to no covariates*

As shown in the box-plot, the distribution shifts downward after finetuning, indicating improved forecasting accuracy overall. The median WAPE decreases. In both configurations, the mean exceeds the median, indicating right-skewed error distributions caused by a small number of high-error products. However, the smaller gap between mean and median after finetuning suggests a reduction in extreme errors. The improvement is not uniform across products, as reflected in the subtle change in distribution shape and the differing shifts of individual observations.

This demonstrates that finetuning has an effect on the overall performance of TTM and indicates that finetuning improves model quality.

To further analyze the heterogeneous distribution of performance change through finetuning, the WAPE of the zero-shot model per product is directly compared to the WAPE of the finetuned TTM in figure 4.12.

The change in forecasting performance is not uniformly beneficial. The next-day prediction degrades for two of the products, while it improves for all other observed products. No obvious linear relationship can be detected between the size of the error and the improvement due to

Table 4.5: Impact of external covariates on Prophet forecast accuracy. Positive Δ values indicate a worsening of the model (higher error). 'Share Improved' denotes the percentage of SKUs where the error decreased.

Variant (vs. Base)	Median Δ WAPE	Mean Δ WAPE	Share Improved (%)
+ Weather	0.0008	0.0013	34.1%
+ Holidays	0.0017	0.0017	25.0%

Note: Δ WAPE = (Variant WAPE - Base WAPE). A share of $< 50\%$ implies the covariate harmed accuracy for the majority of products.

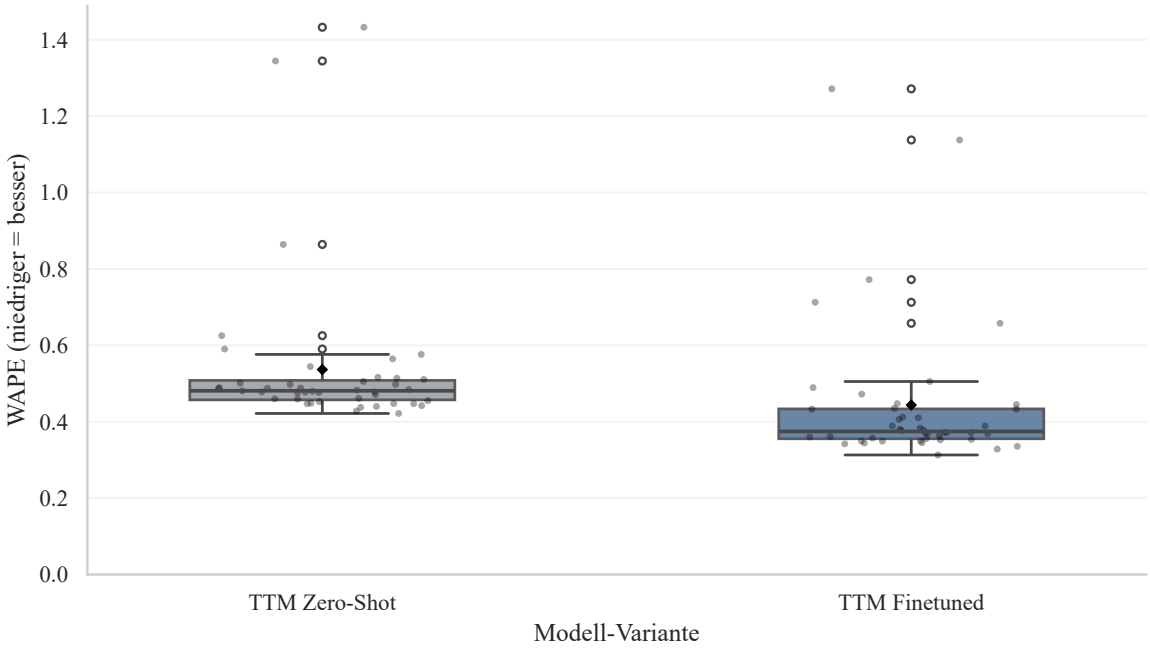


Figure 4.9: *Distribution of WAPE scores for TTM before and after finetuning*

finetuning.

While the effect of finetuning on model performance is neither uniform in size nor direction, most deviations cluster around the mean change of 0.10 WAPE difference between zero-shot and finetuned, with only 10 % of all cases changing by more than 0.13 WAPE in absolute terms. The maximum amount of absolute WAPE difference is capped at 0.20. The median WAPE delta is only slightly higher than the mean WAPE change with an improvement of 0.09 in WAPE.

This change in model performance, as well as the reduction in error rate through finetuning is statistically highly significant with p-values of under 0.0001 both in both cases.

Figure 4.10 revisits the analysis of the relationship between sales volume quantiles and aggregated WAPE from section 4.1 Figure 4.5. The finetuned variant of TTM is highlighted in blue, the base TTM is coloured orange. TTM (Finetuned) consistently outperforms TTM (Zero-shot) across all volume bins. For the quantiles ranging from Q2 to Q5 the distance in median WAPE remains stable. In contrast to all other models –including the zero-shot variant of TTM – , the error rate of the finetuned TTM does not decrease from Q1 to Q2. The WAPE value within the lowest volume quantile is the lowest value among all observed models tied with Moirai MoE.

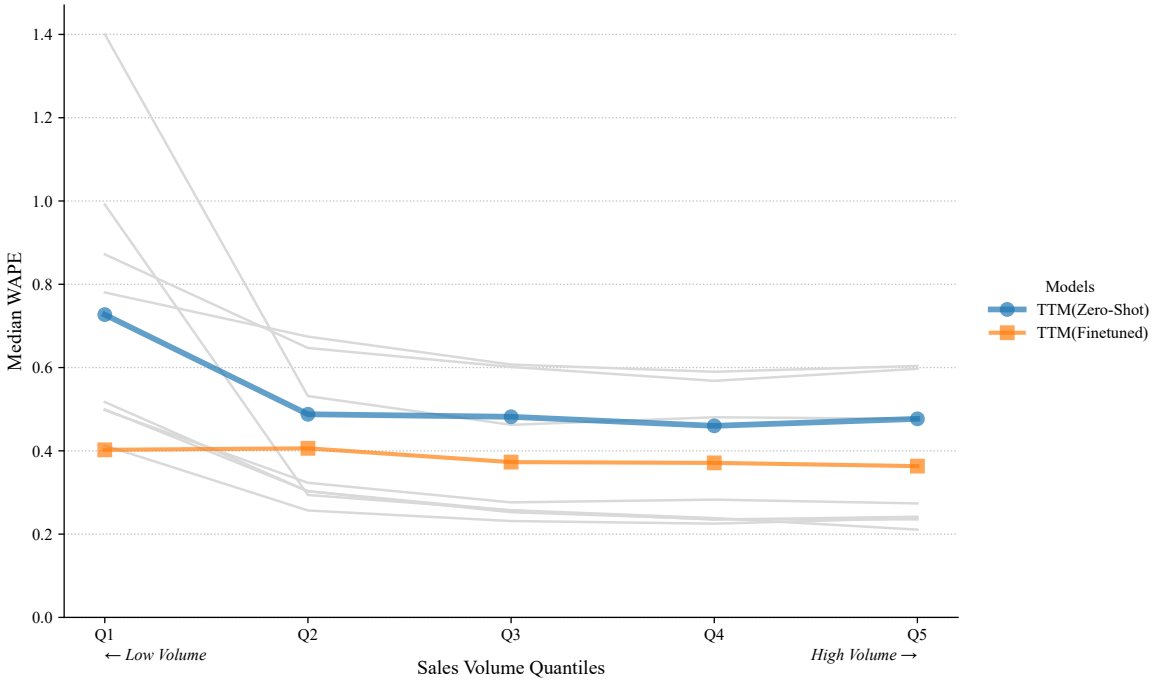


Figure 4.10: *WAPE scores over volume quantiles. TTM Zero-shot and TTM Finetuned highlighted*

