Master's Programme in Life Science Technologies: Complex Systems

Forecasting Prescription Medication Utilization: A Comparative Study of SARIMA, Prophet, XGBoost and LSTM Models

Leveraging Machine Learning and Statistical Models for Predictive Analysis of Pharmaceutical Utilization in Sweden: A Comparative Evaluation Using RMSE, MAE, and MAPE Metrics

---

Vili Ripatti

Aalto University

MASTER'S THESIS

Aalto University

MASTER'S THESIS 2025

Forecasting Prescription Medication Utilization: A Comparative Study of SARIMA, Prophet, XGBoost and LSTM Models

Leveraging Machine Learning and Statistical Models for Predictive Analysis of Pharmaceutical Utilization in Sweden: A Comparative Evaluation Using RMSE, MAE, and MAPE Metrics

Vili Ripatti

Thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Technology.

Helsinki, 10 February 2024

Supervisor: Dr. Riku Linna
Advisors: Ph.D Ornela Bardhi

Aalto University
School of Science
Master's Programme in
Life Science Technologies

Aalto University

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Abstract

Author
Vili Ripatti

Title
Forecasting Prescription Medication Utilization: A Comparative Study of SARIMA, Prophet, XGBoost and LSTM Models

School School of Science

Master's programme Life Science Technologies

Major Complex Systems **Code** SCI3060

Supervisor Dr. Riku Linna

Advisors Ph.D Ornela Bardhi

Level Master's thesis **Date** 10 February 2025 **Pages** 64 **Language** English

Abstract

Predictive analytics in healthcare is essential for improving patient outcomes, managing limited resources, and addressing the growing complexities of drug demand and supply. Accurate forecasting of prescription drug utilization enables healthcare providers and policymakers to make informed decisions, reduce costs, and ensure timely availability of medications. This study evaluates the performance of four widely used forecasting models, SARIMA, Prophet, XGBoost, and LSTM, in 97 cardio-related ATC codes using a dataset from the Swedish National Board of Health and Welfare (2009–2023). The effectiveness of the models was compared by using root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE) as evaluation metrics.

The results indicate that SARIMA consistently outperforms other models, accurately capturing seasonal patterns, abrupt shifts, and long-term trends, securing first place in three quarters of cases. LSTM demonstrated strong performance in handling datasets with nonlinear dependencies but occasionally struggled with abrupt trend shifts. XGBoost delivered reasonable results, particularly for simpler datasets, yet struggled with high volatility, leading to reduced accuracy in unstable time series. Prophet, while robust to missing data, was less effective in capturing complex temporal dynamics, leading to higher errors in datasets with irregular trends.

These findings underscore the importance of aligning model selection with dataset characteristics to optimize forecasting outcomes. By highlighting the strengths and limitations of these models, this research contributes to the growing body of knowledge on predictive modeling in healthcare, offering valuable insights for practitioners aiming to enhance decision-making processes in drug utilization management.

Keywords Prescription Medication, Forecasting, Time Series Analysis, Machine Learning in Healthcare, SARIMA, Prophet, LSTM Networks, XGBoost, Hyperparameter Tuning, Exogenous Variables

Aalto University

Aalto-Yliopisto, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Tiivistelmä

Tekijä
Vili Ripatti

Työn nimi
Reseptilääkkeiden käytön ennustaminen: vertaileva tutkimus SARIMA-, Prophet-, XGBoost- ja LSTM-malleista

Koulu Perustieteiden korkeakoulu

Koulutusohjelma Life Science Technologies

Pääaine Complex Systems **Pääainekoodi** SCI3060

Työn valvoja TkT Riku Linna

Työn ohjaaja Ph.D Ornela Bardhi

Taso Diplomityö **Päivämäärä** 10.2.2025 **Sivumäärä** 64 **Kieli** suomi

Tiivistelmä

Terveydenhuollossa käytettävät ennustemallit ovat keskeisen tärkeitä potilastulosten parantamiseksi, rajallisten resurssien hallitsemiseksi sekä lääkkeiden kysynnän ja tarjonnan kasvavien haasteiden ratkaisemiseksi. Reseptilääkkeiden määräämisen täsmällinen ennustaminen mahdollistaa terveydenhuollon toimijoiden sekä päättäjien tietoon perustuvien päätöksien tekemisen, kustannusten vähentämisen sekä lääkkeiden saatavuuden varmistamisen. Tässä tutkimuksessa arvioidaan neljän laajalti käytetyn ennustemallin – SARIMA:n, Prophet:in, XGBoost:in ja LSTM:n – suorituskykyä 97:ssä sydän- ja verisuonitauteihin liittyvässä ATC-koodissa hyödyntäen Ruotsin terveys- ja hyvinvointiviraston aineistoa vuosilta 2009–2023. Mallien onnistumista mitataan jäännösvirhehajonnalla (RMSE), keski-itseisvirheellä (MAE) sekä keskimääräisellä absoluuttisella virheellä (MAPE).

Tulokset osoittavat, että SARIMA suoriutuu yksiselitteisesti muita malleja paremmin. SARIMA onnistuu mallintamaan tarkasti kausittaisia vaihteluita, äkillisiä muutoksia sekä pitkän aikavälin trendejä, saavuttaen parhaan sijan kolmessa neljäsosassa tapauksista. LSTM osoitti vahvaa suorituskykyä epälineaaristen riippuvuussuhteiden käsittelyssä, mutta kohtasi ajoittain haasteita äkillisten trendimuutosten kanssa. XGBoost tuotti kohtalaisia tuloksia erityisesti yksinkertaisemmissa aineistoissa, mutta kärsi korkean volatiliteetin aiheuttamista haasteista, mikä johti heikentyneeseen tarkkuuteen epävakaissa aikasarjoissa. Prophet, vaikkakin mukautuu tehokkaasti puuttuvaan dataan, oli heikko tulkitsemaan monimutkaisia ajallisen dynamiikan aikasarjoja. Tämä johti suuriin virheisiin aineistoissa, joissa oli epäsäännöllisiä trendejä.

Nämä löydökset korostavat mallivalinnan ja aineiston ominaispiirteiden yhteensovittamisen merkitystä ennustetarkkuuden optimoimiseksi. Tuomalla esiin mallien vahvuudet ja rajoitteet tämä tutkimus kartuttaa terveydenhuollon ennustavan mallintamisen tietoperustaa, tarjoten arvokkaita näkemyksiä ammattilaisille, jotka pyrkivät tehostamaan lääkkeiden käytön hallintaan liittyvää päätöksentekoa.

Avainsanat Reseptilääkkeet, Ennustaminen, Aikasarja-analyysi, Koneoppiminen terveydenhuollossa, SARIMA, Prophet, LSTM-verkot, XGBoost, Hyperparametrien viritys, Eksogeeniset muuttujat

Contents

Preface

Hassua, nyt se loppuu. Kiitos Success Clinic mahdollisuudesta tehdä töitä ja tämä työ. Kiitos ohjaajalleni Ornelalle, joka auttoi auttamatonta ja tuki jokaisella askeleella. Kiitos valvojalleni Rikulle kärsivällisyydestä sekä vastaamisesta iltamyöhään kirjoitettuihin avunpyyntöihin.

Minulla oli ihan todella hauskaa opiskellessa. Toivottavasti minulla on hauskaa myös tämän ajan jälkeen.

Lopuksi muutama lause inspiroivilta ja unohtumattomilta diplomi-insinööreiltä, kauppatieteiden maisterilta sekä arkkitehdiltä:

Kiitos Ilmari, Emil, Sekseli, Silli, Stas ja Vade unohtumattomista perjantaista (Siljander, 2023).

Kiitos ystävilleni tukemisesta (Rasku, 2023).

Lisäksi, haluan kiittää ystäviäni (Perttula, 2023) Kallea, Anttia ja Anttonia niin vietetyistä hetkistä, kuin seikkailuista (Niskanen, 2024).

Erityisesti kiitos koko FTMK yhdeksännelletoista unohtumattomista muistoista (Vänskä, 2023).

Kiitos AYY:lle antoisasta opiskelijaelämän vuodesta. Ja kiitos, Vili, rakkaudestasi ja tuestasi, erityisesti kuluneen vuoden aikana (Leinonen, 2022).

Kiitos Äpy ja Rektorsgatan 1 väki ajoista joita ei saa eikä voi unohtaa sekä lukemattomista ja luetuista hilpeistä hetkistä (Luukkanen, 2023; Riski, 2022).

Kiitos Miki ja Otso, lapsuudesta, nuoruudesta ja elämästä.

Kiitos Allu, Pepe, Äiti ja Faija kasvatuksesta.

Ja Milja, kiitos, että oot ollut tukena aina, aivan jokaisessa asiassa.

Kotona, 10. päivänä helmikuuta 2025

Vili Ripatti

Acronyms and Abbreviations

AI	Artificial Intelligence
EHR	Electronic Health Record
ARIMA	Autoregressive Integrated Moving Average
LSTM	Long Short-Term Memory
SARIMA	Seasonal Autoregressive Integrated Moving Average
XGBoost	eXtreme Gradient Boosting
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ATC	Anatomical Therapeutic Chemical Classification System
RNN	Recurrent Neural Network
AIC	Akaike Information Criterion
ADF	Augmented Dickey-Fuller Test
KPSS	Kwiatkowski-Phillips-Schmidt-Shin Test
MAD	Median Absolute Deviation
SMAPE	Symmetric Mean Absolute Percentage Error
MSLE	Mean Squared Logarithmic Error
DTW	Dynamic Time Warping

p	Order of the autoregressive (AR) component in SARIMA
d	Degree of differencing in SARIMA
q	Order of the moving average (MA) component in SARIMA
P, D, Q	Seasonal counterparts of AR, I, and MA in SARIMA
s	Seasonal period (e.g., 12 for monthly data)
$\phi(B)$	Autoregressive polynomial in SARIMA
$\theta(B)$	Moving average polynomial in SARIMA
$\Phi(B^s)$	Seasonal autoregressive polynomial in SARIMA
$\Theta(B^s)$	Seasonal moving average polynomial in SARIMA
ϵ_t	White noise error term in SARIMA
$y(t)$	Observed time series value at time t in Prophet
$g(t)$	Trend component in Prophet
$s(t)$	Seasonal component in Prophet
$h(t)$	Holiday or external event effects in Prophet
K	Number of boosting iterations in XGBoost
$\mathcal{L}(y_i, \hat{y}_i)$	Loss function in XGBoost
$\Omega(f_k)$	Regularization term in XGBoost
f_t	Forget gate activation in LSTM
i_t	Input gate activation in LSTM
o_t	Output gate activation in LSTM
C_t	Cell state in LSTM
h_t	Hidden state in LSTM
AIC	Akaike Information Criterion
Δy_t	First difference of the time series in ADF test
ϵ_t	Error term in stationarity tests
τ_t	Deterministic trend component in KPSS test

1. Introduction

1.1 Background

Artificial Intelligence (AI) has revolutionized predictive analytics across various domains, leveraging its capacity to process vast datasets and uncover complex patterns that traditional methods often miss. The ability of AI to analyze unstructured data, such as text, images, and videos, has enabled breakthroughs in fields ranging from finance to environmental monitoring. For instance, machine learning algorithms can predict stock market trends, optimize supply chains, and anticipate weather patterns with remarkable precision, offering valuable insights for decision-making in highly dynamic environments [1, 2].

In healthcare, AI-powered predictive analytics extends its transformative potential to areas like disease diagnosis, personalized treatment planning, and resource management. Algorithms trained on electronic health records (EHRs) and diagnostic imaging have demonstrated superhuman performance in identifying early signs of diseases such as cancer and cardiovascular conditions, enabling timely interventions [3]. Moreover, the integration of AI into predictive systems facilitates the proactive management of critical resources, such as hospital staffing and medication inventories, ensuring operational efficiency while enhancing patient care outcomes. As data continues to grow in volume and complexity, AI's role as a critical enabler of predictive analytics will only expand, driving innovation and improving outcomes across diverse sectors [2].

Predictive analytics in healthcare is revolutionizing patient care and operational efficiency by enabling data-driven decision-making. The increasing availability of electronic health records (EHRs), diagnostic imaging, and

sensor data provides a foundation for advanced analytics to predict patient trajectories, resource needs, and critical events [1, 2]. However, despite the transformative potential of EHRs, predictive models leveraging this wealth of information remain underdeveloped in several critical areas, including prescription drug utilization and proactive patient management. Before the COVID-19 pandemic, limited attention was given to applying machine learning and time series methodologies to EHR data, resulting in a lack of understanding of how these models can effectively support prescriptive and preventative healthcare.

Time series forecasting methods like ARIMA and LSTM play a key role in healthcare analytics. ARIMA models are well-suited for identifying linear trends and seasonal patterns, as demonstrated by Dritsakis and Klazoglou in their study on forecasting U.S. healthcare expenditures [4]. In contrast, LSTM networks excel in modeling non-linear dynamics and long-term dependencies, making them particularly effective for predicting clinical events and patient conditions in real time [5, 6]. However, there is a notable gap in systematic studies that integrate these advanced models with robust hyperparameter tuning specifically tailored to healthcare data. This limitation often restricts the adaptability and precision of these models, especially when incorporating exogenous variables such as socio-economic factors, seasonal variations, and policy changes.

Integrating traditional statistical models with machine learning approaches offers a robust framework for healthcare forecasting. Hybrid models combining ARIMA's strengths in short-term forecasting with the adaptability of LSTM and XGBoost have shown promise in capturing complex temporal patterns and improving accuracy [7, 3]. These models are critical for applications such as resource allocation, staffing optimization, and medication utilization [8, 6]. However, the high variability and external influences impacting healthcare data, including sudden demand surges or demographic shifts, present challenges that require further research into model optimization and application.

Forecasting models are increasingly used to predict patient demand and improve operational planning. Accurate predictions enable hospitals to manage resources efficiently, particularly during peak demand periods like flu seasons. These tools also support early interventions, helping prevent adverse events and optimize patient outcomes [1, 5]. Yet, prescription drug forecasting—a critical component of healthcare delivery—remains

underexplored. Predictive models for prescription demand must comply with dynamic factors such as regional health policies, demographic trends, and external shocks. Enhancing machine learning approaches in this domain could significantly improve resource management and patient care.

