VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

FACULTY OF fundamental sciences

DEPARTMENT OF information systems

Mirmatlab Karimli

**Research on Application of machine learning methods for optimization of business processes**

Master’s degree Thesis

Information Systems Software Engineering study programme, state code 6211BX017

Not specified specialisation

Informatics Engineering study field

**Supervisor** Prof. Dr. Diana Kalibatiene

Vilnius, 2025

Contents

[FACULTY OF fundamental sciences 1](#_Toc201749774)

[1. Introduction 6](#_Toc201749775)

[1.1. Investigation Object 6](#_Toc201749776)

[1.2. The Aim and Tasks of the Thesis 7](#_Toc201749777)

[1.3. Novelty of the Topic 7](#_Toc201749778)

[1.4. Relevance of the Topic 7](#_Toc201749779)

[1.5. Research Methodology 8](#_Toc201749780)

[1.6. Scientific Value of the Thesis 8](#_Toc201749781)

[1.7. Main Results of the Thesis 9](#_Toc201749782)

[1.8. Structure of the Work 9](#_Toc201749783)

[2. Related Works Analysis 10](#_Toc201749784)

[2.1. Main Concepts 10](#_Toc201749785)

[2.2. Review of the Main Related Works 10](#_Toc201749786)

[2.3. Systematization of the Related Works 13](#_Toc201749787)

[2.4. Main Results of the 2nd Section 22](#_Toc201749788)

[3. Proposed Approach 24](#_Toc201749789)

[3.1. Introduction to the Proposed Method 24](#_Toc201749790)

[3.2. BPMN Diagram: Drug Sale Prediction using XGBoost (Seasonal Focus) 24](#_Toc201749791)

[3.3. BPMN Diagram: Drug Sale Prediction Process in Pharmacy Stores 30](#_Toc201749792)

[3.4. Advantages of the Proposed Method 34](#_Toc201749793)

[3.4.1. Comparison of Models 34](#_Toc201749794)

[3.5. Main Results of the 3rd Section 35](#_Toc201749795)

[3.6. Future Work 36](#_Toc201749796)

[4. Initial Analysis of Drug Sale Dataset in Pharmacy Store 37](#_Toc201749797)

[4.1. General information about datasets 37](#_Toc201749798)

[4.1.1. 1st dataset (Turkish Pharmacy Store’s Data): 37](#_Toc201749799)

[4.1.2. 2nd dataset (Polish and German Pharmacy Stores’ Data): 38](#_Toc201749800)

[4.2. Dataset Analysis 39](#_Toc201749801)

[4.3. The Spearman rank-order correlation coefficient 40](#_Toc201749802)

[4.3.1. Description of the Spearman Rank-Order Correlation Matrix 42](#_Toc201749803)

[4.4. Seasonal Trends of Drug Sales 43](#_Toc201749804)

[4.5. Main Results of the 4th Section 45](#_Toc201749805)

[5. Conclusion 48](#_Toc201749806)

[References 49](#_Toc201749807)

**List of Tables**

[Table 2.1. Summary of Application of machine learning methods for drug sale predictions 14](#_Toc201749760)

[Table 3.1. The differences in main aspects between algorithms 34](#_Toc201749761)

[Table 4.1. Table of Dataset Comparison 39](#_Toc201749762)

List of Figures

[Figure 1. BPMN Table of Drug Sale Prediction using XGBoost 25](#_Toc201702976)

[Figure 2. Sub-process: Preprocess Data 28](#_Toc201702977)

[Figure 3. Sub-process: Feature Engineering 29](#_Toc201702978)

[Figure 4. BPMN Table of Business Process 30](#_Toc201702979)

[Figure 5. Spearman rank-order correlation matrix 41](#_Toc201702980)

[Figure 6. Seasonal trends more sold drugs 42](#_Toc201702981)

**Abbreviations**

**ML** – Machine Learning

**ARIMA** – Autoregressive Integrated Moving Average

**XGBoost** – Extreme Gradient Boosting

**LSTM** – Long Short-Term Memory

**MAPE** – Mean Absolute Percentage Error

**RMSE** – Root Mean Squared Error

**MAE** – Mean Absolute Error

**R2** – R-squared

**SVR** – Support Vector Regression

**CNN-LSTM** – Convolutional Neural Network - Long Short-Term Memory

**RF** – Random Forest

**SARIMA** – Seasonal Autoregressive Integrated Moving Average

**RNN** – Recurrent Neural Network

**RFID** – Radio Frequency Identification

**DRL** – Deep Reinforcement Learning

# Introduction

The pharmaceutical industry requires accurate sales forecasting and efficient inventory management to ensure that they maintain optimum levels of stock and satisfy customer demand. Common forecasting models, such as ARIMA, linear regression, etc. do not always capture the non-linear, dynamically changing behaviour of pharmaceutical sales, particularly in the case of seasonal fluctuations in demands and external influences. Thus, the implementation of advanced machine learning (ML) algorithms is a potential way to improve the correctness of predictions and facilitate business operations in pharmacy management (Dutta, Das, & Chatterjee, 2022; Manna, Kolpe, & Mhalungekar, 2023; Pall, Gauthier, Auer, & Mowaswes, 2022).