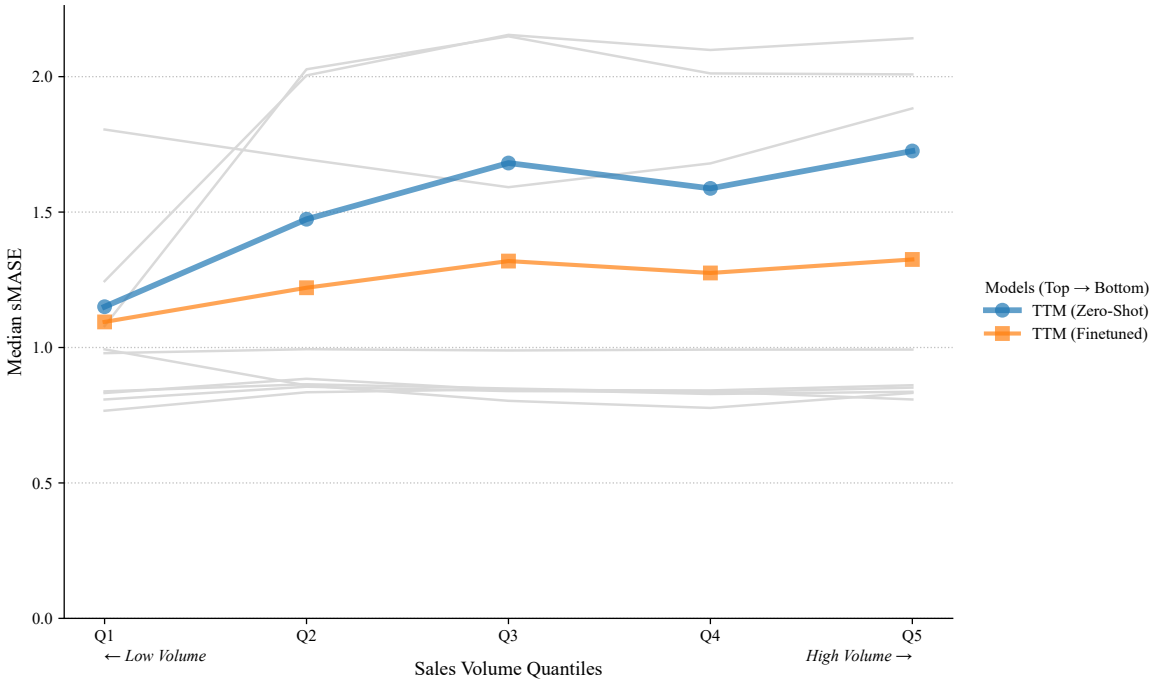


Figure 4.11: *Relationship of sMASE to Sales Volume over Sales Volume quantiles. Smaller values indicate improved accuracy. Finetuned vs. Zero-Shot highlighted.*

To verify that the stratified effect of finetuning on model performance is robust to metric artefacts the same relationship is plotted in figure 4.11 with sMASE instead of WAPE. The finetuned error curve highlighted in orange is consistently lower than the zero-shot variant colored in blue across all sales quantiles. The difference between the two models is constant across all observed bins. Outside of Q1 the sMASE of the finetuned model is substantially over 1, indicating a worse performance in comparison to the reference baseline of the naive weekly seasonality.

Despite these improvements, the relative, aggregated ranking of the models found in section 4.1

does not change as seen in Table 4.6 with TTM (Finetuned) highlighted in bold. The median WAPE of the finetuned model with a value of 0.37 lies equally as far from the 0.27 WAPE of Prophet as it lies from the Zero-shot foundation model with a median WAPE of 0.48.

Without the inclusion of Sundays this value improves slightly from 0.37 to 0.34. Finetuning increases the tendency of TTM to over-forecast observed in 4.3 substantially from -0.12 to a bias of -1.47. Excluding Sundays from the computation of the mean error, the direction of the bias flips and increases in absolute terms in comparison to the zero-shot model from 1.34 to 1.83.

Taken together, these results demonstrate that finetuning substantially improves the performance of the TTM foundation model, and reduces extreme errors, but does not fully close the performance gap to the strongest classical approaches. Bias increases with finetuning.

Table 4.6: Impact of finetuning on TTM forecast accuracy. Positive Δ values indicate an improvement of the model. 'Share Improved' denotes the percentage of SKUs where the error decreased.

Finetuning vs. Zero-shot	**Median Δ WAPE**	**Mean Δ WAPE**	**Share Improved (%)**
Finetuning	0.1027	0.0942	95.4%

Note: Δ WAPE = (Zero-Shot TTM WAPE- Finetuned TTM WAPE). A share of $> 50\%$ implies the finetuning benefitted accuracy for the majority of products.

Table 4.7: Performance comparison of all forecasting models, including finetuned variants. Models are sorted by WAPE.

Model Name	**WAPE**
Moirai MoE	0.24
4 Weeks Median	0.25
4 Weeks Mean	0.26
ETS	0.26
Prophet	0.27
Naive Last Week	0.29
TTM (Finetuned)	**0.37**
TTM (Zero-Shot)	0.48
Naive Last Year	0.52
Naive Yesterday	0.62
XGBoost	0.65