Forecasting plays a pivotal role in healthcare management, enabling policymakers and administrators to make informed, strategic decisions by reducing uncertainty [9]. Accurate predictions of key healthcare metrics, such as fertility, morbidity, and mortality rates, provide invaluable insights for proactive planning and interventions, serving as indicators of societal health [9, 8].

Healthcare data often exhibit complex patterns, requiring sophisticated models to improve predictive accuracy. Short-term forecasting supports rapid responses to fluctuations in patient admissions or emergency care demands, while long-term forecasting aids in planning for chronic disease management and resource allocation years in advance [8]. These forecasts enable healthcare managers to optimize medication inventory, reduce waste, and ensure the availability of critical resources.

The selection of forecasting models depends on data characteristics, desired accuracy, and decision-making contexts [8]. Integrating predictive models with decision-support systems, such as those incorporating fuzzy logic, further enhances their applicability by combining insights with expert knowledge to guide resource prioritization and policy development [9]. By bridging predictive analytics with operational decision-making, forecasting models empower healthcare organizations to shift from reactive to proactive management, improving both patient outcomes and system sustainability.

Despite these challenges, advancements in predictive analytics highlight the promise of combining domain-specific knowledge with cutting-edge techniques. As Obermeyer and Lee emphasize, predictive models should complement clinical decision-making rather than replace it. This requires close collaboration between healthcare professionals with domain expertise and data scientists developing the algorithms [2]. Addressing existing gaps through targeted research on EHR-based predictions, hyperparameter tuning, and advanced forecasting methodologies could establish a comprehensive framework for healthcare analytics, supporting both resource optimization and proactive healthcare delivery.

1.2 Hypothesis

The purpose of this study is to determine the most suitable forecasting model for predicting prescription drug utilization. This research seeks to explore how different time series forecasting models—SARIMA, Prophet, XGBoost, and LSTM—perform when applied to healthcare data with seasonal patterns, abrupt changes, and non-linear dependencies. A key question guiding this study is whether a prescription dataset, characterized by temporal trends and seasonal variations, is appropriate for evaluating these forecasting models. Additionally, this study investigates how standard error metrics (RMSE, MAE, MAPE) influence the evaluation of model performance and what methodological considerations, such as data preprocessing, hyperparameter tuning, and train-test splitting, play a role in the accuracy of predictions. By addressing these questions, this research aims to provide a comparative analysis that will contribute to the development of more reliable forecasting techniques for healthcare applications.

The analysis uses a comprehensive dataset provided by Success Clinic from the Swedish National Board of Health and Welfare, encompassing individual-level prescription data categorized by 97 ATC codes. The data spans from January 2009 to December 2023 and includes prescriptions for a variety of conditions, such as heart failure and type-2 diabetes. This study will evaluate the models' predictive accuracy across diverse drug categories, providing insights into the optimal approach for forecasting medication demand and informing future healthcare planning.

1.3 Limitations

While this study aims to provide a comprehensive comparison of SARIMA, Prophet, XGBoost, and LSTM models for predicting prescription medication, several limitations must be acknowledged.

First, the dataset used in this research comprises high-quality individual-level prescription data for 97 Anatomical Therapeutic Chemical (ATC) codes in Sweden, spanning from January 2009 to December 2023. Although the data required minimal preprocessing due to its quality, the scope is limited to Sweden's healthcare system. As prescribing behaviors, healthcare policies, and population demographics vary globally, the generalizability of these findings to other regions may be limited.

Second, the models employed in this study rely on specific assumptions and hyperparameters. Despite efforts to optimize each model for the dataset, alternative configurations or methodologies could yield different results. Additionally, while these models excel at uncovering temporal patterns, they do not incorporate exogenous variables such as socio-economic changes, seasonal health trends, or policy shifts, which could significantly influence prescription patterns.

Furthermore, although the dataset's temporal range provides a robust basis for analysis, it may not fully capture the impact of unforeseen events, such as pandemics or sudden changes in medical practices, that could affect prescription behavior in the future. This limitation highlights the potential need for adaptive models capable of accounting for such factors.

Finally, computational constraints present another challenge. Advanced models like LSTM and XGBoost require substantial resources for training and hyperparameter tuning, potentially limiting the exploration of additional data features or configurations. The evaluation of model performance using metrics such as RMSE and MAPE provides a clear comparison framework; however, these metrics may not fully capture the operational implications of prediction errors, such as overstocking or understocking medications.

Despite these limitations, the study's high-quality data and rigorous methodology contribute to meaningful insights into prescription medication forecasting. Future research could address these constraints by incorporating external variables, expanding geographic scope, and exploring additional model configurations to enhance the robustness and applicability of the findings.

1.4 Thesis structure

This thesis is structured as follows:

- Chapter 1 provides an introduction to the research topic, outlining the background, hypothesis, practical implications, limitations, and the structure of the thesis.

- Chapter 2 focuses on the theoretical framework, exploring traditional and advanced forecasting models and their relevance to healthcare appli-

cations.

- Chapter 3 details the data and methodology, including data preprocessing, model configurations, and evaluation metrics.

- Chapter 4 presents the results of the comparative analysis, showcasing the performance of each model across different evaluation criteria.

- Chapter 5 discusses the implications of the findings, identifies limitations, and provides recommendations for future research.

- Chapter 6 concludes the thesis by summarizing the key contributions and discussing potential applications of the findings in healthcare forecasting.

1.5 Summary

The above introduction outlined the importance of predictive analytics in healthcare, focusing on forecasting prescription drug utilization using SARIMA, Prophet, XGBoost, and LSTM models. Leveraging a high-quality dataset of 97 ATC codes from Sweden, this thesis compares these models to identify the most effective approach for accurate predictions.

The study highlights the transformative role of AI in healthcare, particularly in analyzing electronic health records (EHRs) for improved decision-making. However, gaps remain in applying predictive models to dynamic areas like prescription demand, with challenges in addressing external variables and optimizing configurations.

Despite limitations such as the regional scope of the data and the exclusion of exogenous factors, this research contributes valuable insights into forecasting methodologies. The thesis is structured to cover theoretical foundations, methodologies, results, and implications, culminating in a comprehensive evaluation of forecasting models for healthcare applications.

2. Theory

Time series forecasting plays a crucial role in healthcare analytics, particularly in predicting prescription medication utilization. This chapter provides a theoretical foundation for the key forecasting models used in this study: Seasonal Autoregressive Integrated Moving Average (SARIMA), Prophet, XGBoost, and Long Short-Term Memory (LSTM) networks. Each method is rooted in different theoretical principles, offering unique advantages and challenges in predictive modeling.

2.1 Seasonal Autoregressive Integrated Moving Average

Seasonal Autoregressive Integrated Moving Average (SARIMA) is a widely used statistical model for time series forecasting that extends the standard Autoregressive Integrated Moving Average (ARIMA) framework to accommodate seasonality. Given that prescription medication data often exhibit both trend and seasonal components, SARIMA provides an effective way to model these patterns explicitly [10, 11].

SARIMA incorporates autoregressive (AR) terms, differencing (I), and moving average (MA) terms, along with seasonal counterparts. The model is denoted as SARIMA$(p, d, q)(P, D, Q)_s$, where p represents the number of lagged observations, d denotes the degree of differencing to achieve stationarity, and q specifies the number of lagged forecast errors. The seasonal terms P, D, Q correspond to seasonal autoregressive, differencing, and moving average components, respectively, while s represents the seasonal period.

The general SARIMA model is formulated as:

$$\phi(B)\Phi(B^s)\Delta^d\Delta_s^D Z_t = \theta(B)\Theta(B^s)\epsilon_t, \tag{2.1}$$

where $\phi(B)$ and $\theta(B)$ are polynomials in the backshift operator B, $\Phi(B^s)$ and $\Theta(B^s)$ are seasonal components, and ϵ_t represents white noise. SARIMA effectively captures periodic behaviors and underlying trends, making it a robust choice for forecasting time-dependent medication utilization patterns [12].

2.2 Prophet

Prophet, developed by Facebook, is an additive regression model designed for time series forecasting, particularly suited to datasets with strong seasonality and irregular trends [13]. Unlike SARIMA, which explicitly models time series components using statistical properties, Prophet decomposes time series into distinct components:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t, \tag{2.2}$$

where $g(t)$ represents the trend component, $s(t)$ captures seasonality through Fourier series, $h(t)$ models the effects of external events such as holidays, and ϵ_t accounts for random noise. Prophet allows for flexible trend modeling, supporting both piecewise linear and logistic growth functions, enabling adaptability in situations where abrupt changes occur in prescription trends due to external factors such as policy changes or healthcare interventions.

2.3 XGBoost

XGBoost (eXtreme Gradient Boosting) is an advanced machine learning technique that leverages an ensemble of decision trees to enhance predictive accuracy. Unlike traditional time series models, XGBoost constructs successive trees to minimize residual errors, making it highly effective for capturing complex, non-linear dependencies [14]. The objective function in XGBoost consists of a loss function and a regularization term:

$$\text{Obj} = \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{2.3}$$

where $\mathcal{L}$ represents the loss function, $\Omega(f_k)$ is a regularization term controlling model complexity, and K denotes the number of boosting iterations.

The model's ability to handle large-scale datasets and capture intricate interactions between features makes it well-suited for healthcare applications, where various exogenous factors can influence medication utilization patterns.

2.4 Long Short-Term Memory

Long Short-Term Memory (LSTM) networks are a specialized type of Recurrent Neural Network (RNN) designed to model sequential dependencies in data. Traditional RNNs suffer from vanishing gradient problems when learning long-term dependencies, but LSTMs mitigate this issue through memory cells and gating mechanisms [15].

Each LSTM cell consists of three primary gates that regulate information flow:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \tag{2.4}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{2.5}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{2.6}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C), \tag{2.7}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t, \tag{2.8}$$

$$h_t = o_t * \tanh(C_t), \tag{2.9}$$

where x_t is the input at time t, h_t is the hidden state, C_t is the cell state, and σ denotes the sigmoid activation function. The LSTM architecture enables the retention of information over extended sequences, making it particularly useful for forecasting time series with long-term dependencies. In healthcare analytics, LSTMs provide a powerful framework for capturing intricate prescription utilization patterns that evolve over time.

2.5 Evaluation Metrics

In predictive modeling and machine learning, accurately assessing the performance and accuracy of models is critical. This is often accomplished through various evaluation metrics that quantify the prediction errors. Three widely used metrics are the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE), each

providing different insights into the error characteristics of predictive models.

2.5.1 Root Mean Squared Error

Root Mean Squared Error (RMSE) is a popular metric that measures the average magnitude of the prediction error. It is defined as the square root of the mean of the squared differences between predicted values and actual values, mathematically expressed as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - A_i)^2},$$

where P_i and A_i are the predicted and actual values, respectively, and n is the number of observations. RMSE is particularly sensitive to large errors due to its squaring of prediction errors, which weights larger errors more heavily than smaller ones. This characteristic makes it suitable for applications where large errors are particularly undesirable [16].

2.5.2 Mean Absolute Error

Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of predictions, without considering their direction (positive or negative errors). It is calculated as the mean of the absolute differences between predicted and actual values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|P_i - A_i|.$$

MAE is less sensitive to large errors compared to RMSE, making it a robust measure of errors across various datasets, particularly when the dataset contains outliers [16].

2.5.3 Mean Absolute Percentage Error

Mean Absolute Percentage Error (MAPE) expresses accuracy as a percentage, making it easy to interpret, especially when communicating with non-technical stakeholders. It is calculated as the average of the absolute percentages of errors relative to actual values:

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{P_i - A_i}{A_i}\right|.$$

MAPE is particularly useful for problems where the magnitude of the prediction is significant, such as forecasting and budget planning. However, MAPE can lead to distortions when dealing with values close to zero, as it inherently gives higher weight to percentage errors in low-volume data [17].

2.6 Parameter Optimization Methods

This section covers the theoretical aspects of the Akaike Information Criterion (AIC), the Augmented Dickey-Fuller (ADF) test, and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. These methodologies are crucial for statistical model selection and time series analysis.

2.6.1 Akaike Information Criterion

The Akaike Information Criterion (AIC), developed by Hirotugu Akaike, is a tool for model selection among a finite set of models. It is based on the concept of information entropy, providing a measure of the relative quality of statistical models for a given dataset. The AIC is defined as:

$$\text{AIC} = 2k - 2\ln(L),$$

where k is the number of parameters in the model and L is the maximum likelihood of the model. The criterion estimates the information lost when a given model is used to represent the process that generated the data. A model with the minimum AIC value is preferred as it balances complexity and goodness of fit [18].

2.6.2 Augmented Dickey-Fuller Test

The Augmented Dickey-Fuller (ADF) test is used to test the null hypothesis that a unit root is present in an autoregressive model of a time series. It is an augmentation of the Dickey-Fuller test to include higher order regressive processes in the model:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \phi_i \Delta y_{t-i} + \epsilon_t,$$

where Δy_t is the first difference of the series, α is a constant, βt the trend component, and ϵ_t the error term. The null hypothesis of the ADF test is that $\gamma = 0$, implying a unit root is present [19].

2.6.3 Kwiatkowski-Phillips-Schmidt-Shin Test

The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, unlike the ADF test, tests for the null hypothesis that an observable time series is stationary around a deterministic trend (i.e., it has no unit root). The test statistic compares the series against the model:

$$y_t = \tau_t + \epsilon_t,$$

where τ_t is the deterministic trend (if included in the model) and ϵ_t is the stationary error component. The KPSS statistic is derived from the regression residuals, testing for the stationarity of the series [20].

2.7 Summary

This chapter presents the theoretical foundation of time series forecasting models applied to prescription medication utilization. The Seasonal Autoregressive Integrated Moving Average (SARIMA) model extends the ARIMA framework to account for seasonality, making it well-suited for capturing periodic trends in pharmaceutical data. By incorporating autoregressive, differencing, and moving average components, SARIMA effectively models both short-term fluctuations and long-term trends [10, 11, 12].