In this thesis, I would like to look into the use of machine learning to help optimize drug sales forecasting in pharmacy stores. In enhancing the accuracy of the forecasts, the research aims to enable the pharmacies to optimize their levels of stock, minimize operational expenses and minimize customer waiting time. The research is concentrated on dealing with complicated time-series data, which are usually affected by unstable sales and outside influences, including promotions and seasonal changes (Saena & Suttichaya, 2020; Zeng, Yang, Wang, Zhu, & Feng, 2025).

The novelty of the proposed research is the use of advanced ML techniques to solve the problem of pharmaceutical sales forecasting features. The past research applying ML techniques does not fully optimize them to the problem of pharmacies, i.e. non-linear and seasonal sales trends. The proposed research will address this gap, as it will concentrate on such methods as XGBoost and LSTM that can learn such complicated patterns (Fourkiotis & Tsadiras, 2024; Gurnani, Korkey, Shahz, Udmalex, Sambhe, & Bhirudk, 2017).

The problem is relevant because appropriate sales forecasting may result in a decrease in costs and in more competent use of inventory (Zeng et al., 2025). The aim is proving that machine learning is able to increase the accuracy of forecasting, which will minimize waste and stocks as well as increase customer satisfaction in pharmacies. The results might as well be implemented in other retailing sectors having the same problem (Mousa & Al-Khateeb, 2023).

The study methodology is about the comparison of the effectiveness of different machine learning models to the traditional forecasting methods on pharmaceutical sales data. The evaluation will be carried out with the help of such metrics as Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE) (Pall et al., 2022; Saena & Suttichaya, 2020). Also, the hybrid machine learning model will be suggested to achieve optimal sales forecasting to perform better inventory management and optimize the business processes in the pharmaceutical industry (Manna et al., 2023; Gurnani et al., 2017).

## Investigation Object

The investigation object is machine learning techniques and their use in drug sales prediction.

## The Aim and Tasks of the Thesis

The aim of this research is to enhance the accuracy of drug sales predictions in pharmacy stores by using machine learning techniques.

1. To analyze current machine learning frameworks and techniques for drug sales forecasting.
2. To propose a machine learning-based method for optimizing drug sales predictions in pharmacy stores.
3. To develop and evaluate a prototype of the suggested method and confirm its efficiency.

## Novelty of the Topic

Using machine learning methods to predict drug sales in pharmacy stores is becoming a big issue for the pharmaceutical industry. Models such as ARIMA normally used for forecasting sales data usually miss the unique, complex movements and non-linear patterns seen in medicine sales. Besides, the standard approaches mostly deal with little information or simple models, which makes their predictions less secure. This work will help create a new understanding of drug sales by introducing and using machine learning techniques that look at sales during seasonalities. For example, Saena et al. (2025) carried out experiments with Deep Learning (DL) models such as LSTM and CNN-LSTM to forecast results but did not look closely at how those methods could help optimize sales of drugs in pharmacy stores (Saena et al., 2025). In a similar manner, Mousa and Al-Khateeb (2023) looked into machine learning methods for predicting drug demand. However, they did not discuss how to implement ML forecasts efficiently in pharmacy stores. Nevertheless, Pall et al. (2023) took on predicting shortages of drugs but did not examine how to use machine learning to forecast daily sales in pharmacies (Pall et al., 2023).

In contrast to previous research, this research has a major difference by including XGBoost, LSTM, and hybrid models in machine learning, helping to predict better and oversee the pharmacy store's inventory.

## Relevance of the Topic

Managing drug sales predictions in pharmacy stores is very important for the pharmaceutical industry since it is more difficult today to predict sales, maintain the life of products, and handle stock. If sales predictions are correct, pharmacy stores can keep the right amount of stock, use less of what they do not sell, and never lack drugs for patients. With more pharmaceutical products being introduced and some medications affected by the seasons, usual forecasting methods do not always identify the details, which leads to problems in drug sales.

Various factors show why this topic is still important. One of the main challenges is that the industry may lose money if its sales predictions are not correct. As explained by Saena and Suttichaya (2025), in their study, machine learning is being investigated to forecast drug sales, but it is also acknowledged that most other models are not able to handle the difficulties of pharmaceutical sales data. In addition, Mousa and Al-Khateeb (2023) mention that since deep learning methods hold much potential, researchers should explore their use in the pharmaceutical industry. Finally, Pall et al. (2023) mention that sales forecasting is relevant for coping with drug shortages since predicting sales can control inventory issues that are key to the healthcare system.

They show that enhancing drug sales prediction models helps in cutting down expenses and enhances that the pharmacy stores can meet public health needs in a fast and effective way. The aim of this thesis is to fill a gap by applying more advanced machine learning models for predicting drug sales of pharmacy stores, which helps both the financial and healthcare parts of the pharmaceutical industry.

## Research Methodology

This research follows a systematic approach to optimize drug sales prediction in pharmacy stores by using machine learning (ML) techniques. Existing machine learning techniques were assessed through comparative analysis and library research, and holes were found and ideas were generalized with the aid of systematization and logical induction. The study was done using the case method to determine the benefits of the new framework in predicting sales and controlling needs. The model integrates ML algorithms for sales forecasting, stock level optimization, and replenishment planning. Logical induction and data-driven evaluation guide the refinement of this approach.