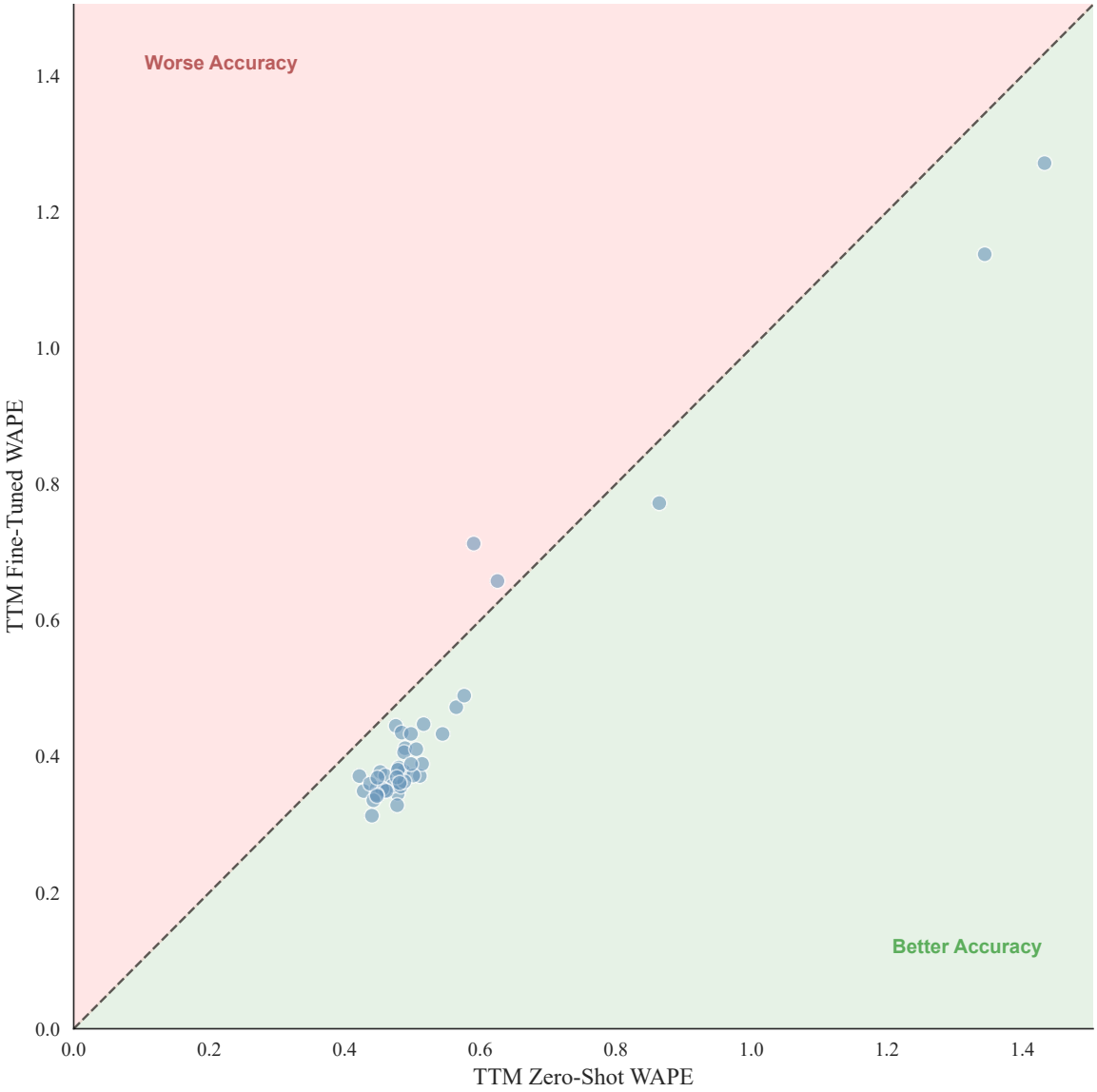


Figure 4.12: *WAPE scores for TTM before and after finetuning. Direct comparison per product*

5

Discussion

This chapter discusses the results from Chapter4 and compares them to results in the literature. The study compared time series foundation models to classical forecasting methods for the task of next-day demand forecasting in the context of an organic bakery, and further examined the role of exogenous covariates and finetuning on model performance.

5.1 Comparative Forecasting Performance Across Models

The strong performance of the 4 Weeks Median heuristic and the statistical model ETS in Table 4.1 suggests that demand in this context is largely driven by stable seasonal structure. Further evidence comes from the near-identical forecasting performance of the top four models on non-Sunday days. As shown in Table 4.2, forecasting performance on non-Sunday days is almost identical among the top four models. Since demand on Sundays is structurally zero, including these observations affects aggregate error metrics and the estimated bias; in particular, models that fail to predict exact zeros exhibit systematic over-forecast bias.

Within the broader design objective that this empirical study is part of, over- and under-forecast bias lead directly to waste or stockouts within the context of perishable bakery goods production. Operationally, the Sunday over-forecast bias does not result in waste because the forecasting system is not used for production planning on closed days. When only operational days are considered, TTM and XGBoost exhibit a tendency to under-forecast demand, which could lead to stockouts and lost sales. In contrast, ETS, Prophet, and Moirai MoE systematically over-forecast demand, potentially resulting in moderate waste of perishable goods.

The convergence of performance across forecasting approaches with fundamentally different inductive biases, including a simple heuristic such as 4 Weeks Median/Mean, is consistent with diminishing returns from additional model complexity in this dataset, implying that much of the extractable predictive signal is captured by seasonal structure. Such findings echo skepticism toward complex Transformer-based LTSF models expressed in Zeng et al. (2023). As an empirical study intended to inform a subsequent design and evaluation cycle, these findings indicate that model selection should emphasize operational factors such as interpretability, stability, computational cost, and the acceptable balance between waste and stockouts, rather than aggregate accuracy alone.

Moirai MoE's relative advantage is concentrated in the lowest sales-volume quantile, when measured by WAPE. Together with the relatively small shift in Moirai MoE's performance after excluding Sundays, this pattern is consistent with the hypothesis that differences between models are partly driven by how well intermittent and zero-demand regimes are handled, rather than uniformly improved accuracy across all demand levels. These structural challenges of intermittent demand predictions are consistent with Fildes et al. (2022).

The comparatively weak performance of XGBoost may be explained by the fact that tree-based models do not inherently encode temporal dependence. They rely on engineered features to represent temporal patterns and seasonality. However, their inability to capture the regular zero-demand pattern on Sundays remains notable. Possible explanations include insufficient representation of calendar structure in the feature set, limited generalization of learned decision rules, or suboptimal hyperparameter configuration.

Alone, however, none of these fully explains the failure to learn a clear calendar effect. Together with the findings from RQ2, one plausible explanation is that the dataset is too small to reliably learn the relevant temporal patterns.

To investigate whether the results were specific to XGBoost's implementation, the setup was replicated using LightGBM, which produced nearly identical results. The near-identical outcomes suggest that the shared structural bias of tree-based models, rather than implementation-specific factors, is the primary driver of the observed underperformance. Qualitative reports from informal discussions with stakeholders involved in prior forecasting projects describe similar challenges when deploying XGBoost on comparable sales data from the same company. While no historical error metrics from those empirical comparisons are available, these observations partly corroborate the quantitative findings of this study.

The underperformance of XGBoost in this study stands in contrast to Wikamulia et al. (2023), who report strong predictive accuracy for XGBoost in a bakery demand forecasting context. However, the data used in the aforementioned study covers a substantially larger time-frame in comparison, includes information about multiple, regionally diverse locations and the most influential features, such as transaction quantities and grand totals, suggest that target leakage may partly explain the forecasting results. These differences suggest that the strong results reported by Wikamulia et al. (2023) may not be directly transferable to the empirical results of this study.

The divergence between mean- and median-based 4-week baselines, together with the sales concentration shown in Figure 3.4, is consistent with a right-skewed demand distribution with a heavy tail, where mean-based aggregation is more sensitive to high-volume outlier days.

The comparatively low IQR of the foundation models (despite their differing architectures) indicates more uniform performance across SKUs than several classical approaches exhibit. The pattern may reflect the transfer of product-independent temporal structure learned during pretraining, though alternative explanations such as a structural bias towards smoother or more conservative predictions cannot be ruled out. From a practical standpoint, stable performance across products with a robust lower error bound is arguably more attractive than a marginally lower aggregate WAPE, particularly when differences fall within 1 percentage point. Finally, the stronger performance of Moirai MoE relative to TTM is consistent with the relevance of domain-appropriate pretraining data, suggesting

that transfer from generic time-series repositories may be limited when retail-specific demand characteristics are absent.