Prophet, a decomposition-based model, offers flexibility in handling irregular trends and seasonal patterns through an additive regression framework. Its ability to incorporate external factors, such as policy changes, enhances its adaptability in healthcare forecasting [13]. In contrast, XGBoost employs gradient boosting techniques to capture complex, non-linear dependencies, leveraging an ensemble of decision trees to improve predictive accuracy [14]. The inclusion of regularization ensures robust performance across diverse datasets.

Long Short-Term Memory (LSTM) networks, a class of recurrent neural networks, are particularly effective in modeling long-term dependencies within sequential data. Through gated memory cells, LSTMs mitigate vanishing gradient issues, enabling them to learn temporal patterns in prescription utilization over extended time horizons [15].

Evaluation metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) provide different perspectives on model accuracy. RMSE penalizes large errors,

MAE offers a straightforward measure of average prediction deviation, and MAPE expresses accuracy as a percentage, making it useful for comparative analysis across datasets [16, 17].

Additionally, this chapter covers parameter optimization methods, including the Akaike Information Criterion (AIC) for model selection, the Augmented Dickey-Fuller (ADF) test for stationarity assessment, and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, which evaluates trend stationarity [18, 19, 20]. These statistical tools are essential for ensuring robust model specification in time series analysis.

3. Methodology

This section outlines the data used in this study, the forecasting models applied, and the technical environment in which the experiments were conducted. The data preparation process, selection of models (SARIMA, Prophet, XGBoost, and LSTM), parameter tuning, and model evaluation metrics will be detailed. Finally, the hardware and software infrastructure used for analysis will be discussed.

3.1 Data Description

The data in this study is sourced from the Swedish National Board of Health and Welfare. The dataset contains prescription medication utilization data from January 2009 to December 2023, focusing on Anatomical Therapeutic Chemical Classification (ATC) codes related to heart failure and type-2 diabetes. To ensure a robust analysis, only data columns with at least 25% non-zero values were included, minimizing the impact of sparsely used drugs on the overall results. The dataset was preprocessed to ensure completeness and accuracy.

3.2 Model Selection

In this section we evaluate the forecasting models employed in this study—SARIMA, Prophet, XGBoost, and LSTM—highlighting their advantages, limitations, and justification for inclusion in the context of healthcare time series forecasting.

The **SARIMA** model is a foundational tool in time series analysis, valued for its simplicity, interpretability, and statistical robustness. SARIMA combines trend components, cyclical factors, and random errors, making it highly adaptable and suitable for short-term predictions [21]. Its ability

to account for changing trends, periodic variations, and random disturbances has made it one of the most widely used models in fields such as epidemiology and disease forecasting [22]. However, SARIMA's reliance on stationarity and linearity restricts its applicability to datasets with non-linear patterns or external influences, limiting its utility in capturing the complexities of prescription demand.

Prophet, developed by Facebook, provides a flexible framework that incorporates seasonality, trends, and holidays [23]. Its ability to handle non-stationary data, robustness to missing values, and support for incorporating domain knowledge make it particularly effective for irregular healthcare datasets. Additionally, Prophet does not require evenly spaced time series measurements, making it suitable for handling real-world datasets. However, its reliance on additive decomposition may limit its ability to model non-linear interactions, and it may underperform when compared to machine learning models in datasets with intricate dependencies [24].

XGBoost, a machine learning technique based on gradient boosting, has gained prominence in time series modeling for its ability to handle non-linear relationships and complex datasets. Its minimal need for preprocessing, robust feature extraction, and high predictive accuracy provide clear advantages [25]. XGBoost is particularly effective in healthcare forecasting, where external factors such as policy changes and seasonality significantly influence predictions. Its scalability and efficiency in handling large datasets further enhance its appeal, but it requires preprocessing to capture temporal dependencies and careful hyperparameter tuning to avoid overfitting [26, 27].

LSTM, a type of Recurrent Neural Network (RNN), is specifically designed to model sequential data by retaining long-term dependencies. Memory gates enable LSTM to identify patterns in complex, non-linear datasets, making it highly effective for healthcare applications [6]. Studies have shown that LSTM consistently outperforms traditional models like SARIMA, with improvements in prediction accuracy by up to 85% on average [6]. However, LSTM's computational intensity and susceptibility to overfitting small datasets necessitate robust regularization techniques and careful tuning [15, 28].

Justification for Model Selection

The inclusion of SARIMA, Prophet, XGBoost, and LSTM models is motivated by their complementary strengths in addressing the challenges of healthcare time series forecasting. SARIMA serves as a reliable statistical benchmark for linear trends and seasonality [21, 22], while Prophet provides adaptability for non-stationary data and irregular patterns [23]. XGBoost excels in managing non-linear relationships and external influences, making it suitable for tackling the complexities of healthcare forecasting [25, 26]. LSTM's ability to capture long-term dependencies and model intricate non-linear patterns positions it as a powerful tool for sequential healthcare data [6].

Together, these models form a comprehensive framework for evaluating forecasting accuracy and addressing the intricacies of healthcare data. Their inclusion ensures a balanced analysis of traditional statistical methods and advanced machine learning techniques, contributing valuable insights into the optimal approach for predicting prescription drug demand.

3.3 Evaluation Metrics

Evaluating the performance of forecasting models requires robust and interpretable metrics that reflect the accuracy and reliability of predictions. This study utilizes three widely adopted error metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics were chosen for their complementary strengths and relevance to healthcare forecasting, ensuring a comprehensive assessment of model performance.

RMSE is particularly effective in scenarios where larger prediction errors carry a greater penalty, as it emphasizes significant deviations by squaring the differences between predicted and actual values before averaging. This sensitivity to large errors is beneficial in healthcare contexts, such as medication demand forecasting, where extreme inaccuracies can disrupt resource allocation. However, RMSE's tendency to overemphasize outliers requires careful interpretation, especially in datasets with substantial variability [29].

In contrast, MAE offers a balanced perspective by averaging the absolute differences between predicted and observed values without disproportionately weighting larger errors. Its straightforward interpretation and

robustness to outliers make it suitable for applications requiring consistent error quantification across diverse prediction categories [16]. In healthcare, MAE provides an intuitive measure of average forecast accuracy, aiding decision-makers in evaluating model performance across different medication types or time horizons.

MAPE further enhances interpretability by expressing errors as a percentage of observed values, making it particularly useful for understanding prediction accuracy relative to the scale of the data. This metric is widely applied in fields requiring clear communication of forecast accuracy, including healthcare planning and policy-making [29, 16]. While MAPE's reliance on percentage errors can pose challenges in datasets with values close to zero, its ability to standardize errors across different scales ensures its applicability in this study's analysis of prescription medication demand.

The combination of RMSE, MAE, and MAPE allows for a nuanced evaluation of forecasting models, addressing the varying priorities of error sensitivity, robustness, and interpretability. By leveraging these metrics, this study aims to provide a comprehensive comparison of model performance, contributing valuable insights to healthcare forecasting and decision-making.

3.4 Hardware and Software

The experiments were conducted on a machine equipped with an AMD64 Family 23 processor running at 3.99 GHz, with 16GB of RAM. No GPU acceleration was utilized, but parallel processing was employed for Prophet to optimize computational efficiency. The software environment for the analysis was configured as follows:

- **Operating System:** Microsoft Windows 10 (Version 10.0.19045, 64-bit)
- **Processor:** AMD Ryzen 7 PRO 3700
- **Python Version:** 3.10.11
- **Libraries and Versions:**
  - `pandas` (v2.2.2) for data manipulation and analysis

- `numpy` (v1.26.4) for numerical computations and array handling
- `matplotlib` (v3.7.0) for data visualization
- `statsmodels` (v0.14.0) for time series modeling (ARIMA/SARIMA)
- `pmdarima` (v2.0.4) for automated ARIMA order selection
- `scipy` (v1.10.0) for scientific computing and statistical functions
- `prophet` (v1.1.5) for forecasting with holiday and seasonality effects
- `scikit-learn` (v1.5.0) for machine learning and evaluation metrics
- `xgboost` (v2.0.3) for gradient-boosting decision trees
- `tensorflow` (v2.17.0) for constructing and training neural networks, particularly LSTM models

This methodology provides the computational foundation for running all experiments and conducting the evaluations of the forecasting models on prescription medication utilization.

3.5 Summary

This study investigates optimal forecasting models for prescription medication demand using data from the Swedish National Board of Health and Welfare. The dataset spans January 2009 to December 2023, focusing on Anatomical Therapeutic Chemical (ATC) codes for heart failure and type-2 diabetes. Preprocessing ensured high-quality data by including only columns with at least 25% non-zero values.

Four models—SARIMA, Prophet, XGBoost, and LSTM—were chosen for their complementary strengths. SARIMA serves as a statistical benchmark for linear trends, Prophet handles non-stationary data and irregular patterns, XGBoost captures non-linear relationships, and LSTM models complex sequential dependencies. These models provide a balanced evaluation of traditional and advanced forecasting methods.

Performance was assessed using RMSE, MAE, and MAPE, ensuring a comprehensive evaluation of predictive accuracy. RMSE emphasizes large errors, MAE offers an intuitive average error, and MAPE standardizes errors for easy comparison.

Experiments were conducted using Python 3.10.11 on an AMD Ryzen 7 PRO processor, leveraging libraries like `pandas`, `xgboost`, and `tensorflow`. This methodology ensures a systematic and reliable comparison of forecasting models.

4. Results

4.1 Implementation

This section outlines the implementation details for each predictive model used in this study: SARIMA, Prophet, XGBoost, and LSTM. Each model was configured, trained, and evaluated on historical data for selected ATC codes related to cardiovascular medications. The hyperparameters for each model were tuned based on performance metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The objective was to select the optimal configuration for accurate forecasting of prescription trends.

SARIMA

The implementation of the Seasonal Autoregressive Integrated Moving Average (SARIMA) model in this study followed a systematic process to identify optimal parameters and ensure robust forecasting performance. The procedure was based on the pseudo code outlined in Appendix **??** and applied to the prescription medication dataset, encompassing multiple ATC codes.

The core steps of the SARIMA implementation included data preprocessing, parameter optimization, model fitting, and performance evaluation. Data preprocessing involved filtering the dataset to include only columns with at least 25% non-zero values, ensuring that sparsely used medications did not skew the results. Date columns were converted into a time-series format, with the date serving as the index.

For parameter optimization, initial configurations of SARIMA orders

(p, d, q) and seasonal orders (P, D, Q, m) were generated using the `auto_arima` function from the `pmdarima` library. This automated method tested multiple combinations to minimize the Akaike Information Criterion (AIC). To promote robustness, additional random parameter sets were generated and evaluated to account for potential local minima in the AIC search process. Stationarity tests, such as the Augmented Dickey-Fuller (ADF) and KPSS tests, guided the selection of differencing parameters.

The model was fitted to each ATC code using the best parameter configuration identified during the optimization process. Forecast accuracy was evaluated using three error metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics provided a comprehensive assessment of model performance, emphasizing both error magnitude and relative accuracy.

Post-implementation, results were stored in structured files for analysis, and graphical comparisons of actual versus predicted values were generated to visualize model performance. The entire pipeline was executed in Python, leveraging libraries such as `pandas`, `numpy`, `statsmodels`, and `pmdarima` for data handling, modeling, and optimization.

A detailed pseudo code for the SARIMA implementation process is included below, providing a clear framework for replicating the methodology employed in this study.

1. **Data Preparation:**

- Load the dataset into a DataFrame.
- Filter columns with at least 25% non-zero values.
- Convert the date column into a time-series index.

2. **Parameter Optimization:**

- Use `auto_arima` to identify initial (p, d, q) and (P, D, Q, m) parameters by minimizing the AIC.

- Generate additional random parameter sets to ensure robustness.

3. **Model Fitting and Evaluation:**

- Fit the SARIMA model for each ATC code using optimal parameters.
- Calculate error metrics (RMSE, MAE, MAPE) for each fitted model.

4. **Result Storage and Visualization:**

- Save the model performance metrics and predictions to structured files.
- Generate plots comparing actual and predicted values for visualization.

This structured approach ensured the SARIMA model was robustly implemented and evaluated, contributing to a comprehensive analysis of prescription medication forecasting.

Prophet

The implementation of the Prophet model in this study was designed to optimize forecasting performance for prescription medication demand. The process included data preparation, hyperparameter tuning, model fitting, and evaluation to ensure robust predictions for multiple ATC codes.

Data preparation involved filtering the dataset to include only columns with at least 25% non-zero values, ensuring that rarely prescribed medications did not distort results. The data was partitioned into training and testing sets, with training covering the period from January 2009 to December 2022. Dates were formatted to align with Prophet's requirements, and the model's robustness to missing data eliminated the need for additional imputations.

Hyperparameter tuning was performed using a grid search approach, testing combinations of changepoint prior scale ({0.05, 0.1, 1.0}), seasonality prior scale ({0.01, 0.1, 1.0}), holidays prior scale ({0.01, 0.1, 1.0}), and fixing

the seasonality mode to 'additive.' This grid search allowed the model to adapt to variations in trends, seasonality, and holiday effects. Cross-validation with a rolling window method was employed to evaluate each parameter combination, and the configuration yielding the lowest RMSE was selected as the optimal setup.

For model fitting, the best-performing parameters were applied to train the Prophet model for each ATC code. The model incorporated monthly seasonality using a custom seasonal component with a Fourier order of 5, alongside Swedish public holidays to account for variations influenced by holidays. Predictions were generated for the testing period, spanning January to December 2023.

Model performance was evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics provided a comprehensive assessment of the model's predictive accuracy. The results and optimal parameters were stored for further analysis, while visual comparisons of actual versus predicted values were generated to illustrate performance.

The implementation was conducted in Python, utilizing the Prophet library for forecasting and complementary libraries such as pandas and matplotlib for data handling and visualization. This structured approach ensured that the model was both rigorously tuned and robustly evaluated to deliver reliable forecasts tailored to the characteristics of the dataset.

XGBoost

The implementation of the XGBoost model in this study aimed to capture complex, non-linear relationships within the prescription medication dataset. XGBoost was selected for its ability to handle high-dimensional data, efficiency in feature selection, and scalability, making it suitable for the diverse and intricate patterns observed in the time-series data.