## Scientific Value of the Thesis

This study's primary scientific contribution is the creation and assessment of an optimization technique based on machine learning that is specifically suited for pharmaceutical sales prediction. In forecasting pharmaceutical sales, the study offers a comparative analysis of current machine learning frameworks, highlighting their advantages and disadvantages. This model offers a creative, data-driven way to increase operational efficiency in pharmaceutical sales forecasting by combining ML-driven demand forecasting, stock replenishment, and expiration minimization techniques. The research advances our understanding of ML-driven business process optimization by showing how ML approaches can improve sales predictions, lower inventory costs, and minimize shortages.

## Main Results of the Thesis

The goal of the thesis is to set up a machine learning approach to improve predicting drug sales in pharmacy stores. It was observed during the analysis that models like ARIMA have difficulties dealing with non-linear patterns that change over time and fluctuate seasonally, which are common in pharmaceutical data. The models, XGBoost and LSTM were better suited for drug sales since they gave the most reliable results with datasets that had changing trends over time. According to the analysis, a method that merges XGBoost and LSTM was created to address both complicated trends and seasonal effects.

The study creates a machine learning-based optimization technique especially suited for pharmaceutical sales prediction based on the comparison analysis. By cutting-edge demand forecast models, the strategy enhances drug sales, minimizes shortages and ensures optimal stock levels. Through case study analysis, the suggested method's practical evaluation demonstrates its efficacy in lowering operating expenses and improving sales predictions. These findings support the idea that machine learning (ML) can enhance pharmaceutical sales prediction for business process optimization.

## Structure of the Work

The second section deals with the examination of current machine learning methods for pharmaceutical sales prediction. The third section examines the creation of the suggested machine learning-based framework for pharmaceutical sales prediction. The fourth chapter describes the framework's implementation and assessment using testing and prototyping on both artificial and real-world datasets.

# 

# Related Works Analysis

## Main Concepts

Sales forecasting is the main idea of this thesis, which predicts future drug sales to optimize inventory management in pharmacy stores. The correct forecasting supports keeping the appropriate quantity of stock besides preventing excess and stock-outs. The pharmaceutical industry is a business where forecasts are hard because of the seasonality and the perishability of the products. According to Saena and Suttichaya (2025), traditional approaches, such as ARIMA, may not perform well in such complexities because of their incapability to model non-linear patterns and seasonality of drug sales.

Using machine learning algorithms is also the topic of this thesis, especially XGBoost and LSTM. XGBoost is stronger in large datasets, and non-linear relations, whereas LSTM is suitable in time-series predictions, which aim at representing long-term dependencies in sales data. Pall et al. (2023) argue that these machine learning models are superior to the traditional approaches since they offer more reliable predictions and can process complex., large volumes of information more effectively.

Besides, the study examines hybrid models, where XGBoost and LSTM are integrated to take advantage of the benefits of the 2 methods. Such a hybrid method provides a higher accuracy in forecasting since it can capture both time-series dynamics and non-linear dynamics. The studies by Saena and Suttichaya (2025) and Pall et al. (2023) demonstrate that hybrid models can outperform single algorithms, being able to address several issues in medicine sales prediction.

## Review of the Main Related Works

The related works review indicates that different machine learning techniques are applied to predict drug sales. Kravets et al. (2018) used random forest models and included such factors as availability, seasonality, and price for sales prediction. According to Saena and Suttichaya (2019), a CNN-LSTM model demonstrated better results than other methods of sales forecasting. Also, Dutta et al. (2019) discussed machine learning use in other fields, including disease diagnosis that can be transformed to apply to predicting pharmaceutical sales. Such works outline the usefulness of machine learning, particularly random forest and LSTM, in increasing accuracy in sales forecasting.

There have been several machine learning algorithms used in the prediction of pharmaceutical drug sales in the literature survey. As another example, Mehmet Ali Balcı and Ömer Akgüller (2021) reported approximately 90% accuracy in predicting drug sales using the Support Vector Regression (SVR). Some other researchers like Hunneman et al. (2021) utilized Generative Adversarial Networks (GAN) to make predictions as well as rely on historical data to make sales predictions. Also, XGBoost has been popular and its application in predicting sales was successful in studies like that conducted by Liu et al. (2021), as it outperformed the conventional models of ARIMA and LSTM. These papers demonstrate how there is an increasing application of more sophisticated machine learning algorithms in order to maximize the accuracy of the predictions on drug sales.

The related works reviewed in this article focus on the use of the machine learning method to predict drug sales. As an example, Janardhanan and Barrett (2017) applied LSTM and ARIMA to time series forecasting of CPU usage demonstrating that LSTM can be successfully applied to make accurate predictions. Also, Selvin et al. (2017) discussed stock price prediction based on LSTM, RNN, and CNN with a sliding window technique that outperformed the ARIMA, emphasizing the effectiveness of the deep learning models. Likewise, Kaneko and Yada (2016) used deep learning to forecast sales in retail stores, which also demonstrated the promising nature of such models in enhancing the accuracy of different forecasting tasks. All these studies show the rising use of advanced machine learning algorithms, mainly LSTM and CNN based models, to make accurate predictions in sales forecasting.