The divergence between WAPE and sMASE suggests that the apparent WAPE improvement for weaker models at higher volume quantiles is at least partly attributable to the denominator scaling of WAPE due to increasing sales volume rather than a genuine reduction in forecast difficulty or model superiority. Because WAPE expresses absolute error relative to total sales volume, higher-volume strata mechanically reduce WAPE even when absolute errors remain large, making otherwise weak models appear more competitive. The sMASE analysis corrects for this and reveals that the relative ranking between model classes is more stable across volume strata than the WAPE trend implies, with the performance gap between the top cluster and weaker approaches persisting across all quantiles.

The outlier observed for Moirai MoE in Q3 of Figure 4.2, which runs counter to the otherwise diminishing advantage at higher forecasting difficulty, may be partly explained as a measurement artefact. Because sMASE scales error relative to the seasonal naive baseline, unusually poor baseline performance for the median product in this bin reduces the denominator and inflates the scaled error differential. The apparent advantage in this quantile may therefore reflect baseline instability rather than a substantive improvement in forecast accuracy.

Similarly, the apparent divergence in TTM (Zero-Shot) performance between Figures 4.11 and 4.10 may reflect a shift in the naive baseline rather than a deterioration in TTM's absolute accuracy. If the seasonal naive baseline produces lower errors in higher-volume strata for the finetuned comparison, the sMASE denominator shrinks and the scaled gap widens, irrespective of changes in TTM's actual forecasts. Together, these observations reinforce that metric choice and baseline behaviour interact with demand structure in ways that can obscure genuine model differences, and that no single metric should be interpreted in isolation.

Across the results in RQ1 the general trend of heterogeneity holds. No one model is obviously superior to alternative forecasting approaches across all dimensions of analysis and all strata. This implies that a forecasting solution within this case context could benefit from a hybrid architecture combining multiple models for different product and demand regime segments. Hybridization as a beneficial forecasting approach is consistent with results from forecasting benchmarks, such as M4 (Makridakis et al., 2020).

Additionally, simple forecasting methods relying on transparent baselines and heuristics have shown competitive results, with small bias, small dispersion of error, low computational costs and high interpretability for both developer and end-user. Such results suggest they may represent a well-justified default for a future forecasting system within this context. Notably, while forecasting competitions such as M4 (Makridakis et al., 2020) also document the competitiveness of simple approaches, the effect appears stronger within this case context than in those broader benchmarks.

While complex models do not substantially outperform simple seasonal heuristics in this context, this empirical study makes a strong case for automation in demand planning within the broader design cycle. An automated implementation of a simple baseline addresses the auditability, time cost, and key-person dependency weaknesses identified in Chapter 1, without requiring the computational overhead of foundation models.

5.2 Impact of Exogenous Variables on Prophet Forecast Accuracy

Overall, exogenous factors do not improve forecasting quality, as shown in Figure 4.8. The finding holds for both weather-related and holiday features, and contrasts with stakeholder expectations ('Brotsüchtig – Digitale Lebensmittelrettung', 2024) as well as the broader literature, where both factors have been associated with changes in retail demand (Huber & Stuckenschmidt, 2020; Rose & Dolega, 2022). However, the studies most relevant to this comparison differ substantially in scale, product portfolio breadth, aggregation level, and geographic context. Rose and Dolega (2022) and C.-L. Yang and Sutrisno (2018) operate at a scale and aggregation level where weather effects are more likely to manifest as detectable signal, and the latter focuses on intra-day demand evolution rather than next-day prediction. Findings from these contexts may therefore not transfer to this empirical study.

One primary explanation is that demand structure in this context is dominated by stable weekly seasonality, which leaves limited explanatory residual variance for exogenous factors. Beyond this, the features as currently implemented may not sufficiently capture the potentially non-linear relationship between weather or holidays and demand. The additive regressor structure within Prophet is a further constraint, as it may be unable to model non-linear interactions between covariates and baseline demand levels. However, the poor forecasting performance of XGBoost, which is not subject to the additive structural constraint of Prophet and can in principle capture non-linear covariate relationships, indirectly suggests that the size of this non-linear effect is limited in comparison to the seasonal signal.

For operational usage in a bakery with no dedicated IT team and limited capacity for exhaustive feature engineering, these findings suggest that the complexity cost of covariate integration may outweigh the marginal forecasting benefit in this context.

5.3 Effect of Finetuning on TTM Forecast Performance

Finetuning substantially improves TTM forecast accuracy, a result that goes beyond the conditional improvements reported in prior benchmarking and fine-tuning work, which suggests that adaptation gains are heterogeneous across settings and not guaranteed (Li et al., 2025; Liang et al., 2024; Qiao et al., 2025). The pattern suggests that finetuning primarily adapts the model to store-level demand structure and enables exploitation of cross-series information, rather than learning highly product-specific patterns.

Support for this interpretation comes from the strong performance of TTM (Finetuned) in the lowest sales-volume quantile (Figure 4.5). Low-volume items are difficult to forecast in isolation due to intermittency and noise, but shared temporal structure across products can provide informative signals when learned jointly.

The trend-defying increase in WAPE from the lowest to the second lowest quantile is consistent with Moirai MoE also improving the least, suggesting that those models, which both are excellent with intermittent demand, already extracted all possible information compared to alternative approaches, which seem to need more sales volume to find the patterns.

The remaining performance gap relative to the top-performing models may reflect both limited

data available for finetuning and the influence of pretraining-induced inductive biases, which can constrain adaptation to highly regular seasonal patterns already well captured by classical methods.

Notably, finetuning increases the magnitude of forecast bias compared to the zero-shot configuration. The increase suggests the adapted model produces more systematic predictions rather than near-neutral forecasts, likely reflecting learned structural patterns in the training data. The observation is consistent with cautions raised by Qiao et al. (2025) regarding overfitting risk during finetuning on small datasets. By contrast, baseline methods exhibit near-zero bias not because they are better calibrated in a learned sense, but because they reflect aleatoric demand noise through simple rules that do not encode systematic directional errors.

6

Conclusion

This chapter synthesizes the main findings across the research questions evaluated in Chapter 4 and discussed in Chapter 5. Subsequently, limitations of the analysis and its conclusions are brought up and suggestions for further work are made.

6.1 Summary

The objective of this study was to evaluate how TSFMs compare with classical forecasting approaches beyond aggregate accuracy metrics in the task of next-day retail demand forecasting within the context of an organic bakery. Additionally, the impact of exogenous variables (weather and holidays) and the effect of model finetuning were examined. Demand structure within this context is dominated by stable weekly seasonality, with forecasting difficulty decreasing as sales volume increases across most models. Multiple approaches achieve similar aggregate WAPE values, suggesting diminishing returns from additional model complexity and indicating that no single approach is universally superior on this dataset. Top-performing models span a wide range of complexity, from simple heuristics and statistical methods such as ETS and Prophet to the foundation model Moirai MoE.