The implementation process began with data preprocessing, where the dataset was filtered to include columns with at least 25% non-zero values. The data was partitioned into training (2009 to 2022) and testing (2023) subsets. For each ATC code, temporal features such as month, quarter, and year were extracted to capture seasonality and long-term trends in prescription patterns. These features formed the input for the XGBoost model.

Hyperparameter tuning was conducted using a grid search combined with time-series cross-validation to account for the sequential nature of the data. The hyperparameters optimized included: - **Learning Rate** ({0.01, 0.05, 0.1, 1.0}): This controlled the step size in the gradient descent process, balancing convergence speed and stability. - **Max Depth** ({3, 5, 7}): The maximum depth of trees allowed the model to capture varying levels of complexity in the data. - **Number of Estimators** ({100, 200}): This defined the number of boosting rounds to prevent overfitting while ensuring robust performance.

The grid search was performed for each ATC code using the ‘GridSearchCV‘ function from Scikit-learn, with negative RMSE as the evaluation metric. This ensured the identification of hyperparameter configurations that minimized prediction errors while respecting temporal dependencies. The best parameter set for each ATC code was selected based on the lowest RMSE achieved during cross-validation.

Using the optimal parameters, the XGBoost model was fitted to the training data for each ATC code. Forecasts for 2023 were generated by constructing a future DataFrame with temporal features extending from January to December 2023. Predictions were evaluated against actual values, and error metrics including RMSE, MAE, and MAPE were calculated to assess model performance comprehensively.

The implementation was executed in Python, leveraging the ‘xgboost‘, ‘pandas‘, and ‘matplotlib‘ libraries. The results, including optimal parameters and error metrics, were stored in structured files, and plots comparing actual and predicted values were generated for visual analysis. This robust implementation ensured that the XGBoost model was well-tuned and effectively captured the dynamics of the prescription medication dataset.

LSTM

The implementation of the Long Short-Term Memory (LSTM) neural network in this study was designed to model sequential dependencies in the prescription medication dataset, leveraging its capacity to capture long-term temporal patterns. The implementation followed a temporal split approach, with data from 2009 to the end of 2022 used for training, and data from 2023 reserved for testing. This approach preserves the sequential nature of the time series, ensuring that predictions are made on unseen

future data and that the results across all the different models remain coherent.

The data was normalized using the `MinMaxScaler`, scaling all values to a range between 0 and 1. A sliding window approach with a window size of 12 months was employed to create input-output pairs, enabling the model to account for seasonality and dependencies over a one-year look-back period.

The LSTM model was implemented using two layers. The first layer contained LSTM cells configured to return sequences, allowing for the extraction of temporal dependencies, while the second layer processed these sequences to generate predictions. A dense output layer was added to produce single-step forecasts. The Adam optimizer was selected for its ability to handle sparse gradients and adapt learning rates dynamically, enhancing the convergence of the model during training.

A grid search was conducted to tune the hyperparameters, ensuring the model was well-suited to the dataset's unique characteristics. The hyperparameters tuned included: - **Units**: The number of memory cells in the LSTM layers, tested with values of 50 and 100. - **Dropout Rate**: A regularization parameter to mitigate overfitting, tested with values of 0.2 and 0.3. - **Learning Rate**: Controlled the step size during gradient descent, tested with values of 0.001, 0.01, and 0.1. - **Epochs and Batch Size**: The number of training iterations was fixed at 50, with batch sizes of 32 and 64 tested to balance training efficiency and model performance.

Predictions for 2023 were generated iteratively using the most recent 12 months of observed or predicted data as input for each subsequent prediction. The model's performance was evaluated using RMSE, MAE, and MAPE, providing a comprehensive assessment of prediction accuracy across different error dimensions. The configuration yielding the lowest RMSE for each ATC code was selected as the optimal model.

The implementation was carried out using Python, leveraging `TensorFlow` for model training and `pandas` and `numpy` for data preprocessing. Results, including predictions and error metrics, were saved for each ATC code, enabling detailed performance analysis. Additionally, visualizations were created to compare actual and predicted values, offering insights into the temporal dynamics captured by the LSTM model.

4.2 Overall Results

SARIMA consistently demonstrated the highest accuracy across the 97 evaluated ATC codes, achieving the lowest RMSE, MAE, and MAPE values. Its ability to explicitly model seasonality and trend dependencies enabled robust forecasting performance, particularly in datasets with stable periodic fluctuations. The model's interpretability and statistical foundation reinforced its reliability for prescription forecasting, especially when historical patterns remained consistent over time.

LSTM also delivered competitive results, particularly in capturing nonlinear dependencies and sequential patterns within the data. Its ability to retain long-term temporal relationships allowed for effective tracking of evolving trends, though its performance was less stable in cases with abrupt demand shifts. The computational complexity of LSTM, combined with its sensitivity to hyperparameter tuning, highlighted both its potential and its limitations in healthcare forecasting applications.

XGBoost performed well in structured datasets but exhibited greater variability when confronted with noisy or highly volatile data. While its tree-based boosting mechanism effectively captured complex interactions, the model's reliance on feature engineering and hyperparameter optimization contributed to inconsistencies in performance. In datasets with abrupt structural changes, XGBoost tended to exhibit higher forecasting errors compared to SARIMA and LSTM.

Prophet, despite its ease of implementation and resilience to missing data, consistently ranked as the least accurate model. Its additive regression framework struggled with high-frequency variations and sudden shifts in prescription trends, leading to larger errors in volatile datasets. While Prophet remains a useful exploratory tool, its limitations in handling non-stationary or highly irregular time series diminished its effectiveness for precise forecasting in this context.

Overall, SARIMA emerged as the most reliable model for prescription forecasting, particularly in datasets exhibiting strong seasonal and trend components. LSTM and XGBoost demonstrated the capacity to model more complex relationships but required careful tuning to optimize their effectiveness. Prophet, while valuable for exploratory analysis, exhibited limitations in adaptability and predictive accuracy. These findings underscore the necessity of model selection based on dataset characteristics, balancing

statistical rigor, computational efficiency, and forecasting precision.

4.3 Detailed Time Series Analysis

Following the general assessment, this section provides a detailed analysis of selected ATC codes to illustrate the performance nuances of each model. Specific ATC code examples demonstrate the models' abilities to handle varied data characteristics, including trend stability, seasonality, and abrupt shifts.

Analyzed ATC Codes and Their Active Ingredients

- **A10AB01**: Insulin (human)
- **A10BB12**: Glimepiride
- **A10AB06**: Insulin glulisine
- **C01CA01**: Etilefrine
- **C09AA05**: Ramipril
- **C09DA06**: Candesartan and Diuretics

4.3.1 ATC Code A10AB01

Prophet, SARIMA, and LSTM closely capture overall prescription trends for A10AB01, effectively tracking the decrease in values and seasonal fluctuations with minimal deviations, as shown in Figure 4.1. SARIMA demonstrates strong accuracy, particularly in reflecting long-term trends and short-term variations, while Prophet and LSTM perform similarly well but exhibit slight overpredictions in 2023. In contrast, XGBoost overestimates prescription values during the prediction period, deviating significantly from observed data. The closer view of the 2020–2024 period in Figure 4.2 highlights these differences, Prophet maintaining the best alignment with the actual values, followed by SARIMA and LSTM, which remain competitive despite minor deviations.

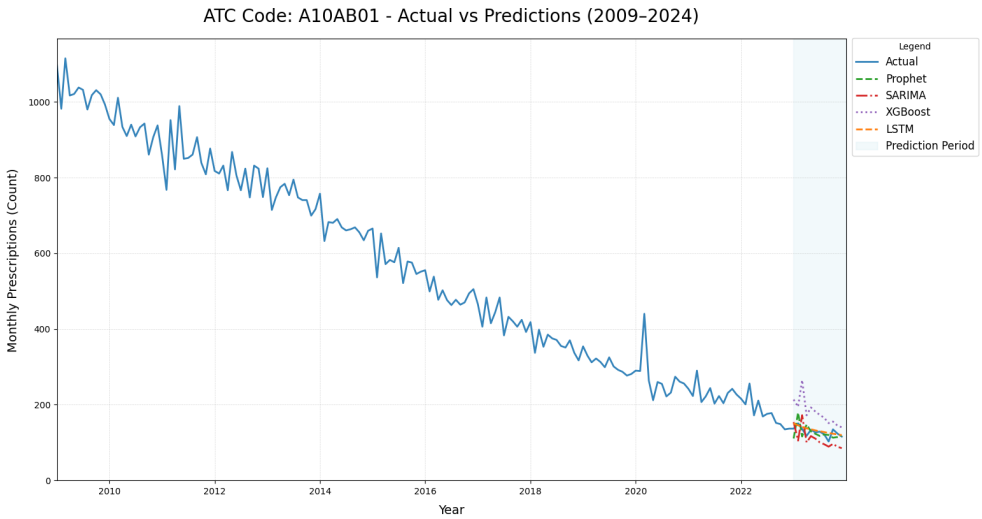

Figure 4.1. Predictions vs Actual Values for ATC Code A10AB01 (2009-2023)

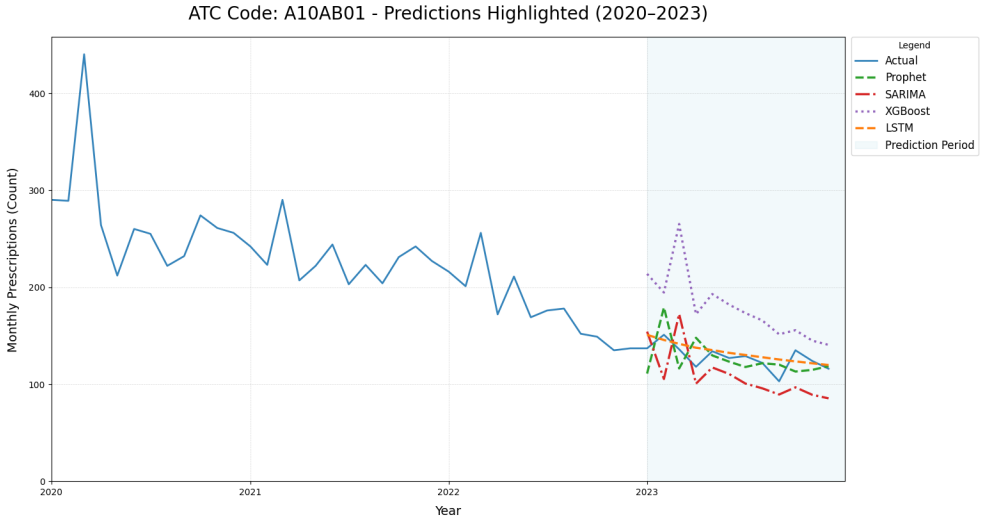

Figure 4.2. Predictions vs Actual Values for ATC Code A10AB01 (2020-2023)

4.3.2 ATC Code C01CA01

As illustrated in Figure 4.3, SARIMA effectively predicts the overall trend for C01CA01, accurately reflecting both the general pattern and the inherent volatility in the data. LSTM captures the broader trend but fails to account for some of the finer market fluctuations, resulting in a less precise fit. In contrast, Prophet and XGBoost tend to underpredict consistently throughout the prediction period. A closer inspection in Figure 4.4 confirms these observations, showing SARIMA's superior alignment with the actual values, particularly in tracking the peaks and troughs.

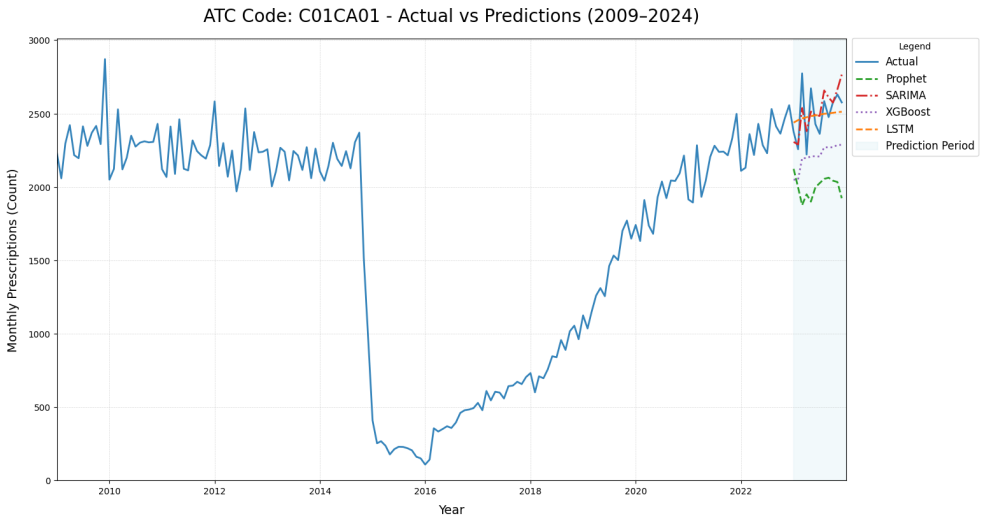

Figure 4.3. Predictions vs Actual Values for ATC Code C01CA01 (2009-2023)

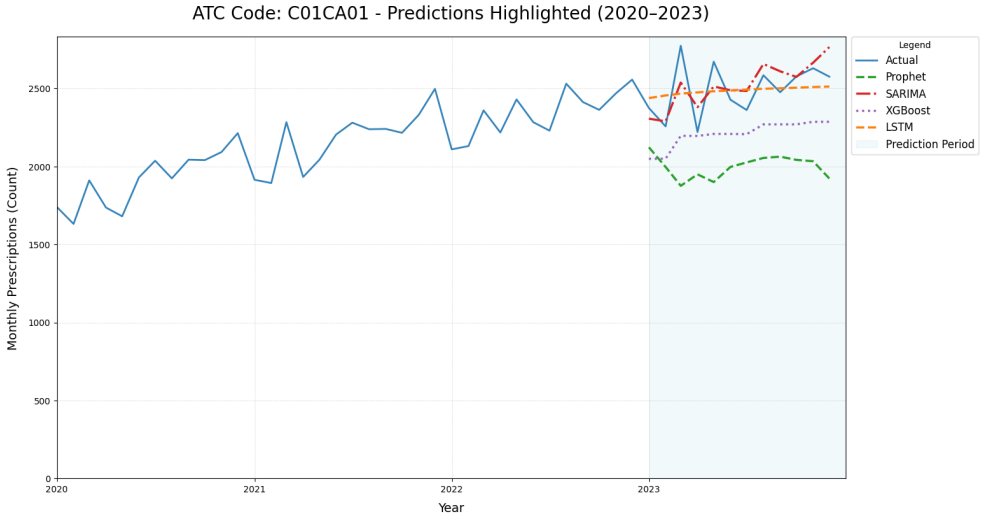

Figure 4.4. Predictions vs Actual Values for ATC Code C01CA01 (2020-2023)

4.3.3 ATC Code C09DA06

For the ATC code C09DA06, all four models provide satisfactory predictions, as shown in Figure 4.5. SARIMA and Prophet converge closely in their predictions toward the end of 2023, demonstrating their ability to align with the overall prescription trend. LSTM follows the mean value of the actual data more consistently, while XGBoost, despite slightly underestimating the prescription values, produces predictions that are relatively well-aligned with the observed data. The closer look provided in Figure 4.6 highlights the comparable performance of all models in capturing the general trend, with minor differences in precision and variability across the models.