When analyzing the data on drug sales, various machine learning algorithms are used to predict drug sales, in particular, linear regression, random forest, neural networks, and support vector machines. Some past research, such as Zadeh et al. (2014) has been conducted on how effective machine learning is in sales prediction, specifically through a data mining method such as support vector regression. Besides that, the Levenberg-Marquardt algorithm is employed to make more precise predictions in time series, as shown by Al-Gunaid et al. (2017), who used neural network-based models in energy consumption forecasting. The research then comes to the conclusion that random forest becomes the best strategy to apply in pharmaceutical sales forecasting as it shows a high level of accuracy in comparison with other models that were established.

This article, related works review is devoted to using machine learning (ML) and deep learning methods to forecast pharmaceutical demand. Different ML models such as Support Vector Regression (SVR), Random Forest (RF), and Recurrent Neural Networks (RNN) can be used to predict drug sales, as it has been done by Kravets et al. (2018), who accomplished sales prediction in the territory of the Volgograd region with RF models. Besides, more advanced deep learning models such as Long Short-Term Memory (LSTM) networks are mentioned due to their capability to learn non-linear relationships in time-series data, which Amalnick et al. (2020) demonstrate. The models play a significant role in enhancing the accuracy of the predictions especially when it comes to the management of the pharmaceutical supply chain and optimization of inventory levels with a view to minimizing costs and ensuring that shortages are avoided.

Within the scope of the related works review, the article explains the great value of machine learning techniques in predicting pharmaceutical sales with a focus on diverse techniques, such as XGBoost and LSTM networks. As an illustration, Fourkiotis and Tsadiras (2024) emphasize that XGBoost demonstrates the best results in terms of pharmaceutical sales prediction and shows the lowest values of mean absolute percentage error (MAPE) in different classes of drugs, including 17.89% anti-inflammatory drugs (M01AB) and 16.92% anti-inflammatory drugs (M01AE). The research uses these models in comparison to the traditional models such as the ARIMA which turned out to be less efficient in modelling the seasonal effects and complicated trends within the data. In addition, the study also highlights the essentiality of data preparation, seasonality adjustment, and application of superior algorithms, such as LSTM, in dealing with long-term dependencies in pharmaceutical sales prediction.

Different machine learning models such as ARIMA, Auto Regressive Neural Network (ARNN), XGBoost, Support Vector Machine (SVM) as well as hybrid models were compared in terms of sales forecasting in this article. ARIMA was applied first but missed to capture nonlinear trends, so they applied more sophisticated models such as ARNN that worked well with nonlinearity (Gurnani et al., 2017). The hybrid models which include ARIMA with ARNN, XGBoost, and SVM performed better, especially the ARIMA-ARNN hybrid. More so, the Standard Template Library (STL) decomposition, which separates time series into seasonal, trend, and remainder showed the best performance in all other models, including hybrids, with the least Root Mean Squared Error (RMSE). The findings indicate the benefit of employing the STL decomposition to process both linear and nonlinear sales data components (Gurnani et al., 2017).

In this article, there are machine learning and ensemble learning are used in applying to drugstore sales forecasting, and impressive progress is made. Zeng et al. (2025) propose a new TS-LGBM model that uses a self-attention mechanism and integrates LightGBM to model temporal dependencies in sales data and outperforms the former methods in terms of prediction accuracy, including XGBoost and LightGBM. Such other models as LSTM are also compared, and TS-LGBM demonstrates the best result in both root mean square error (RMSE) and symmetric mean absolute percentage error (SMAPE) measures. This creates the significance of the increased need to incorporate deep learning methods to enhance sales forecasting in the pharmaceutical sector.

This article uses machine learning to forecast drug shortage, which is a major problem in the drug business. Pall et al. (2023) applied historical data on drug shortages in Canadian pharmacies and shortage reports to train predictive models that can predict drug shortages with an accuracy of 69%. The models have taken into account several variables that include the days of supply per patient, the patient number, past shortage, and therapeutic classes of drugs. By doing this they not only assist in predicting the shortages but also allow the pharmacies to manage their stock and order models in a much better way and eventually minimize the effects of shortages on the patients and the work processes.

In this paper, the authors suggest a machine learning model of sales forecasting with XGBoost. Features engineering techniques that the model uses include time features, statistical features, and lag features, which are extracted through the sales data of Walmart. Experiment results on this dataset, which contains 1913 days of data on Walmart stores in the U.S., show that the XGBoost model provides better results than the classical regression models such as Linear and Ridge Regression. In particular, the XGBoost model has a Root Mean Squared Scaled Error (RMSSE) score of 0.655 which is much lower than 0.783 (Linear Regression) and 0.774 (Ridge Regression) (Xie & Zhang, 2021). It means that XGBoost is not only more efficient but also more accurate in terms of sales prediction.