Despite similar aggregate performance, more detailed analysis reveals meaningful differentiation between model classes. Foundation models and heuristic approaches exhibit lower error dispersion across products, indicating more consistent performance, while learning-based models show greater variability in forecast error. A major differentiator of model quality among top-performing models was the adaptability to structural zero demand days. Simple baseline methods maintain near-neutral bias, whereas models that learn from the training data introduce systematic directional errors. Moirai MoE demonstrates particular strength for low-volume and intermittent demand products, and remains robust to the inclusion of structural zero-demand days, suggesting beneficial inductive biases derived from pretraining.

Given the perishable nature of bakery goods, bias and error dispersion carry direct operational relevance. Systematic over-forecasting translates to overproduction and waste, while systematic under-forecasting increases the risk of stockouts and lost sales, while high dispersion implies single-day losses that cannot be recovered within the planning horizon. The lower dispersion of foundation models and heuristic approaches, alongside the near-neutral bias of simple baselines, therefore suggests operational advantages that aggregate accuracy metrics alone would not reveal.

The observed heterogeneity across different strata and different product segments suggests that a hybridization strategy based on SKU-level segmentation may be more beneficial than a single global forecasting approach. These findings highlight the importance of evaluating forecasting models beyond aggregate metrics, as distributional behaviour, robustness, and bias characteristics can be operationally more relevant than marginal differences in mean error.

Features engineered beyond basic temporal structure have no consistent positive cross-product effect on Prophet's forecasting performance of Prophet. This further strengthens the hypothesis that weekly structure dominates the forecasting signal. This counter-intuitive finding highlights that extensive feature engineering might have diminishing returns on model quality and may thus be operationally undesirable. At the SKU level, there seem to be some products which improve due to the inclusion of some of the exogenous factors, further suggesting product segmentation may lead to increased forecasting quality.

Finetuning improves the performance of TTM on the dataset across almost all products. This improvement is especially strong for low-volume products relative to alternative approaches. The bias of the model ,however, increases at the same time. Despite the improvement in aggregated WAPE, TTM cannot close the gap to the high-performing models, which may be due to a strong, detrimental inductive bias and limited training data. Taken together, these findings suggest that demand structure and seasonality are more important than model sophistication in explaining forecasting performance within this context. Some foundation models offer competitive results and are particularly well-suited to sparse and intermittent demand. The absence of consistent covariate effects underlines that feature engineering does not automatically translate into improved forecast quality. Finetuning improves aggregate accuracy but introduces systematic bias and is insufficient to elevate TTM into the group of top-performing models. Collectively, these results support a context-aware forecasting strategy that prioritises model simplicity, operational interpretability, and demand-segment-level differentiation over marginal gains in aggregate accuracy.

6.2 Limitations

The findings of this study are bounded by its empirical scope and should thus be carefully interpreted within this context. The analysis is based on data from a single location of a local organic bakery chain and a time span of 3.5 years, which limits the generalizability of the results to other retail environments, geographic regions, time spans or longer forecasting horizons.

Classical and machine-learning models were evaluated largely in their standard configurations. Extensive hyperparameter optimization and advanced feature engineering, including feature selection, were not performed for these models, meaning the comparison reflects practical out-of-the-box performance rather than fully optimal model results.

Due to computational constraints, the study evaluated a subset of 48 SKUs from a portfolio exceeding 300 items. While these products account for a substantial share of total sales volume, the sample may not fully capture the range of demand patterns present across the broader portfolio. Similarly, only two time-series foundation models were examined; conclusions regarding the TSFM model class should therefore be interpreted as indicative rather than exhaustive.

SKU identifiers were pseudonymised, preventing linkage between forecast performance and product

identity. This made a SKU-level interpretation of results, such as identifying which specific product categories benefit most from particular forecasting approaches, impossible. Thus, pseudonymisation limits the actionability of findings at the individual product level.

Finally, only the TTM model was finetuned due to computational and budget constraints. Other foundation models may exhibit different adaptation behavior.

6.3 Further Work

The most direct extension of this study is broadening the empirical scope: including additional locations, other bakery chains, longer time horizons, and a larger share of the product portfolio would strengthen the generalizability of the findings. Testing in other retail domains would further clarify whether the observed patterns are specific to organic bakery demand, reflect broader characteristics of retail or even the model classes themselves. A more systematic evaluation of TSFMs, which includes the finetuning and assessment of additional foundation models (especially models requiring GPU access), would allow for more complete conclusions about this model class. Given the budget and computational constraints of this study, this question remains both open and practically highly relevant.

On the feature engineering side, the weather covariates used in this study were limited to linear representations of temperature, precipitation, and wind. Non-linear transformations and threshold-based encodings may better reflect the true relationship between weather and demand and warrant further empirical investigation. Incorporating promotion data would address a key gap identified in the limitations and allow for a more comprehensive analysis of exogenous demand drivers. The role of exceptional high-sales days, beyond the binary spike feature employed here, could similarly be explored in greater depth.

Finally, hierarchical modeling approaches may improve forecast accuracy by leveraging cross-product structure and increase in volume.

Separately, a more granular SKU-level error analysis enabled by access to non-pseudonymised identifiers could further the understanding of where and why forecasting approaches diverge, and inform the segmentation-based hybridization strategy suggested by the findings of this study, and constitute a natural input to the design and evaluation cycle within DSR that this empirical work is intended to support.

A

Complete Listing of Columns

Table A.1: Overview of all columns in the raw point-of-sale dataset

Column Name	Category	Description
rechnung_nummer	Invoice metadata	Unique invoice identifier
rechnung_datum	Invoice metadata	Date of invoice creation
rechnung_zbon_nummer	Invoice metadata	Daily closing report number linked to invoice
buchung_datum	Invoice metadata	Booking / transaction date
lieferschein_nummer	Invoice metadata	Delivery note number
lieferschein_datum	Invoice metadata	Delivery note date
rechnung_zahlungsart	Invoice metadata	Payment method used
rechnung_stornoDatum	Invoice metadata	Timestamp of invoice cancellation
rechnung_stornoGrund	Invoice metadata	Reason for cancellation
storno_bezeichnung	Invoice metadata	Cancellation type label
storno_zeitstempel	Invoice metadata	Exact cancellation timestamp
zbon_nummer	Invoice metadata	Daily closing report number
buchung_zbon_datum	Invoice metadata	Daily report date associated with booking
beleg_typ	Invoice metadata	Document / receipt type
rechnung_interneReferenzRechnung	Invoice metadata	Internal cross-reference to related invoice
rechnung_ersteller	Invoice metadata	User who created the invoice

Continued on next page

Table A.1 – continued from previous page

Column Name	Category	Description
rechnung_empfaenger	Invoice metadata	Invoice recipient designation
rechnung_steuerfrei	Invoice metadata	Flag indicating tax-exempt invoice
rechnung_summe	Invoice metadata	Total invoice gross amount
rechnung_nettoSumme	Invoice metadata	Total invoice net amount
rechnung_steuerSumme	Invoice metadata	Total invoice tax amount
rechnung_anzahlAusdrucke	Invoice metadata	Number of times the invoice was printed
rechnung_bezahlt	Invoice metadata	Flag indicating whether invoice is paid
rechnung_bezahltAm	Invoice metadata	Date and time of payment
buchung_ersteller	Invoice metadata	User who created the booking entry
artikel_nummer	Product information	Internal product identifier
artikel_seriennummer	Product information	Serial number of the article
artikel_ablaufdatum	Product information	Product expiry / best-before date
artikel_bezeichnung	Product information	Product name / description
artikel_kommentar	Product information	Free-text comment on the line item
warengruppe_bezeichnung	Product information	Product group classification
auswertung_bezeichnung	Product information	Reporting / evaluation category name
artikel_menge	Product information	Quantity sold
artikel_einheit	Product information	Unit of measurement
artikel_steuersatz	Product information	Tax rate applied to the product
kontonummer	Product information	Chart-of-accounts number for the article
artikel_preisProEinheit	Pricing	Price per unit (gross)
artikel_nettoPreisProEinheit	Pricing	Price per unit (net)
artikel_summe	Pricing	Total gross line price