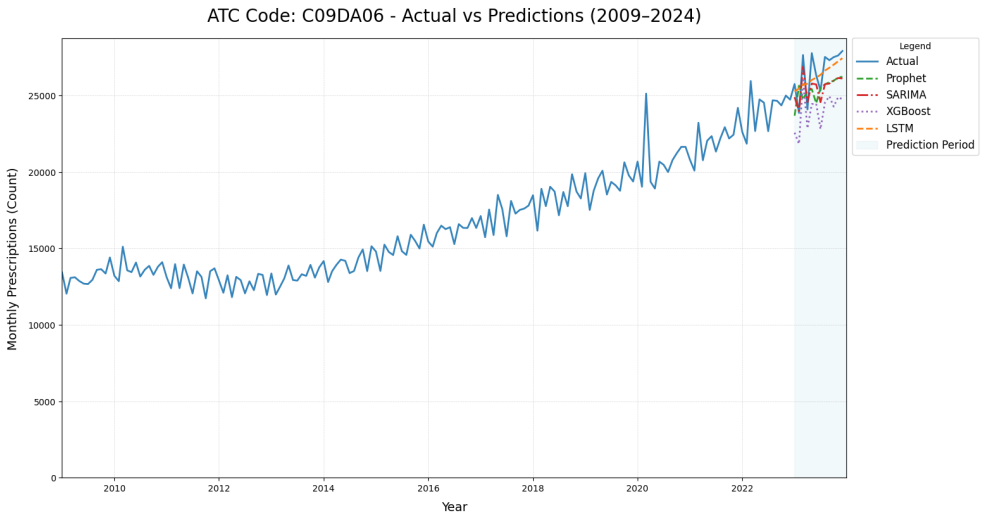

Figure 4.5. Predictions vs Actual Values for ATC Code C09DA06 (2009-2023)

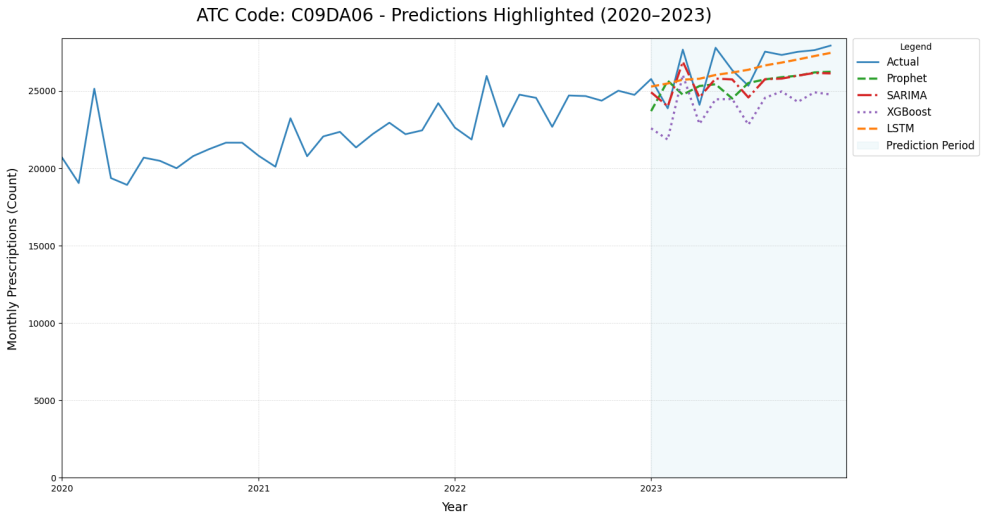

Figure 4.6. Predictions vs Actual Values for ATC Code C09DA06 (2020-2023)

4.3.4 ATC Code A10BB12

The predictions for A10BB12 highlight notable variation among the models, as illustrated in Figure 4.7. The actual data exhibits a stable trend through 2021 and 2022, followed by a sharp decline in 2023. Among the models, SARIMA accurately captures this abrupt downward shift, aligning closely with the observed data. In contrast, Prophet and XGBoost predict an upward trend, resulting in significant overestimations. LSTM, while demonstrating stability, fails to detect the sharp drop in 2023, leading to a less precise representation of the actual trend. Figure 4.8 provides a closer examination of the forecasting period, highlighting these differences in model performance.

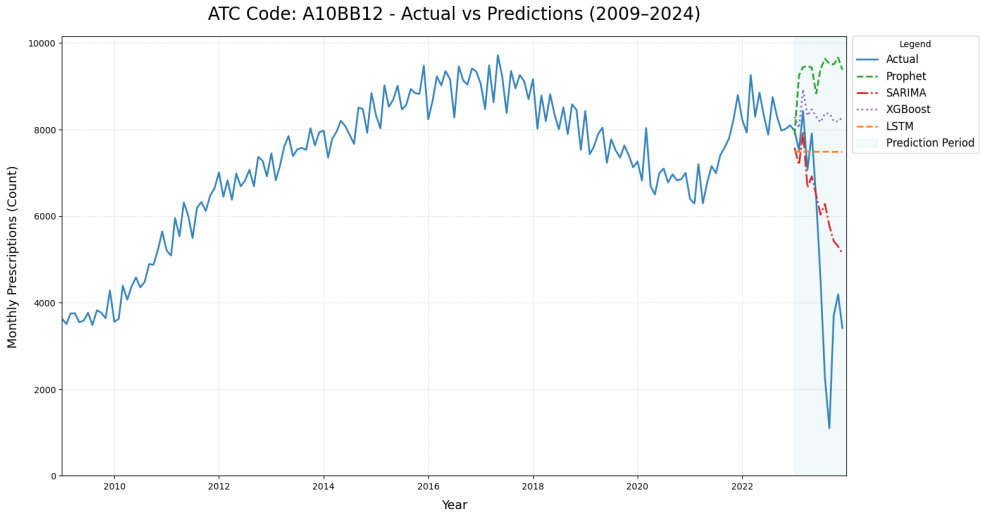

Figure 4.7. Predictions vs Actual Values for ATC Code A10BB12 (2009-2023)

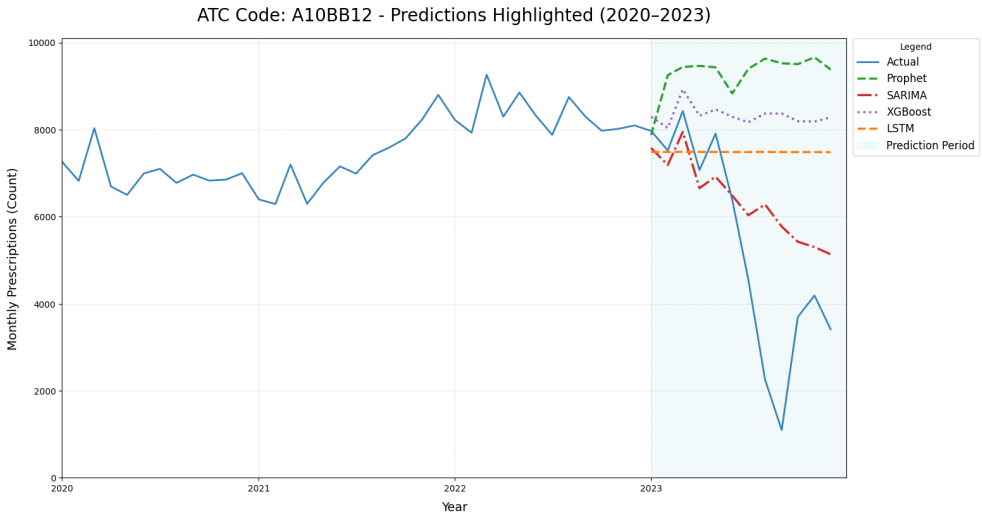

Figure 4.8. Predictions vs Actual Values for ATC Code A10BB12 (2020-2023)

4.3.5 ATC Code C09AA05

For the ATC code C09AA05, SARIMA demonstrates strong performance, as shown in Figure 4.10. It accurately captures both short-term fluctuations and the overall long-term trends in the data. Prophet and LSTM also effectively model the seasonality inherent in the data, although their predictions slightly deviate during periods of sharp changes. XGBoost, while providing a more conservative fit, still aligns reasonably well with the overall trend. Figure **??** offers a detailed view of the forecasting period, further illustrating the relative accuracy of SARIMA compared to the other models.

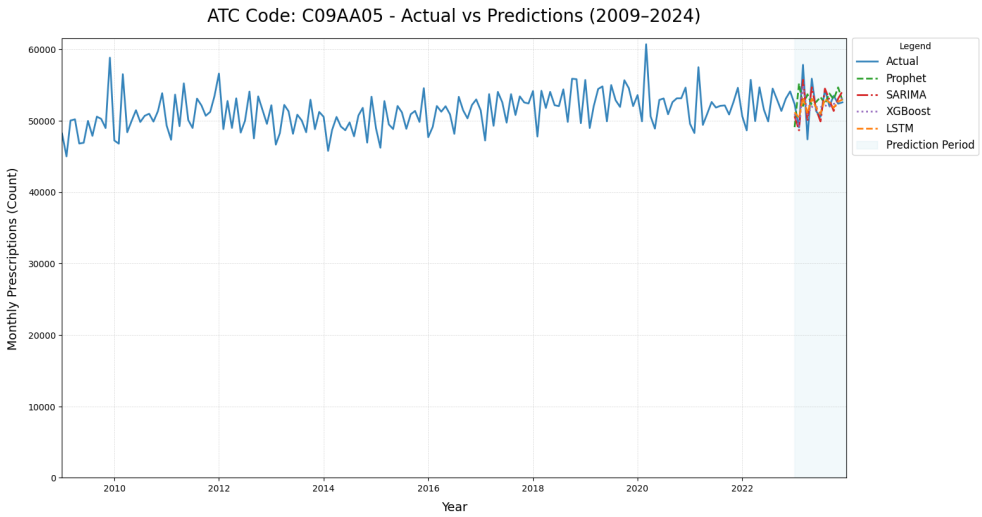

Figure 4.9. Predictions vs Actual Values for ATC Code C09AA05 (2009-2023)

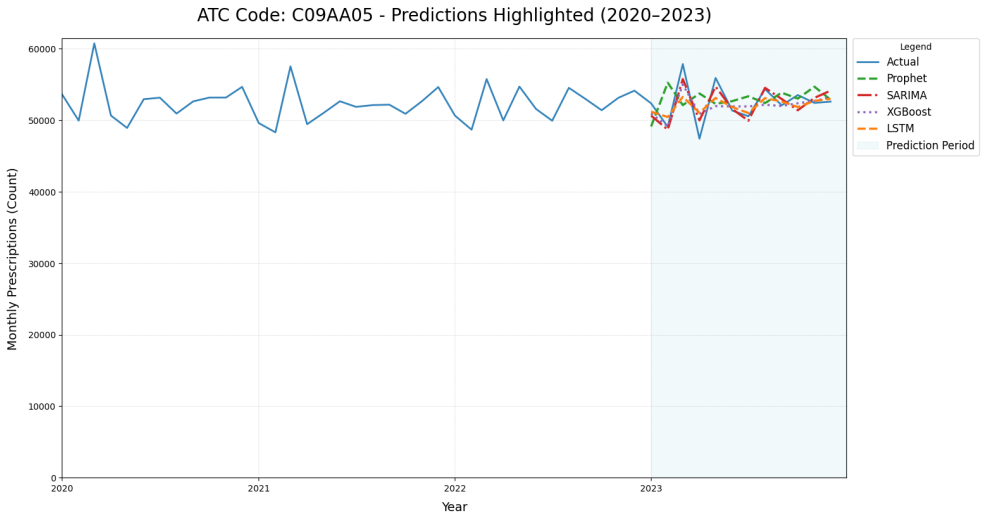

Figure 4.10. Predictions vs Actual Values for ATC Code C09AA05 (2020-2023)

4.3.6 ATC Code A10AB06

For the ATC code A10AB06, LSTM demonstrates the most accurate performance, effectively capturing the stable trend in the data, as shown in Figures 4.12 and **??**. XGBoost and SARIMA exhibit similar variance patterns; however, SARIMA tends to underpredict while XGBoost slightly overpredicts the trend. Prophet underperforms in this case, failing to align with the stable trend and instead predicting a significant decline that does not reflect the actual data. The detailed view in Figure **??** highlights these differences during the prediction period.