## Systematization of the Related Works

This table consists of nine columns that summarize various research studies on dynamic resource allocation in business processes. The columns are as follows: References (R); Year of Publication (Year); Main Research Question/Problem (MRQ); Approach (A); Field Studied/Application Domain (F); Dataset Used (D); Attributes Used for Prediction (AUP); Evaluation of the Approach (EA); Comparison with Other Works (COW); and Results (R). Each column presents key aspects of the research, offering insights into how various studies address business process optimization challenges in drug sale prediction in pharmacy stores. It includes machine learning techniques, sales forecasting methods, pharmacy store sales datasets, and evaluation strategies, demonstrating their impact on optimizing drug sale predictions and improving business process efficiency in pharmacy management.

Table 2.. Summary of Application of machine learning methods for drug sale predictions

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Reference** | **Main research question / problem** | **Used approach** | | **Field Studied / Application domain** | **Dataset used** | **Attributes used for prediction** | **Evaluation of the approach** | **Comparison with other works** | **Result** |
| (Dutta et al. 2022) | To predict the sales of pharmaceutical products using machine learning methods. | | Linear Regression, ARIMA, Seasonal ARIMA, Holt-Winters, Seasonal Naïve Model. | Pharmaceutical Sales. | Pharmaceutical product dataset from GitHub. | Historical sales data, product price, seasonal data. | MAPE of 19.07%, compared with other models (Holt-Winters, Seasonal ARIMA). | Linear Regression outperformed Holt-Winters and other models. | Linear Regression performed best with MAPE, outperforming other models like Holt-Winters |
| (Manna et al., 2023) | To predict pharmaceutical drug sales using machine learning algorithms. | | Linear Regression, Random Forest (RF), SVR, XGBoost. | Pharmaceutical Drug Sales. | Sales data of various drugs (antipyretics, antihistamines, etc.). | Hourly, weekly, monthly, and yearly sales data. | XGBoost achieved the highest accuracy for M01AB and N02BA. | XGBoost outperformed Linear Regression, SVR, and RF in sales prediction. | XGBoost provided the highest accuracy compared to other models with accuracy. |
| (Saena & Suttichaya, 2023) | To predict medication purchase amounts for inventory management using ML. | | MLP, LSTM, CNN-LSTM, Rolling Windows. | Pharmaceutical sales forecasting. | Sales data from AV group medicines. | Monthly, 3-month, and 6-month sales data. | CNN-LSTM provided the best forecasting results | CNN-LSTM model outperformed other models like MLP and LSTM | CNN-LSTM showed superior performance for 1-month predictions for specific medicines. |
| (Al-Gunaid et al., 2018) | To compare ML methods for predicting pharmaceutical drug sales. | | Linear regression, Random Forest, Neural Network, SVM, Levenberg-Marquardt Algorithm. | Pharmaceutical drug sales prediction. | Sales data of 2 medications. | Various sales data such as pricing, sales per week, exchange rates. | Random Forest and Neural Network models gave promising results. | Comparison of models showed that Random Forest had the highest accuracy. | RF performed well but it still requires improvement to predict more accurately. |
| (Mousa & AL-Khateeb 2023) | Using deep learning to forecast pharmaceutical sales demand. | | Review of predicting methods using ML and DL. | Forecasting pharmaceutical sales and demand. | Numerous databases, such as Coop service, Rossman store sales, and Pharmal sales, among others. | Prescription data, economic factors, and past sales. | Measures of accuracy include MSE, RMSE, and R2. | Compared RF, LSTM, ARIMA, XBoost, and linear regression. | Forecasting accuracy was increased by DL techniques, however data limits are still an issue. |
| **Reference** | **Main research question / problem** | | **Used approach** | **Field Studied / Application domain** | **Dataset used** | **Attributes used for prediction** | **Evaluation of the approach** | **Comparison with other works** | **Result** |
| (Fourkiotis & Tsadiras, 2024) | To forecast pharmaceutical sales using machine learning methods. | | ARIMA, LSTM, XGBoost, Seasonal Naïve. | Pharmaceutical sales forecasting.. | 600,000 sales records from a pharmacy. | Historical sales data, product categories, seasonality. | XGBoost achieved the lowest MAPE values for various drug categories. | Compared with ARIMA and LSTM, XGBoost outperformed in all categories. | XGBoost provided the lowest MAPE and MSE, outperforming ARIMA, LSTM. |
| (Gurnani et al., 2017) | To evaluate hybrid machine learning models for sales forecasting. | | ARIMA, ARNN, XGBoost, SVM, STL Decomposition. | Drugstore sales forecasting. | Rossmann sales data. | Weekly sales data, promotional events, holidays. | STL Decomposition provided the best results. | Hybrid models performed better than individual models like ARIMA. | STL Decomposition outperformed all models with the lowest RMSE. |
| (Zeng et al., 2025) | To forecast sales for chain drugstores using a novel model. | | TS-LGBM (Self-attention + LightGBM), XGBoost, LSTM. | Chain drugstore sales forecasting. | Z-chain drugstore data. | Temporal characteristics, nearby drugstores, sales data. | TS-LGBM outperformed all other models. | TS-LGBM outperformed TS-XGB, LightGBM, XGBoost, LSTM. | TS-LGBM demonstrated superior prediction accuracy. |