Continued on next page

Table A.1 – continued from previous page

Column Name	Category	Description
artikel_nettoSumme	Pricing	Total net line price
artikel_steuerSumme	Pricing	Tax amount on the line item
artikelRabatt_bezeichnung	Pricing	Article-level discount description
artikelRabatt_summe	Pricing	Article-level discount gross amount
artikelRabatt_nettoSumme	Pricing	Article-level discount net amount
rechnungsRabattAliquot_summe	Pricing	Pro-rated invoice discount (gross)
rechnungsRabattAliquot_nettoSumme	Pricing	Pro-rated invoice discount (net)
einkaufspreis_netto	Pricing	Purchase / cost price (net)
einkaufspreis_brutto	Pricing	Purchase / cost price (gross)
marge_netto	Pricing	Net margin on the line item
marge_brutto	Pricing	Gross margin on the line item
kundennummer	Customer information	Customer account number
kunde_firma	Customer information	Customer company name
kunde_uid	Customer information	Customer VAT identification number
kunde_vorname	Customer information	Customer first name
kunde_nachname	Customer information	Customer last name
kunde_strasse	Customer information	Customer street address
kunde_plz	Customer information	Customer postal code
kunde_stadt	Customer information	Customer city
kunde_land	Customer information	Customer country
tisch_kunde	Customer information	Customer assigned to a table
lieferant_name	Supplier information	Supplier company name
lieferant_vorname	Supplier information	Supplier contact first name
lieferant_nachname	Supplier information	Supplier contact last name
lieferant_strasse	Supplier information	Supplier street address

Continued on next page

Table A.1 – continued from previous page

Column Name	Category	Description
lieferant_plz	Supplier information	Supplier postal code
lieferant_stadt	Supplier information	Supplier city
retourbuchung_boolean	Flags / Booleans	Indicates a return / reversal booking
trainingsmodus_boolean	Flags / Booleans	Indicates a training-mode transaction
product_id	System identifiers	System-internal product identifier
productGroup_id	System identifiers	System-internal product group identifier
productCategory_id	System identifiers	System-internal product category identifier
table_id	System identifiers	Table identifier in the POS system
tableArea_id	System identifiers	Table area / zone identifier
employee_id	System identifiers	Employee responsible for the transaction
bill_id	System identifiers	System-internal bill / invoice identifier
deliveryBill_id	System identifiers	System-internal delivery bill identifier
dailyReport_id	System identifiers	System-internal daily report identifier
__source_file	System identifiers	Source file from which the record originates
location	System identifiers	Store location identifier

B

Spike Feature Creation

B.1 Feature List

Table B.1: Input features used for the spike probability classifier

Category	Feature
Identity & Temporal	location_code, artikel_code s
	store_idx, sku_idx
	day_of_week, dow
	month
Calendar & Closures	is_sunday
	is_closed, is_closed_tomorrow, was_closed_yesterdays
	is_holiday, is_day_before_holiday
	school_holiday, start_school_holiday
	fenstertag, christmas_eve, new_years_eve, karfreitag
Meteorological	weathercode e
	temperature_2m_max, temperature_2m_min
	precipitation_hours

B.2 Code Implementation

Listing B.1: Generation of the P$_spikefeatureusinganXGBoostclassifier$

```
1
2  import xgboost as xgb
3  from typing import List
4
5  IDS = ["location", "artikel_bezeichnung"]  # group definition
6  SPIKE_QUANTILE = 0.90
7
8  SPIKE_QUANTILE = 0.90
9  DATE = "date"
10 Y = "artikel_menge"
```

```
11
12 # date column is datetime
13 df_w_naive = df_w_naive.copy()
14 df_w_naive[DATE] = pd.to_datetime(df_w_naive[DATE])
15 df_w_naive = df_w_naive.sort_values(IDS + [DATE])
16
17 # Define train/test split (first 912 days = train)
18 train_end_date = df_w_naive[DATE].min() + pd.DateOffset(days=912)
19 train_mask = df_w_naive[DATE] < train_end_date
20
21 print(f"Total days: {len(df_w_naive)}")
22
23 CLASSIFIER_FEATURES = [
24     'location_code', 'artikel_code', 'store_idx', 'sku_idx',
        'day_of_week',
25     'is_sunday', 'is_holiday', 'is_day_before_holiday', 'school_holiday',
26     'start_school_holiday', 'christmas_eve', 'new_years_eve',
        'karfreitag',
27     'fenstertag', 'is_closed', 'is_closed_tomorrow',
        'was_closed_yesterday',
28     'weathercode (wmo code)', 'temperature_2m_max ( C )',
29     'temperature_2m_min ( C )', 'precipitation_hours (h)', 'dow', 'month'
30 ]
31
32 # minimum number of spike days required to train classifier
33 MIN_SPIKE_DAYS = 10
34
35 df_w_naive['P_spike'] = np.nan
36 df_w_naive['is_spike_day_all'] = np.nan
37
38 for group_key, group_idx in df_w_naive.groupby(IDS).groups.items():
39     g = df_w_naive.loc[group_idx].sort_values(DATE)
40     g_train = g[g[DATE] < train_end_date].copy()
41
42     if len(g_train) == 0:
43         continue
44
45     # training-only spike threshold
46     thr = g_train[Y].quantile(SPIKE_QUANTILE)
47
48     df_w_naive.loc[group_idx, 'is_spike_day_all'] = (g[Y] >=
        thr).astype(int).values
49
50     # training labels
51     y_train = (g_train[Y] >= thr).astype(int)
52     n_pos = int(y_train.sum())
53     n_neg = int((1 - y_train).sum())
54
55     # fallback: constant spike probability
56     if n_pos < MIN_SPIKE_DAYS or n_neg == 0:
```

```
        df_w_naive.loc[group_idx, 'P_spike'] = float(y_train.mean())
        continue

    X_train = g_train[CLASSIFIER_FEATURES]

    xgb_clf = xgb.XGBClassifier(
        n_estimators=100,
        max_depth=3,
        learning_rate=0.1,
        objective='binary:logistic'
    )

    xgb_clf.fit(X_train, y_train)

    X_all = g[CLASSIFIER_FEATURES]
    df_w_naive.loc[group_idx, 'P_spike'] =
        xgb_clf.predict_proba(X_all)[:, 1]


print("\n--- New Feature Generation Complete (per-group models) ---")
print(f"Spike days (Target=1): {df_w_naive['is_spike_day_all'].sum()}")
print(df_w_naive['P_spike'].describe().to_markdown())

print("\n--- P_spike Comparison (0=Non-Spike, 1=Spike) ---")

print("\nSpike Day P_spike Distribution:")
print(df_w_naive[df_w_naive['is_spike_day_all'] ==
    1]['P_spike'].describe().to_markdown())

print("\nNon-Spike Day P_spike Distribution:")
print(df_w_naive[df_w_naive['is_spike_day_all'] ==
    0]['P_spike'].describe().to_markdown())
```