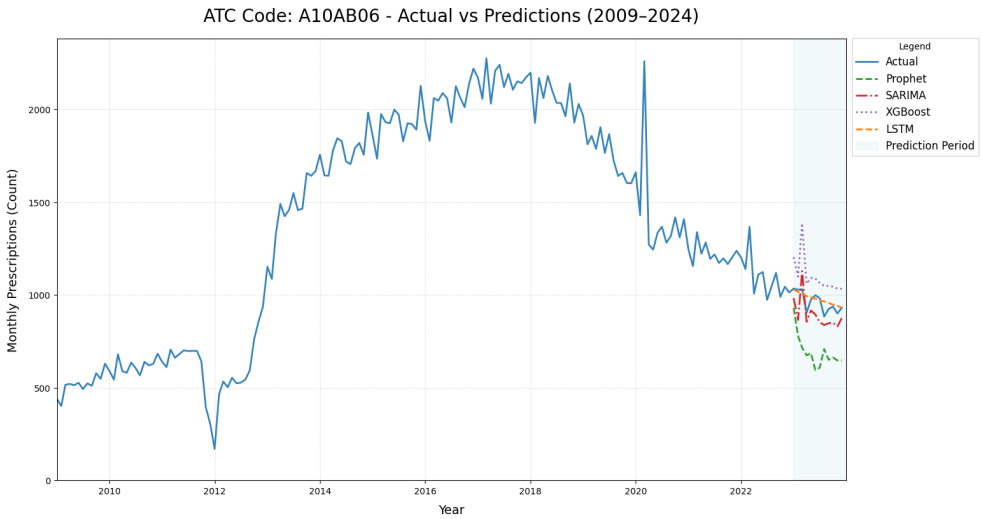

Figure 4.11. Predictions vs Actual Values for ATC Code A10AB06 (2009-2023)

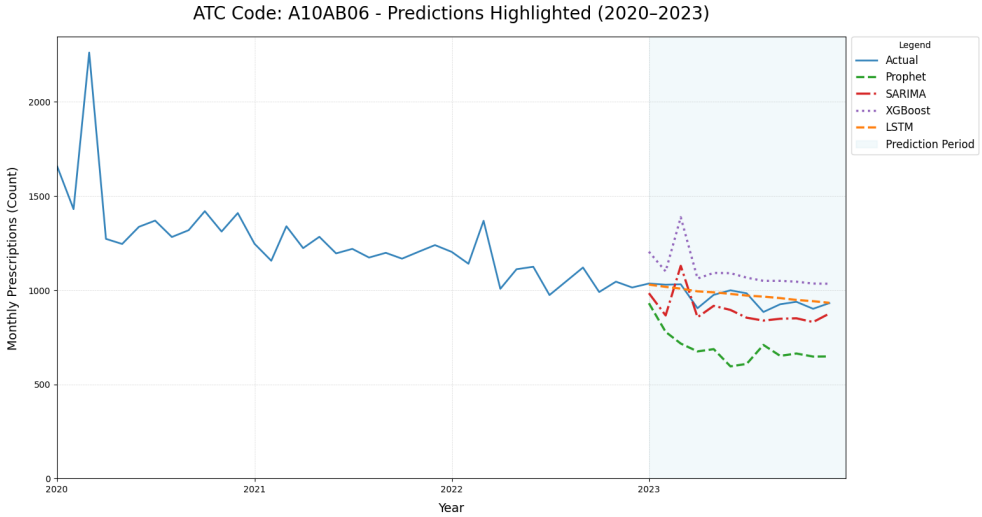

Figure 4.12. Predictions vs Actual Values for ATC Code A10AB06 (2020-2023)

4.4 Model Comparison

This section presents a detailed evaluation of the predictive performance of SARIMA, Prophet, XGBoost, and LSTM across the selected ATC codes. The comparison is supported by quantitative error metrics (RMSE, MAE, and MAPE), parameter configurations for each model, visual comparisons using normalized bar charts, and an aggregated ranking system. The results highlight the relative strengths and limitations of each model, offering insights into their applicability to different datasets and scenarios.

Quantitative Results

The detailed error metrics for each model and ATC code are summarized in Tables 4.1, 4.4, 4.3, and 4.2. These tables outline the RMSE, MAE, and MAPE values along with the best parameter configurations for each model, providing a comprehensive view of their performance.

SARIMA demonstrates its superiority across most ATC codes, achieving the lowest error metrics in three quarters of cases. For instance, in ATC codes such as `C01CA01` and `C09DA06`, SARIMA effectively captures both short-term fluctuations and long-term trends, achieving RMSE values of 123.67 and 1170.92, respectively. Its MAPE performance is consistently low, particularly for datasets with stable seasonality or limited abrupt changes. These results underscore SARIMA's ability to model both seasonality and temporal dependencies explicitly, making it a versatile and reliable choice for healthcare forecasting. Its performance demonstrates the continued relevance of traditional time-series methods in domains where interpretability and seasonality are crucial.

Table 4.1. Selected parameters and resulting errors for SARIMA model

ATC Code	Best Params	RMSE	MAE	MAPE
A10AB01	(5, 1, 1), (0, 1, 1, 12)	27.73	25.81	2.0
A10AB06	(1, 1, 2), (1, 1, 1, 12)	77.31	65.88	7.0
A10BB12	(6, 2, 0), (2, 1, 0, 12)	2022.58	1454.83	6.6
C01CA01	(1, 1, 0), (0, 1, 2, 12)	123.67	85.48	3.0
C09AA05	(0, 0, 0), (0, 1, 0, 12)	1418.01	1187.25	2.0
C09DA06	(3, 1, 2), (3, 1, 0, 12)	1170.92	1071.64	4.0

LSTM ranks as the second-best model, excelling in datasets with non-linear temporal dynamics. For example, in ATC code `A10AB01`, LSTM achieves an RMSE of 10.66, outperforming all other models. Similarly, in `C09DA06`, it achieves an MAE of 949.08. These results highlight LSTM's capability to generalize across datasets with complex patterns, particularly where non-linear dependencies play a significant role. However, LSTM's

performance occasionally lags in datasets with abrupt trend shifts, such as A10BB12, where it struggles to capture sharp changes in the data. This limitation suggests that while LSTM excels in flexibility, it requires careful tuning and significant computational resources, particularly for large datasets.

Table 4.2. Selected parameters and resulting errors for LSTM model

ATC Code	Units	Dropout Rate	Learning Rate	Epochs	Batch Size	RMSE	MAE	MAPE
A10AB01	50	0.00208	0.00069	50	32	10.66	8.19	6.74
A10AB06	100	0.00208	0.00069	50	32	39.64	28.38	3.07
A10BB12	50	0.00208	0.00069	50	32	3178.22	2425.30	102.05
C01CA01	50	0.00208	0.00069	50	32	155.58	130.87	5.27
C09AA05	100	0.00139	0.00069	50	32	2052.15	1552.37	2.93
C09DA06	100	0.00139	0.001	50	32	1126.41	949.08	3.63

XGBoost, while competitive in certain cases, generally underperforms compared to SARIMA and LSTM. Its difficulty in handling complex temporal dependencies is evident in A10BB12, where it exhibits a high RMSE of 3756.93. Despite this, XGBoost performs adequately in datasets with simpler seasonal patterns or minimal volatility. This model's computational efficiency and ease of tuning make it a viable option for less complex datasets, but its struggles with datasets featuring abrupt shifts and intricate seasonality limit its applicability in certain healthcare scenarios.

Table 4.3. Selected parameters and resulting errors for XGBoost model

ATC Code	Learning Rate	Max Depth	N Estimators	RMSE	MAE	MAPE
A10AB01	0.1	3	200	58.87	51.68	40.30
A10AB06	0.1	3	100	157.20	139.74	14.51
A10BB12	0.05	3	200	3756.93	2951.60	120.33
C01CA01	0.01	5	200	317.01	285.96	11.20
C09AA05	0.01	5	200	1971.59	1504.21	2.84
C09DA06	1.0	3	100	2608.59	2520.47	9.44

Prophet consistently ranks last across all metrics. Its limitations are most apparent in datasets with abrupt changes or high volatility. For example, in C09AA05, Prophet's RMSE (3512.56) and MAPE (5.52) substantially lag behind SARIMA and LSTM. While Prophet's reliance on additive seasonal-

ity and its simplified approach to modeling temporal patterns contribute to its weaker performance, it remains an accessible tool for practitioners seeking a straightforward forecasting method. Its computational efficiency and ease of implementation make it suitable for exploratory analyses or cases where model interpretability is a priority.

Table 4.4. Selected parameters and resulting errors for Prophet model

ATC Code	Changepoint Prior Scale	Seasonality Prior Scale	Holidays Prior Scale	Seasonality Mode	RMSE	MAE	MAPE
A10AB01	0.1	0.01	0.01	Additive	19.62	15.07	11.90
A10AB06	1.0	0.01	0.1	Additive	133.48	121.32	12.57
A10BB12	1.0	0.01	0.1	Additive	4366.55	3519.40	138.20
C01CA01	1.0	0.01	0.01	Additive	228.64	201.25	8.25
C09AA05	0.1	0.01	0.01	Additive	3512.56	2877.22	5.52
C09DA06	0.1	0.01	0.01	Additive	1717.05	1617.97	6.09

Trade-offs and Computational Considerations

The comparative performance of these models highlights trade-offs between accuracy, computational cost, and ease of use. SARIMA emerges as the most robust model overall, particularly for datasets with clear seasonal and trend components. However, its performance in highly non-linear cases is occasionally surpassed by LSTM, which demonstrates superior adaptability to datasets with complex patterns. This adaptability comes at the cost of longer training times and higher computational demands, making LSTM a resource-intensive option.

XGBoost offers a middle ground between computational efficiency and accuracy, performing moderately well in simpler datasets but struggling with high volatility and abrupt trend shifts. Prophet, while less competitive in terms of accuracy, remains a valuable tool for users prioritizing interpretability and rapid implementation. These trade-offs emphasize the importance of model selection based on the specific characteristics of the dataset and the forecasting requirements.

Visual Comparisons

Figures 4.13, 4.14, and 4.15 present heatmaps that summarize the normalized RMSE, MAE, and MAPE values for all models across the selected ATC codes. These heatmaps provide an intuitive way to compare the performance of each model on a relative scale. In the heatmaps, the normalized values are scaled such that 0.0 represents the best performance (lowest

error) and 1.0 represents the worst performance (highest error) for each ATC code. Intermediate values reflect the relative position between the best and worst performers, offering a clear visual representation of each model's effectiveness.

SARIMA consistently achieves the lowest normalized metrics, particularly in ATC codes such as `A10BB12` and `C01CA01`, where its performance is closest to 0.0. LSTM also demonstrates strong performance, especially in `A10AB01` and `C09DA06`, with competitive normalized error values. These results highlight the ability of SARIMA and LSTM to effectively capture the temporal and seasonal dynamics in these datasets.

XGBoost, while moderately successful in certain ATC codes, such as `C09AA05`, shows greater variability in performance. It struggles with datasets featuring abrupt shifts or irregular patterns, as evidenced by its higher normalized error values in `A10BB12`. Prophet consistently ranks last or close to last across all error metrics, with notably high normalized MAPE values in `A10BB12` and `C09AA05`. These results emphasize Prophet's limitations in adapting to complex temporal patterns, especially in datasets with abrupt changes or significant volatility.

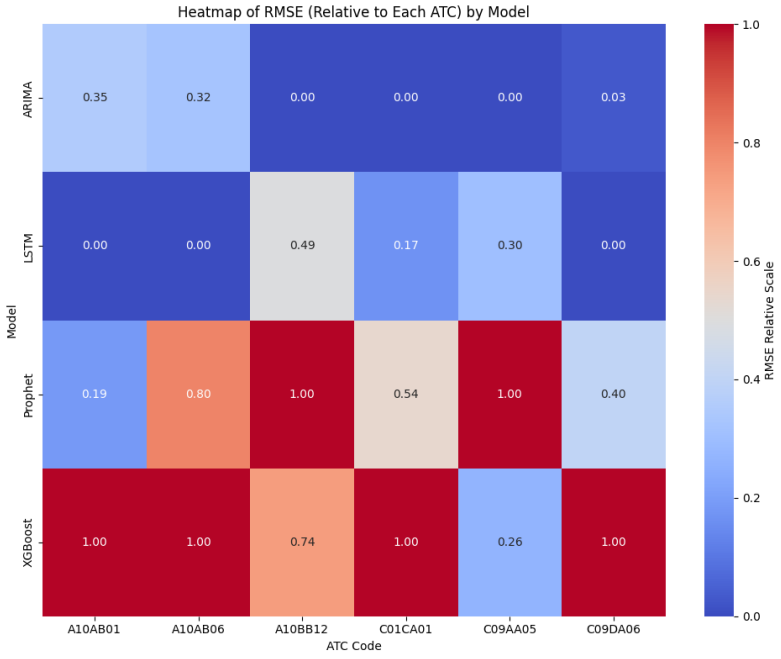

Figure 4.13. Normalized RMSE Comparison Across Models

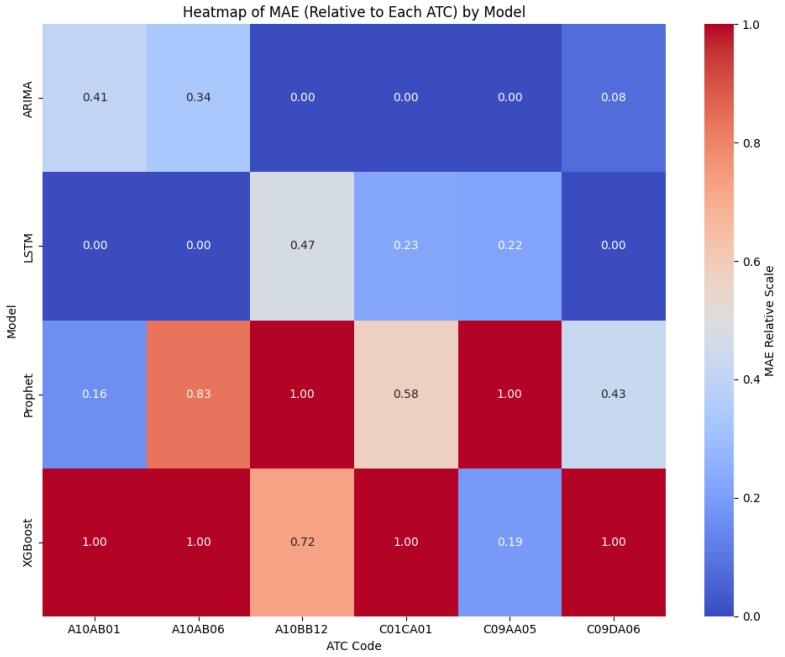

Figure 4.14. Normalized MAE Comparison Across Models

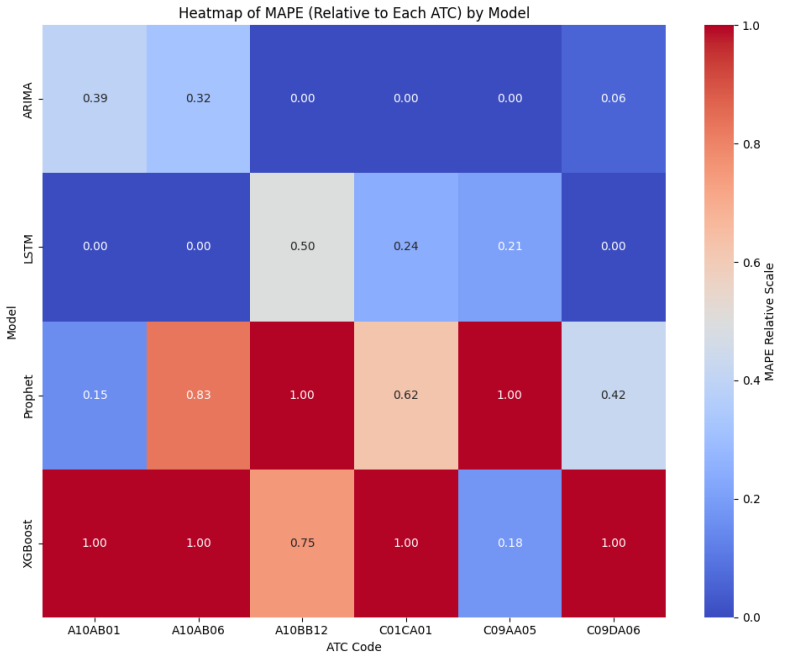

Figure 4.15. Normalized MAPE Comparison Across Models

Overall Model Rankings

The cumulative performance of the models is summarized in the ranking distribution chart (Figure 4.16). Rankings were determined based on each model's relative performance in RMSE, MAE, and MAPE for all 97 ATC codes, with the number of first-, second-, third-, and fourth-place rankings aggregated across all metrics. A normalized scoring system highlights the overall competitiveness of the models.