| (Pall et al. 2023) | Using machine learning and pharmacy data to forecast medication shortages. | | XBoost-based supervised machine learning for classification. | Management of the pharmacy supply chain. | 22 Canadian pharmacies' sales statistics as well as historical data on drug shortages. | Days of supply on average for each patient, days of supply overall, past shortages, and drug hierarchy within therapeutic classes. | 69% precision and 0.44 kappa value for shortage class prediction. | Contrasted with other performance evaluation models | 59% of the most significant shortages were predicted one month in advance. |
| (Xie & Zhang, 2021) | To develop a machine learning model for sales forecasting using XGBoost. | | XGBoost, Feature Engineering. | Retail Sales Forecasting. | Walmart sales dataset (1913 days of data). | Historical sales data, time features (day, week, month), price, promotional events. | The XGBoost model performed well, achieving an RMSSE score. | Compared with Linear Regression and Ridge Regression, XGBoost was more accurate. | XGBoost outperformed Linear Regression and Ridge Regression with an RMSSE. |
| (Saha & Rathore 2024) | Creating an intelligent inventory control system for hospital supply chains that depend on drug demand. | | Reinforcement learning using multiple agents and a semi-Markov decision process model. | Inventory control for the hospital supply chain. | Information from an Indian multispecialty hospital. | Demand levels for medications, drug dependence, and inventory control procedures. | A case study was used to validate the model; order quantities and inventory policies were optimized. | In contrast to conventional can-order and min-max policies. | Better delivery of healthcare services and lower costs through efficient inventory control. |
| **Reference** | **Main research question / problem** | | **Used approach** | **Field Studied / Application domain** | **Dataset used** | **Attributes used for prediction** | **Evaluation of the approach** | **Comparison with other works** | **Result** |
| (Boute et al. 2022) | Examining how deep reinforcement learning (DRL) might be used in inventory control. | | DRL includes multiple algorithmic options: actor-critic, value and policy based models. | Job-shop scheduling and Industry 4.0 scheduling. | Supply chain optimization and inventory management. | Lead times, ordering decisions, demand trends, and inventory levels. | Evaluated with a range of training techniques and hyperparameter values. | DRL strategies including DON, PPO, and A3C were contrasted with conventional inventory control techniques. | DRL can perform better than conventional techniques. Model selection and tuning are essential. |
| (Angula & Dongo 2024) | Al and ML's effects on public pharmaceutical systems' ability to predict drug supply and demand. | | Thorough analysis of current Al/ML forecasting techniques. | Systems of public pharmaceuticals. | Thirteen peer-reviewed studies from Google Scholar and PubMed. | Data on demand and supply, prescription medication records, and consumption. | R2 and RMSE were used to quantify accuracy; Random Forest had the lowest RMSE of 0.74. | Comparison of several AI/ML models, such as ARIMA, ANN, and linear regression. | Demand recasting is improved by Al/ML, but adoption, privacy, and ethics issues still exist. |
| (Bhat et al. 2024) | Medication access optimization with machine learning-based stochastic inventory models. | | Using projective, causal, and machine learning forecasting techniques for stochastic modeling. | Inventory management in public health. | Information about a single medication from 53 medical facilities. | Lead time, inventory levels, patient visits, and patterns of consumption. | Evaluation using RMSE; comparison of projective, causal, and stochastic models. | Compared the stochastic, LSTM, and ARIMA models. | ML models increased forecasting accuracy, decreased shortages, and optimized stock levels. |
| (Ahmadi et al. 2022) | Creating clever inventory control strategies for pharmaceuticals that expire quickly in a healthcare supply chain. | | Stochastic mixed integer programming and reinforcement learning, including Q-learning and Deep Q-network. | Management of healthcare inventories. | Data simulations for a central storehouse and regional hospitals. | Purchase price based on age, product age distribution, order amounts, and remaining life distributions. | Compared with policies utilizing genetic algorithms and CPLEX. | Demonstrated cost-effectiveness in comparison to conventional periodic review procedures. | IIM policies decreased inventory costs, enhanced service quality, and decreased product shortages. |
| (Mostofi & Jain 2021) | Utilizing Industry 4.0 to manage inventories and regulate degrading pharmaceutical products. | | Content analysis of pharmaceutical logistics apps using Industry 4.0. | Supply chain for pharmaceuticals. | Analysis based on literature. | Logistics effectiveness, deterioration rates, and inventory levels. | Evaluation of Industry 4.0 tools qualitatively. | Contrasting Industry 4.0 applications with conventional inventory management. | Industry 4.0 technologies increase inventory tracking, decrease waste, and improve supply chain. |
| **Reference** | **Main research question / problem** | | **Used approach** | **Field Studied / Application domain** | **Dataset used** | **Attributes used for prediction** | **Evaluation of the approach** | **Comparison with other works** | **Result** |