C

Complete List of Features for the Final Dataset

Table C.1: Final feature set used in the analysis

No.	Variable
1	date
2	location
3	artikel_bezeichnung
4	artikel_menge
5	location_code
6	artikel_code
7	store_idx
8	sku_idx
9	day_of_week
10	is_sunday
11	is_holiday
12	is_day_before_holiday
13	school_holiday
14	start_school_holiday
15	christmas_eve
16	new_years_eve
17	karfreitag
18	fenstertag

Continued on next page

Table C.1 – continued from previous page

No.	Variable
19	days_until_christmas
20	is_december_week_4
21	is_fenstertag_friday
22	is_payday_week
23	holiday_proximity_7d
24	is_closed
25	is_closed_tomorrow
26	was_closed_yesterday
27	sales_yesterday
28	sales_yesterweek
29	sales_avg4weeks
30	weathercode (wmo code)
31	temperature_2m_max (°C)
32	temperature_2m_min (°C)
33	precipitation_hours (h)
34	na_lag1
35	na_lag2
36	na_lag3
37	na_lag7
38	na_lag14
39	na_lag21
40	na_lag28
41	na_avg4w_mean
42	na_avg4w_median
43	dow
44	month

Continued on next page

Table C.1 – continued from previous page

No.	Variable
45	P_spike

D

Complete Listing of Model Configurations and Hyperparameters

D.1 XGBoost

Table D.1: XGBoost hyperparameter configuration

Hyperparameter	Value
objective	reg:squarederror
learning_rate	0.05
n_estimators	600
max_depth	7
min_child_weight	2
subsample	0.9
colsample_bytree	0.9
reg_lambda	0.1
random_state	42
tree_method	hist

D.2 TinyTimeMixer

Table D.2: TTM model configuration

Parameter	Value

Parameter	Value
model_path	ibm-granite/granite-timeseries-ttm-r1
architectures	TinyTimeMixerForPrediction

Continued on next page

Table D.2 – continued from previous page

Parameter	Value
model_type	tinytimemixer
context_length	512
prediction_length	96
d_model	192
num_layers	2
num_patches	8
patch_length	64
patch_stride	64
adaptive_patching_levels	3
decoder_d_model	128
decoder_mode	common_channel
decoder_num_layers	2
decoder_adaptive_patching_levels	0
decoder_raw_residual	False
mode	common_channel
num_input_channels	1
num_parallel_samples	100
expansion_factor	2
dropout	0.2
head_dropout	0.2
distribution_output	student_t
loss	mse
scaling	std
norm_mlp	LayerNorm
norm_eps	1e-5
init_std	0.02

Continued on next page

Table D.2 – continued from previous page

Parameter	Value
positional_encoding_type	sincos
frequency_token_vocab_size	5
gated_attn	True
self_attn	False
self_attn_heads	1
patch_last	True
use_decoder	True
use_positional_encoding	False
init_processing	True
post_init	False
resolution_prefix_tuning	False
torch_dtype	float32
transformers_version	4.37.2

D.3 Finetuning Hyperparameters

The hyperparameter optimization was conducted using a predefined search space, summarized as follows:

- **Dropout rate**

$$\text{dropout} \sim \mathcal{U}(0.05,\ 0.30)$$

  where lower values allow reduced regularization.

- **Learning rate**

$$\text{learning\_rate} \sim \text{LogUniform}(10^{-4},\ 10^{-3})$$

  sampled on a logarithmic scale to cover one order of magnitude.

- **Number of training epochs**

$$\text{num\_train\_epochs} \sim \mathcal{U}_Z(15,\ 30)$$

  reflecting initial convergence range.

- **Batch size**

$$\text{per\_device\_train\_batch\_size} = 64$$

Table D.3: TTM finetuning training arguments

Parameter	Value
per_device_train_batch_size	32
learning_rate	5e-4
warmup_steps	0
gradient_accumulation_steps	1
max_grad_norm	1.0
weight_decay	0.01
save_strategy	no
logging_steps	1000

Bibliography

Arango, S. P., Mercado, P., Kapoor, S., Ansari, A. F., Stella, L., Shen, H., Senetaire, H., Turkmen, C., Shchur, O., Maddix, D. C., Bohlke-Schneider, M., Wang, Y., & Rangapuram, S. S. (2025, March). ChronosX: Adapting Pretrained Time Series Models with Exogenous Variables. https://doi.org/10.48550/arXiv.2503.12107

Arunraj, N. S., & Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal of Production Economics*, *170*, 321–335. https://doi.org/10.1016/j.ijpe.2015.09.039

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., ... Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. https://doi.org/10.48550/ARXIV.1806.01261

Beck, N., Dovern, J., & Vogl, S. (2025). Mind the naive forecast! a rigorous evaluation of forecasting models for time series with low predictability. *Applied Intelligence*, *55*(6), 395. https://doi.org/10.1007/s10489-025-06268-w

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2016). *Time series analysis: Forecasting and control* (Fifth edition). Wiley.

Brotsüchtig – digitale Lebensmittelrettung. (2024, July).

Chan, H., & Wahab, M. (2024). A machine learning framework for predicting weather impact on retail sales. *Supply Chain Analytics*, *5*, 100058. https://doi.org/10.1016/j.sca.2024.100058

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. https://doi.org/10.48550/ARXIV.1603.02754

Chowdhury, A. R., & Rozony, F. Z. (2025). A systematic Review of Demand Forecasting Models for Retail E-Commerce enhancing Accuracy in Inventory and Delivery Planning. *International Journal of Scientific Interdisciplinary Research*, *06*(01), 01–27. https://doi.org/10.63125/mbbfw637

Church, K. W., Chen, Z., & Ma, Y. (2021). Emerging trends: A gentle introduction to fine-tuning. *Natural Language Engineering*, *27*(6), 763–778. https://doi.org/10.1017/S1351324921000322

Das, A., Kong, W., Sen, R., & Zhou, Y. (2024, April). A decoder-only foundation model for time-series forecasting. https://doi.org/10.48550/arXiv.2310.10688

Dayama, P., Ekambaram, V., Gifford, W., Jati, A., Kalagnanam, J., Mukherjee, S., Nguyen, N., & Reddy, C. (2024). Tiny Time Mixers (TTMs): Fast Pre-trained Models for Enhanced Zero/Few-Shot Forecasting of Multivariate Time Series. *Advances in Neural Information Processing Systems 37*, 74147–74181. https://doi.org/10.52202/079017-2359

Digitale Lebensmittelrettung: brotsüchtig, WKOÖ und SCCH reduzieren Überproduktion von Brot und Gebäck um 20 Prozent - scch.at. (2024, July).