SARIMA emerges as the clear leader, securing the highest number of first-place rankings, particularly excelling in datasets characterized by seasonal patterns and consistent trends. Its dominant performance demonstrates its robustness and adaptability across diverse healthcare-related time series datasets.

LSTM ranks second overall, showing strong performance in datasets with non-linear dependencies. Its ability to handle complex temporal dynamics makes it a valuable tool for healthcare forecasting, though it occasionally underperforms in datasets with abrupt changes or high volatility. XGBoost, occupying third place, performs reasonably well in datasets with limited complexity but struggles with noisy or volatile time series, highlighting its sensitivity to hyperparameter tuning. Prophet ranks last, with a majority of its rankings falling into third or fourth place. Its limitations in addressing non-linear seasonality and sharp trend shifts are evident, though it remains an accessible option for simpler forecasting tasks.

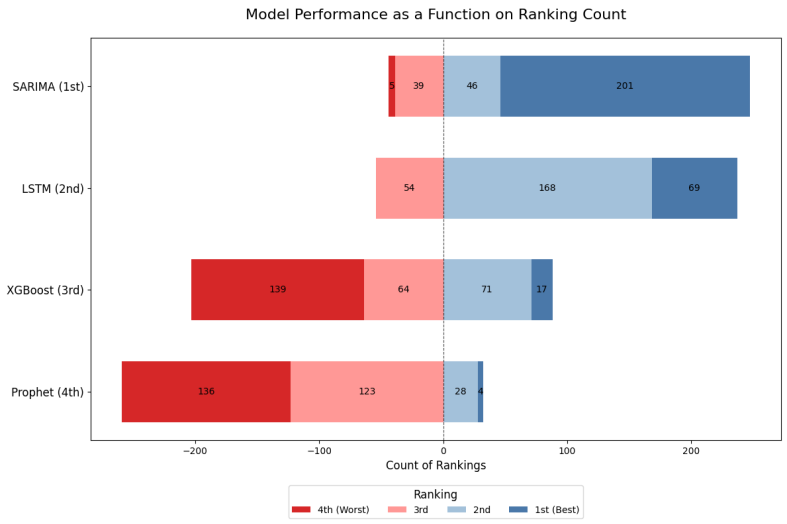

Figure 4.16. Model Performance as a Function of Ranking Count Across ATC Codes

This analysis underscores the importance of selecting a forecasting model that aligns with the dataset's characteristics. While advanced models like LSTM and XGBoost offer flexibility for certain scenarios, SARIMA consistently demonstrates its reliability and effectiveness, solidifying its place as a cornerstone for healthcare time series forecasting.

4.5 Summary

This chapter provides a detailed evaluation of four forecasting models, SARIMA, Prophet, XGBoost, and LSTM, applied to a data set of 97

cardiovascular-related ATC codes. Each model was implemented following a rigorous process, including data preprocessing, hyperparameter optimization, and performance evaluation, with the error metrics RMSE, MAE, and MAPE used to assess accuracy comprehensively.

Among the models, SARIMA demonstrated the highest overall performance, achieving the lowest error metrics in three quarters of cases. Its ability to effectively capture seasonality, trends, and short-term fluctuations underscores its reliability for datasets with periodic or cyclical patterns, particularly in healthcare forecasting. LSTM ranked as the second-best model, excelling in datasets with non-linear temporal dependencies, though its performance was occasionally hindered by abrupt trend shifts. The computational complexity and tuning demands of LSTM were significant trade-offs for its flexibility.

XGBoost showed competitive performance in simpler datasets but struggled with high volatility and abrupt changes, reflecting its sensitivity to noise and dependence on optimized hyperparameters. Prophet, despite its computational efficiency and ease of implementation, ranked last overall due to its reliance on additive seasonality and limitations in handling complex or volatile patterns.

This chapter highlights the trade-offs between model accuracy, computational efficiency, and implementation complexity. SARIMA proved to be the most robust and versatile model, while LSTM and XGBoost offered value in specific contexts. The findings emphasize the importance of selecting models aligned with dataset characteristics and forecasting requirements to achieve optimal results. Visual comparisons and detailed rankings further illustrate the relative strengths and limitations of each approach.

5. Discussion

5.1 General Observations

The comparative analysis of SARIMA, Prophet, XGBoost, and LSTM provides valuable insights into their suitability for healthcare forecasting, particularly in predicting prescription drug utilization. The study revealed that each model demonstrates unique strengths and limitations, influenced heavily by the characteristics of the time series data and the underlying assumptions of the algorithms. This section synthesizes these observations to contextualize their broader scientific and practical implications.

SARIMA emerged as the most reliable model for datasets with strong seasonality and relatively stable trends, excelling in capturing linear and periodic patterns. Its explicit handling of autoregressive, moving average, and seasonal components allowed it to consistently achieve the lowest error metrics in most ATC codes. These results reaffirm SARIMA's relevance in healthcare analytics, where interpretability and the ability to model seasonal dependencies are paramount. However, its reliance on linear assumptions limited its efficacy in datasets with pronounced non-linear dynamics, such as those observed in ATC C01CA01, highlighting the need for complementary approaches.

Prophet, while robust in handling seasonal data and missing values, underperformed in cases with significant volatility or abrupt trend changes. Its additive regression framework proved effective in simpler time series, but it struggled to adapt to complex temporal patterns influenced by external factors such as policy changes or public health events. Despite these limitations, Prophet's ease of implementation and interpretability make it a valuable tool for exploratory analyses and less complex forecasting tasks.

XGBoost demonstrated its ability to capture non-linear relationships, excelling in scenarios with intricate temporal dynamics. However, its susceptibility to overfitting and sensitivity to noise limited its generalizability across certain datasets. The results underscore the importance of careful hyperparameter tuning and feature engineering when applying gradient boosting methods to healthcare data. These findings align with prior studies [26, 30] that have highlighted XGBoost's adaptability in diverse forecasting contexts but also its challenges in balancing complexity and robustness.

LSTM, designed to model long-term dependencies, performed well in datasets with smooth trends but struggled with high volatility and sudden changes. Its capability to learn from sequential dependencies allowed it to track general trends effectively, yet it often failed to capture abrupt fluctuations, as seen in ATC A10BB12. These limitations emphasize the computational and data requirements of deep learning models in healthcare forecasting, while also highlighting their potential for improvement through hybrid approaches or attention mechanisms.

One key observation is the complementarity of these models. No single algorithm consistently outperformed the others across all ATC codes, indicating that the choice of model should be guided by the specific characteristics of the dataset and the forecasting objectives. This aligns with broader findings in predictive analytics, where hybrid methodologies that leverage the strengths of multiple models have shown promise. By combining the linear interpretability of SARIMA with the non-linear adaptability of XGBoost and LSTM, future research can advance the precision and applicability of healthcare forecasting tools.

These findings contribute to the scientific understanding of model performance in healthcare time series, offering a foundation for future research aimed at improving resource management and patient care through advanced predictive analytics. They also underscore the critical importance of tailoring model selection to the unique demands of healthcare datasets, fostering collaboration between domain experts and data scientists to achieve optimal outcomes.

5.2 Evaluation of Error Metrics

In this study, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) were used as the primary metrics to evaluate the performance of the forecasting models. While these metrics are commonly used in time series forecasting due to their ability to measure the differences between predicted and actual values, it is important to assess whether they are optimal for this specific use case and whether other evaluation methods might provide additional insights.

Root Mean Squared Error (RMSE)

RMSE is a widely-used metric that penalizes larger errors more severely due to the squaring of the differences between predicted and actual values. This makes RMSE particularly effective in contexts where large deviations from the true values are highly undesirable. However, its sensitivity to outliers can sometimes make it less suitable for data that exhibits high volatility or sudden shifts, such as in healthcare data influenced by external factors like regulatory changes or public health events [29]. In this study, RMSE proved useful in identifying models that tended to deviate significantly from the true values (e.g., Prophet in ATC A10BB12).

However, in cases where outliers or large variances are present, alternative metrics such as the Median Absolute Deviation (MAD) or Huber loss, which are less sensitive to extreme errors, could provide a more robust evaluation of model performance. These metrics could be considered in future studies to complement RMSE, particularly in datasets with irregular patterns.

Mean Absolute Error (MAE)

MAE provides an intuitive interpretation of the average prediction error, giving equal weight to all errors regardless of their magnitude. This makes MAE less sensitive to outliers compared to RMSE and thus a useful metric in contexts where moderate errors are equally as important as larger ones. For example, in ATC codes such as A10AB06, where the prescription trends are relatively stable, MAE provides a good representation of the overall model accuracy.

However, one drawback of MAE is that it does not provide the same

level of penalization for large errors as RMSE, which might underplay significant deviations [29]. For models like XGBoost, which in some cases overestimated future prescriptions, a combination of MAE and RMSE is essential to capture both moderate and extreme errors.

Mean Absolute Percentage Error (MAPE)

MAPE is a useful metric when evaluating time series data, as it expresses the prediction error as a percentage of the actual value. This makes it easier to compare performance across different scales, which is particularly useful in this study given the varying prescription volumes across different ATC codes. For instance, the high values in ATC C09AA05 are better interpreted with a percentage-based metric like MAPE than with absolute measures like RMSE or MAE [31].

However, MAPE has some well-documented limitations. It tends to produce extremely high values when actual values are close to zero, which can distort the overall evaluation of a model's performance. This is particularly problematic in datasets where there are sudden drops in the prescription rates, as seen in ATC A10BB12. To address this, a logarithmic error metric (such as Symmetric Mean Absolute Percentage Error, SMAPE) could provide a more reliable evaluation when dealing with small actual values, thereby avoiding the skew that MAPE can introduce [31].

Alternative Metrics and Future Considerations

While RMSE, MAE, and MAPE are commonly accepted and effective in many time series forecasting contexts, there are additional metrics that could provide further insights into model performance:

- **Symmetric Mean Absolute Percentage Error (SMAPE)**: This metric resolves some of the issues of MAPE, particularly when dealing with low or zero values, by using a symmetric formula that limits extreme percentages [31]. - **Mean Squared Logarithmic Error (MSLE)**: MSLE can be useful in cases where the data exhibits exponential growth or decline, as it penalizes under-predictions more heavily than over-predictions [32]. - **Dynamic Time Warping (DTW)**: For more complex time series data with shifting patterns, DTW could offer an alternative metric that compares time series based on the shape of the curves rather than point-by-point differences.

Given the high variability and complexity of prescription data, incorporat-

ing a wider array of error metrics could lead to a more holistic evaluation of model performance.

5.3 Train-Test Split Strategy and Hyperparameter Tuning

The choice of train-test split is another key consideration in time series forecasting. In this study, the models were trained using data from the years up to 2022, with predictions made for 2023. While this approach reflects the real-world scenario of forecasting future data based on past trends, it is important to discuss whether this strategy is optimal.

Fixed-Time Split

Using a fixed-time split where the entire 2023 dataset is held out for validation mirrors a practical scenario where future predictions are required. This approach ensures that no future data leaks into the training set, which is critical for avoiding overfitting. However, this strategy may suffer from variance depending on how different the test set is from the training set. For instance, if 2023 contained sudden shifts in prescription behavior due to external factors (e.g., a new health policy), the models trained exclusively on pre-2023 data might not perform as well.

Alternative Split Approaches

An 80/20 split, which randomly divides the dataset into 80% training and 20% testing, is commonly used in many machine learning tasks. However, in time series forecasting, such a random split could lead to leakage of future data into the training set, which would artificially inflate model performance and reduce the ability to generalize [33].

One potential solution is the use of rolling forecast origin (also known as time series cross-validation), where the training set is expanded as each new time point is added to the test set. This method ensures that predictions are made only on unseen data while maximizing the use of available data for training [33].

Hyperparameter Grid Size

A larger hyperparameter grid increases the likelihood of finding the best model configuration but also exponentially increases computation time. In this study, the grid sizes were chosen to balance computational feasibility and the likelihood of finding a good solution. For models like XGBoost,

which are prone to overfitting, a fine-tuned grid that includes regularization parameters (e.g., α and λ) is essential to control for overfitting, particularly on noisy datasets like ATC C09DA06.