| (Kalaichelvan et al. 2024) | Using machine learning and fuzzy theory to optimize EOQ in pharmacy inventory management. | | Kuhn-Tucker optimization, fuzzy logic, and the naïve Bayes classifier. | Optimization of the pharmaceutical supply chain. | Dataset for a simulated pharmaceutical inventory. | Variations in demand, product deterioration, and manufacturing and transportation costs. | High classification accuracy with Naïve Bayes (95.9%). | Fuzzy models and conventional Economic Order Quantity (EOQ) techniques were compared. | Lower inventory expenses and more accurate decision-making. |
| (Rekabi et al. 2023) | Creating supply chains for pharmaceuticals that include perishable goods and soft time windows. | | Demand forecasting and target accomplishment using linear and quadratic regression. | Design of a pharmaceutical supply chain network. | Tehran's historical pharmaceutical demand data. | Delivery windows, variations in demand, and supply chain expenses. | When it came to demand forecasting, quadratic regression fared better than linear regression. | Compared several methods for demand forecasting and optimization. | Enhanced customer satisfaction and lower supply chain expenses. |
| (Rojas et al. 2024) | Bibliometric evaluation of studies on pharmaceutical supply chains' stochastic demand inventory models. | | Bibliometric analysis and a systematic review. | Inventory control in the pharmaceutical supply chain. | 39 studies published between 2018 and 2024. | Collaboration networks, citation counts, and publication trends. | Leading authors, important trends, and research gaps were identified. | Compared various stochastic demand inventory models in the literature. | Emphasized the value of international and interdisciplinary cooperation. |
| (Fernández del Rio et al. 2024) | Improving pharmacy services through behavioral analysis based on reinforcement learning. | | Learning reinforcement and contextual bandits. | Digital pharmacy services and decision-making based on Al. | The SwipeRx app's behavioral data (235,000 users, 45,000 pharmacies). | Stock replenishment, pharmacist involvement, and purchase patterns. | Adaptive treatments raised the variety of products and pharmacy spending. | Сontrasted standard digital treatments with reinforcement learning. | Уnhanced efficiency in pharmacist decision-making and inventory management. |
| (Luu el al. 2023) | Using ML to forecast how many pharmaceutical medications manufacturers will order. | | ML models: XGBoost, Histogram-Based Gradient Boosting and Random Forest. | Forecasting the pharmaceutical supply chain. | Data on 342 drug transactions over 41 months (Vietnam). | Drug consumption quantity, time of consumption, location of consumption. | Random Forest RMSE of 0.95 and R2 of 0.81. | ML models (XGBoost, LGBM, and RF) were compared. | In forecasting medication demand, Random Forest fared better than any other model. |
| (Almentero et al. 2024) | Using machine learning and time series techniques to forecast pharmacy purchase orders. | | Facebook Prophet, Random Forest, XBoost, SARIMA, and LSTM. | Predicting pharmaceutical inventory and purchases. | Data from a pharmaceutical wholesaler in Canada, spanning 15 years. | SKU, date of transaction, past sales, seasonality, and outside variables (holidays, weather). | SARIMA scored better on weekly forecasting than ML models. | contrasted ML techniques with conventional statistical methods. | For pharmacy purchase orders, SARIMA offered the most accurate forecast. |
| **Reference** | **Main research question / problem** | | **Used approach** | **Field Studied / Application domain** | **Dataset used** | **Attributes used for prediction** | **Evaluation of the approach** | **Comparison with other works** | **Result** |
| (Zhu et al. 2021) | Pharmaceutical demand forecasting with supply chain data and machine learning. | | RNN, RF, SVR, and VAR cross-time-series machine learning models. | Supply chain demand forecasting for pharmaceuticals. | Historical demand and inventory data from significant pharmaceutical manufacturers. | Quantity of orders, past demand, and levels of downstream inventory. | The best performance came from RNN, which greatly decreased forecasting mistakes. | contrasted ML with conventional pharmaceutical forecasting (moving average, ARIMA). | Compared to conventional techniques, RNN-based cross-series forecasting increased accuracy. |
| (Chamekh et al. 2017) | Supply chain management for pharmaceuticals using RFID and context-aware middleware. | | RFID, LoT, middleware that is aware of context, and Fasttrack. | Logistics and supply chain security for pharmaceuticals. | RFID data simulations from pharmaceutical production. | Location, time, identity, logistical factors, and drug tracking. | Evaluated with a Fosstrack middleware prototype. | contrasted RFID tracking with conventional logistics. | A context-aware RFID system enhanced logistics management and medicine traceability. |
| (Mehdiyev et al. 2018) | Creating a business process prediction model based on deep learning. | | Deep learning method with multiple stages. | Business process optimization and monitoring. | Helpdesk event logs, BPI Challenge 2012, and BPI Challenge 2013. | Logs of process execution, historical occurrences, and organizational roles. | Enhanced prediction accuracy through the use of n-gram encoding and feature hashing. | Comparing deep recurrent neural networks with traditional classification models. | Compared to current predictive process analytics techniques, deep learning performed noticeably better. |