Elsayed, S., Thyssens, D., Rashed, A., Jomaa, H. S., & Schmidt-Thieme, L. (2021). Do We Really Need Deep Learning Models for Time Series Forecasting? https://doi.org/10.48550/ARXIV.2101.02118

Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting*, *38*(4), 1283–1318. https://doi.org/10.1016/j.ijforecast.2019.06.004

Fries, M., & Ludwig, T. (2024). 'Why are the Sales Forecasts so low?' Socio-Technical Challenges of Using Machine Learning for Forecasting Sales in a Bakery. *Computer Supported Cooperative Work (CSCW)*, *33*(2), 253–293. https://doi.org/10.1007/s10606-022-09458-z

Garza, A., Challu, C., & Mergenthaler-Canseco, M. (2023). TimeGPT-1. https://doi.org/10.48550/ARXIV.2310.03589

Gu, Y., Jazizadeh, F., & Wang, X. (2025). Toward Large Energy Models: A comparative study of Transformers' efficacy for energy forecasting. *Applied Energy*, *384*, 125358. https://doi.org/10.1016/j.apenergy.2025.125358

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hevner, A. R., & Chatterjee, S. (2010). *Design research in information systems: Theory and practice*. Springer.

Historical Forecast API | Open-Meteo.com. (2026).

Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting*, *36*(4), 1420–1438. https://doi.org/10.1016/j.ijforecast.2020.02.005

Hübner, N., Caspers, J., Coroamă, V. C., & Finkbeiner, M. (2024). Machine-learning-based demand forecasting against food waste: Life cycle environmental impacts and benefits of a bakery case study. *Journal of Industrial Ecology*, *28*(5), 1117–1131. https://doi.org/10.1111/jiec.13528

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (Third edition). OTexts.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Kaggle. (2021, October). *State of Data Science and Machine Learning 2021* (tech. rep.).

Kolassa, S., & Schütz, W. (2007). Advantages of the MAD/Mean Ratio over the MAPE. *Foresight: The International Journal of Applied Forecasting*, (6), 40–43.

Lebersorger, S., & Schneider, F. (2014). Food loss rates at the food retail, influencing factors and reasons as a basis for waste prevention measures. *Waste Management*, *34*(11), 1911–1919. https://doi.org/10.1016/j.wasman.2014.06.013

Li, Z., Qiu, X., Chen, P., Wang, Y., Cheng, H., Shu, Y., Hu, J., Guo, C., Zhou, A., Jensen, C. S., & Yang, B. (2025). TSFM-Bench: A Comprehensive and Unified Benchmark of Foundation Models for Time Series Forecasting. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 5595–5606. https://doi.org/10.1145/3711896.3737442

Liang, Y., Wen, H., Nie, Y., Jiang, Y., Jin, M., Song, D., Pan, S., & Wen, Q. (2024). Foundation Models for Time Series Analysis: A Tutorial and Survey. https://doi.org/10.48550/ARXIV.2403.14735

Lima, R. D. T. D., Fernandes, S. M. M., & Lima, S. M. L. D. (2025). Time series forecasting with exogenous variables: A literature review to identify promising gaps in computational research. *Revista Principia*, *62*. https://doi.org/10.18265/2447-9187a2025id8575

Liu, X., Liu, J., Woo, G., Aksu, T., Liang, Y., Zimmermann, R., Liu, C., Savarese, S., Xiong, C., & Sahoo, D. (2024). Moirai-MoE: Empowering Time Series Foundation Models with Sparse Mixture of Experts. https://doi.org/10.48550/ARXIV.2410.10469

Löning, M., Bagnall, A., Ganesh, S., Kazakov, V., Lines, J., & Király, F. J. (2019). Sktime: A Unified Interface for Machine Learning with Time Series. https://doi.org/10.48550/ARXIV.1909.07872

Lovering, C., Jha, R., Linzen, T., & Pavlick, E. (2021). Predicting Inductive Biases of Pre-Trained Models. *International Conference on Learning Representations*.

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. https://doi.org/10.1016/j.ijforecast.2019.04.014

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, *38*(4), 1346–1364. https://doi.org/10.1016/j.ijforecast.2021.11.013

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Python in Science Conference*, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Miller, J. A., Aldosari, M., Saeed, F., Barna, N. H., Rana, S., Arpinar, I. B., & Liu, N. (2024). A Survey of Deep Learning and Foundation Models for Time Series Forecasting. https://doi.org/10.48550/ARXIV.2401.13912

Pietrangeli, R., Eriksson, M., Strotmann, C., Cicatiello, C., Nasso, M., Fanelli, L., Melaragni, L., & Blasi, E. (2023). Quantification and economic assessment of surplus bread in Italian small-scale bakeries: An explorative study. *Waste Management*, *169*, 301–309. https://doi.org/10.1016/j.wasman.2023.07.017

Qiao, Z., Liu, C., Zhang, Y., Jin, M., Pham, Q., Wen, Q., Suganthan, P. N., Jiang, X., & Ramasamy, S. (2025). Multi-Scale Finetuning for Encoder-based Time Series Foundation Models. https://doi.org/10.48550/ARXIV.2506.14087

Ribeiro, B. M. C. (2025). *Improving Multi-SKU Demand Forecasting Through Classification-Driven Modeling: A Hierarchical Approach for Industrial Applications* [Master's thesis, Universidade do Porto].

Rose, N., & Dolega, L. (2022). It's the Weather: Quantifying the Impact of Weather on Retail Sales. *Applied Spatial Analysis and Policy*, *15*(1), 189–214. https://doi.org/10.1007/s12061-021-09397-0

Schneider, J., Meske, C., & Kuss, P. (2024). Foundation Models: A New Paradigm for Artificial Intelligence. *Business & Information Systems Engineering*, *66*(2), 221–231. https://doi.org/10.1007/s12599-024-00851-0

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Python in Science Conference*, 92–96. https://doi.org/10.25080/Majora-92bf1922-011

Taylor, S. J., & Letham, B. (2018). Forecasting at Scale. *The American Statistician*, *72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

Wieringa, R. J. (2014). *Design Science Methodology for Information Systems and Software Engineering*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-43839-8

Wikamulia, N., Jonathan, M., & Isa, S. M. (2023). Bakery Demand Forecasting Using XGBoost and K-Means Clustering. https://doi.org/10.24507/icicelb.14.01.21

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2024). *Experimentation in Software Engineering*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-69306-3

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., & Sahoo, D. (2024). Unified Training of Universal Time Series Forecasting Transformers. https://doi.org/10.48550/ARXIV.2402.02592

Yang, C.-L., & Sutrisno, H. (2018). Short-Term Sales Forecast of Perishable Goods for Franchise Business. *2018 10th International Conference on Knowledge and Smart Technology (KST)*, 101–105. https://doi.org/10.1109/KST.2018.8426091

Yang, W., Cao, D., & Liu, Y. (2025). Foundation Models for Demand Forecasting via Dual-Strategy Ensembling. https://doi.org/10.48550/ARXIV.2507.22053

Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023). Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(9), 11121–11128. https://doi.org/10.1609/aaai.v37i9.26317

Zhang, L., Bian, W., Qu, W., Tuo, L., & Wang, Y. (2021). Time series forecast of sales volume based on XGBoost. *Journal of Physics: Conference Series*, *1873*(1), 012067. https://doi.org/10.1088/1742-6596/1873/1/012067