In future work, the use of automated hyperparameter tuning techniques, such as Bayesian Optimization or Random Search, could help balance computational cost and model performance. These methods would allow for more efficient exploration of the hyperparameter space, leading to potentially better results without the need for exhaustive grid searches.

Evaluation of Hyperparameters Across Models

The relative simplicity of SARIMA and Prophet models means that the tuning process was relatively straightforward, focusing mainly on seasonal and trend parameters. However, in the case of XGBoost and LSTM, the number of possible hyperparameter configurations was much larger, and the performance of these models was highly sensitive to the chosen parameters.

For XGBoost, parameters such as the learning rate, depth of the trees, and number of estimators played a crucial role in its ability to fit the data without overfitting. LSTM, on the other hand, was sensitive to the number of hidden units and dropout rates, which affected its ability to generalize to unseen data. A more exhaustive exploration of these hyperparameters could further improve the models' performance in future studies.

5.4 Comparison with Previous Studies

The comparison of our findings with existing literature reveals interesting insights into the application and performance of forecasting models, particularly in the context of healthcare data. Our results demonstrate that SARIMA outperforms other models, such as Prophet, XGBoost, and LSTM, in forecasting prescription trends, especially for datasets with clear seasonality and stable trends. This aligns with findings by Fattah et al. [7], who highlight the robustness of ARIMA models for time-series forecasting when seasonality and linear patterns dominate.

However, our study also confirms limitations of SARIMA in handling non-linear patterns and abrupt changes, as observed in datasets like ATC code A10BB12. In contrast, LSTM's ability to model complex, non-linear dependencies has been well-documented, particularly in studies like Siami-

Namini et al. [6], where LSTM consistently outperformed ARIMA by reducing error rates by 84-87%. In our results, LSTM ranked second overall, excelling in capturing non-linear dynamics for datasets such as ATC code A10AB01, yet its performance faltered when abrupt shifts occurred, echoing the computational and tuning challenges identified by others [34, 6].

Prophet's underperformance in our study, particularly in volatile datasets, is consistent with its limitations reported in prior research, such as Motamedi et al. [30], where Prophet struggled to capture abrupt shifts in platelet demand despite its simplicity and robustness to missing data. Similarly, studies like Jain et al. [35] found hybrid models, combining ARIMA with LSTM, performed better than Prophet in volatile environments. This suggests Prophet's additive framework may lack the flexibility needed for complex healthcare datasets.

XGBoost, while competitive in simpler datasets, often lagged in performance compared to SARIMA and LSTM. Similar observations were noted in studies like Chen et al. [34], where XGBoost excelled in specific agricultural datasets but faced challenges in datasets with high noise and abrupt changes. The sensitivity of XGBoost to hyperparameter tuning and noise is a recurring theme, further substantiating our findings.

The broader implications of these comparisons highlight the importance of dataset-specific model selection. In healthcare forecasting, where seasonality and trend patterns often dominate, SARIMA continues to be a reliable choice. However, for datasets with significant non-linear components or abrupt trend shifts, hybrid approaches or advanced deep learning models like LSTM offer greater potential. This underscores the conclusions of Motamedi et al. [30] and Jain et al. [35], who advocate for a combination of traditional and modern techniques to address diverse dataset characteristics.

Our findings, while consistent with prior literature, provide a unique contribution by applying these models specifically to prescription datasets categorized by ATC codes. The detailed evaluation of 97 cardiovascular-related ATC codes offers a granular perspective on model performance across diverse data scenarios, enhancing the existing body of knowledge in healthcare forecasting.

5.5 Summary

This chapter discussed the performance of SARIMA, Prophet, XGBoost, and LSTM in forecasting prescription drug utilization, highlighting their unique strengths and limitations. SARIMA consistently excelled in datasets with clear seasonality and stable trends, reaffirming its reliability for healthcare forecasting where interpretability and periodic patterns are crucial. However, it struggled with non-linear dynamics, underscoring the need for hybrid approaches in such cases.

LSTM demonstrated strong performance in capturing non-linear patterns and long-term dependencies but faced challenges with abrupt trend shifts and required significant computational resources. XGBoost, while adaptable to non-linear relationships, was sensitive to noise and overfitting, limiting its robustness in volatile datasets. Prophet, though easy to implement and robust to missing data, underperformed in complex scenarios, reflecting its limited flexibility in addressing irregular patterns.

The findings emphasize that no single model consistently outperformed others, suggesting the need for dataset-specific approaches. Combining the strengths of these models offers a promising avenue for improving accuracy in healthcare forecasting. This study contributes valuable insights into tailoring predictive analytics to meet the unique challenges of healthcare datasets, enhancing decision-making and resource optimization.

6. Conclusion and Future Work

This study evaluated the performance of SARIMA, Prophet, XGBoost, and LSTM models in forecasting prescription trends across 97 cardiovascular-related ATC codes using Swedish prescription data from 2009 to 2023. By systematically analyzing the models' capabilities, this research aimed to identify the most effective approach for improving healthcare resource allocation and decision-making.

The results highlight SARIMA as the most reliable model overall, excelling in datasets with pronounced seasonality and linear trends. Its capacity to capture autoregressive and seasonal components allowed for accurate predictions across the majority of ATC codes, reaffirming its established utility in time series forecasting. However, SARIMA's limitations in handling non-linear dynamics and abrupt changes emphasize the need for complementary approaches when faced with complex datasets.

LSTM demonstrated promise in capturing non-linear patterns and long-term dependencies, particularly in datasets with smoother trends. However, its performance declined in scenarios with high volatility or sudden trend shifts, pointing to challenges in hyperparameter tuning and computational overhead. XGBoost, while effective in some cases, was prone to overfitting, especially in volatile datasets, underscoring the need for advanced regularization techniques. Prophet, though robust in handling missing data and interpretable, underperformed in datasets with non-linear or abrupt changes, limiting its applicability in healthcare forecasting.

The error metrics—RMSE, MAE, and MAPE—provided valuable insights but also revealed limitations in capturing model performance comprehensively. Future work could benefit from incorporating alternative metrics, such as SMAPE or MSLE, to address issues with sensitivity to outliers or small actual values. Additionally, while the fixed-time split used in this

study reflects practical forecasting scenarios, exploring rolling forecast origin or time series cross-validation techniques could further enhance the generalizability of results.

Future Work

The findings of this study suggest several promising avenues for future research. Firstly, exploring hybrid models that combine SARIMA's interpretability with LSTM's adaptability could provide robust solutions for diverse datasets. Incorporating exogenous variables, such as policy changes or demographic data, may improve model performance by accounting for external influences on prescription trends. Moreover, real-time data integration and automated hyperparameter tuning through techniques like Bayesian optimization could enhance model accuracy and scalability.

Lastly, expanding the scope of analysis to include other therapeutic areas or global datasets could validate the findings in diverse healthcare contexts. By addressing these aspects, future research can further bridge the gap between traditional and advanced forecasting methods, contributing to more effective and adaptable healthcare analytics.

In conclusion, this study underscores the critical importance of model selection based on dataset characteristics and forecasting objectives. While SARIMA remains a robust and interpretable option, the potential of machine learning models, particularly in handling complex temporal patterns, offers significant opportunities for advancing healthcare forecasting and improving patient outcomes.

7. References

[1] A. Rajkomar, J. Dean, and I. Kohane. “Machine Learning in Medicine”. In: *New England Journal of Medicine* 380 (2019), pp. 1347–1358.

[2] Z. Obermeyer and T. H. Lee. “Lost in Thought — The Limits of the Human Mind and the Future of Medicine”. In: *New England Journal of Medicine* 377.13 (2017), pp. 1209–1211. DOI: 10.1056/NEJMp1705348.

[3] D. S. Char, N. H. Shah, and D. Magnus. “Implementing Machine Learning in Health Care - Addressing Ethical Challenges”. In: *New England Journal of Medicine* 378.11 (2018), pp. 981–983. DOI: 10.1056/NEJMp1714229.

[4] N. Dritsakis and P. Klazoglou. “Time Series Analysis Using ARIMA Models: An Approach to Forecasting Health Expenditures in USA”. In: *International Economics* 72 (2019), pp. 77–106.

[5] H. Harutyunyan et al. “Multitask learning and benchmarking with clinical time series data”. In: *Scientific Data* 6 (2019), p. 96. DOI: 10.1038/s41597-019-0103-9.

[6] S. Siami-Namini, N. Tavakoli, and A. Siami Namin. “A Comparison of ARIMA and LSTM in Forecasting Time Series”. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 1394–1401. DOI: 10.1109/ICMLA.2018.00227.

[7] J. Fattah et al. “Forecasting of Demand Using ARIMA Model”. In: *International Journal of Engineering Business Management* 10 (2018), p. 1847979018808670.

[8] C. Bui et al. “Time Series Forecasting for Healthcare Diagnosis and Prognostics with the Focus on Cardiovascular Diseases”. In: *Proceedings of the 5th International Conference on the Development of Biomedical Engineering in Vietnam*. 2018, pp. 809–818. DOI: 10.1007/978-981-10-4361-1_138.

[9] C. Purwanto, R. Logeswaran, and C. Eswaran. “A dual hybrid forecasting model for support of decision making in healthcare management”. In: *Advances in Engineering Software* 53 (2012), pp. 23–32. DOI: 10.1016/j.advengsoft.2012.07.006.

[10] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day, 1976.

[11] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. 5th. John Wiley Sons, 2015.

[12] R. J. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. 2nd. Melbourne: OTexts, 2018.

[13] S. J. Taylor and B. Letham. “Forecasting at Scale”. In: *The American Statistician* 72.1 (2018), pp. 37–45. DOI: 10.1080/00031305.2017.1380080.

[14] T. Chen and C. Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[15] S. Hochreiter and J. Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[16] C. J. Willmott and K. Matsuura. “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance”. In: *Climate Research* 30 (2005), pp. 79–82.

[17] D. A. Swanson, J. Tayman, and T. M. Bryan. “MAPE-R: A rescaled measure of accuracy for cross-sectional subnational population forecasts”. In: *Journal of Population Research* 28 (2011), pp. 225–243. DOI: 10.1007/s12546-011-9054-5.

[18] E.-J. Wagenmakers and S. Farrell. “AIC model selection using Akaike weights”. In: *Psychonomic Bulletin & Review* 11.1 (2004), pp. 192–196. DOI: 10.3758/BF03206482.

[19] J. W. Galbraith and V. Zinde-Walsh. “On the distributions of Augmented Dickey-Fuller statistics in processes with moving average components”. In: *Journal of Econometrics* 93 (1999), pp. 25–47. DOI: 10.1016/S0304-4076(98)00097-9.

[20] Y. Shin and P. Schmidt. “The KPSS stationarity test as a unit root test”. In: *Economics Letters* 38 (1992), pp. 387–392. DOI: 10.1016/0165-1765(92)90023-R.

[21] T. Wang et al. “Prevalence of hemorrhagic fever with renal syndrome in Yiyuan County, China, 2005-2014”. In: *BMC Infectious Diseases* 16 (2016), p. 69. DOI: 10.1186/s12879-016-1404-7.

[22] A. K. Sahai et al. “ARIMA modelling forecasting of COVID-19 in top five affected countries”. In: *Diabetes Metab. Syndr. Clin. Res. Rev.* 14 (2020), pp. 1419–1427. DOI: 10.1016/J.DSX.2020.07.042.

[23] S. J. Taylor and B. Letham. *Prophet: Forecasting at Scale*. Facebook Research Blog. 2017. URL: https://research.facebook.com/blog/2017/2/prophet-forecasting-at-scale/.

[24] S. et al. Prajapati. “Comparison of Traditional and Hybrid Time Series Models for Forecasting COVID-19 Cases”. In: *arXiv* 2105.03266 (2021). DOI: 10.21203/rs.3.rs-493195/v1.

[25] M. S. Rahman, A. H. Chowdhury, and M. Amrin. “Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh”. In: *PLOS Glob. Public Health* 2.5 (2022), e0000495. DOI: 10.1371/journal.pgph.0000495.

[26] C. X. Lv et al. “Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model”. In: *BMC Infectious Diseases* 21 (2021), pp. 1–13. DOI: 10.1186/s12879-020-05706-z.

[27] P. D. et al. Talagala. “An ensemble approach to forecasting using time series decomposition”. In: *International Journal of Forecasting* 38 (2022), pp. 1222–1241. DOI: 10.1016/j.ijforecast.2021.10.004.

[28] J. Zhang et al. “Forecasting with LSTM Neural Networks for Time Series: A Case Study on Financial Markets”. In: *Computational Intelligence and Neuroscience* (2017). DOI: 10.1155/2017/9136070.

[29] T. Chai and R. R. Draxler. “Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature”. In: *Geoscientific Model Development* 7 (2014), pp. 1247–1250.

[30] M. Motamedi et al. “Demand Forecasting for Platelet Usage: from Univariate Time Series to Multivariate Models”. In: *PLOS ONE* 18.2 (2023), e0297391. DOI: 10.48550/arXiv.2101.02305.

[31] S. Kim and H. Kim. “A new metric of absolute percentage error for intermittent demand forecasts”. In: *International Journal of Forecasting* 32.3 (2016), pp. 669–679. DOI: 10.1016/j.ijforecast.2015.12.003.

[32] T. O. Hodson, T. M. Over, and S. S. Foks. “Mean Squared Error, Deconstructed”. In: *Journal of Advances in Modeling Earth Systems* 13.11 (2021).

[33] C. Bergmeir, R. J. Hyndman, and B. Koo. “A note on the validity of cross-validation for evaluating autoregressive time series prediction”. In: *Computational Statistics Data Analysis* 120 (2018), pp. 70–83. DOI: 10.1016/j.csda.2017.11.003.

[34] Z. Chen et al. “Automated Agriculture Commodity Price Prediction System with Machine Learning Techniques”. In: *Advances in Science, Technology and Engineering Systems Journal* (2023). DOI: 10.48550/arXiv.2106.12747.

[35] S. Jain et al. “A Novel Ensemble ARIMA-LSTM Approach for Evaluating COVID-19 Cases and Future Outbreak Preparedness”. In: *Health Care Science* 3.6 (2024), pp. 409–425. DOI: 10.1002/hcs2.123.