The first column lists the authors and publication years of several research articles pertaining to business process optimization and machine learning (ML), which displays references in APA format. These references include studies like Dutta et al. (2022), who have applied such models as ARIMA to pharmaceutical sales forecasting; Manna et al. (2023) who have used XGBoost and Random Forest; Saena & Suttichaya (2023) who have employed CNN-LSTM to predict the purchase of medications and Al-Gunaid et al. (2018) who have compared the results of different machine learning models in sales prediction. Mousa & Al-Khateeb (2023) have sampled deep learning techniques like LSTM, Fourkiotis and Tsadiras (2024) and Gurnani et al. (2017) have looked at models like XGBoost and STL Decomposition. The TS-LGBM model suggested by Zeng et al. (2025) demonstrated the best results among the others and Pall et al. (2023) applied XGBoost to forecast pharmacy drug shortages. Furthermore, Angula & Dongo (2024) evaluated AI/ML forecasting methods for public pharmaceutical systems, and Boute et al. (2022) investigated the application of deep reinforcement learning (DRL) for inventory management. Other research focuses on business process optimization. Mehdiyev et al. (2018) investigated deep learning techniques for predictive business process monitoring.

The second column lists the primary research question or issue that each study attempts to answer with an emphasis on the use of machine learning (ML) and optimization approaches across a range of areas. For example, Dutta et al. (2022) forecasted the sale of pharmaceutical products with the help of ML, and Manna et al. (2023) applied such methods as Random Forest, SVR, and XGBoost to predict drug sales. The study is significant because it demonstrates how various approaches are adopted to make sales forecasts. Others, such as Saena & Suttichaya (2023) who have forecasted the purchase of medication based on inventory and Gurnani et al. (2017) have used hybrid machine learning models to test sales forecasting. Pall et al. (2023) sought to predict drug shortages using machine learning models in pharmacy supply chains, and Ahmadi et al. (2022) examined the creation of intelligent inventory control systems for pharmaceuticals with limited shelf life. In a similar vein, Saha and Rathore (2024) aimed to develop a drug-demand-based intelligent hospital inventory management system. Together, these studies use cutting-edge ML techniques to address important issues in drug sales forecasting, business process optimization and strategic decision-making.

The third column highlights an approach or methodology of the details as to which algorithms, models or techniques were applied to predict drug sales or business process optimisation. For example, Dutta et al. (2022) combined Linear Regression, ARIMA, and Seasonal ARIMA models to make sales predictions, whereas Manna et al. (2023) applied such machine learning models as XGBoost, Random Forest, and SVR to perform the same task. Researchers used deep learning networks, including LSTM and CNN-LSTM, in studies, such as Saena & Suttichaya (2023) and Gurnani et al. (2017), whereas Pall et al. (2023) utilized supervised learning through XGBoost to forecast drug outages. To improve hospital inventory control, Saha & Rathore (2024) used a semi-Markov decision process model with reinforcement learning with numerous agents. Boute et al. (2022) investigated many Deep Reinforcement Learning (DRL) techniques, such as actor-critic, value-based, and policy-based models, for inventory optimization in a larger supply chain environment. Together, these papers show how various machine learning methods have been utilized in several areas of drug sales prediction.

The fourth column illustrates the subject or application domain in which each study used machine learning (ML) or optimization techniques. For example, one can consider Dutta et al. (2022) and Manna et al. (2023): in both studies, the authors consider the sales sphere of pharmaceuticals, trying to forecast drug sales with the help of different machine learning models. Saena & Suttichaya (2023) focused on pharmaceutical sales prediction, especially the quantities of medications bought to manage the stock. Such works as Pall et al. (2023) and Gurnani et al. (2017) are related to the drugstore sales forecasting problem, where optimization of the sales process and forecasting of shortages in the work of pharmacies are relevant. Mehdiyev et al. (2018) employed deep learning for process monitoring. These studies show how machine learning can be used to optimize inventory and sales predictions as well as a better means of decision-making in the pharmaceutical and retail industries.

The fifth column lists the sources and kinds of data utilized for training and testing the machine learning models, which also lists the datasets used in each study. This information forms the basis for the formulation of sound forecasts and streamlining of strategies. For example, Fourkiotis & Tsadiras (2024) used 600,000 sales records of a pharmacy, and they were interested in the historical sales data, product category, and seasonality. Gurnani et al. (2017) applied the Rossmann sale data, particularly, the weekly sale data, promotional events, and holidays. Pall et al. (2023) implemented a prediction of drug shortages based on sales statistics and historical data of 22 Canadian pharmacies, whereas Saena & Suttichaya (2023) utilized sales data of AV group medicines. The data in Zeng et al. (2025) was based on Z-chain drugstore data, where the temporal features, the densities of drugstores in the vicinity, and the sales data were considered. Mousa & Al-Khateeb (2023) accessed various pharmaceutical sales databases, such as Coop service, Rossman store sales, and Pharmal sales and Mousa & Al-Khateeb (2023) predicted demands based on prescription data, economic factors, and past sales. The diversity and applicability of the datasets utilized in different papers to solve particular forecasting and optimization problems in the pharmaceutical and retail business are highlighted in this column.

The sixth column lists the key elements of machine learning (ML) models, which describes the qualities used for the factors that have contributed to the results of the forecasting. As an illustration, in Dutta et al. (2022), the features comprised past sales figures, product prices, and seasonal figures, which were useful in forecasting future selling trends of drugs. As predictive variables, Manna et al. (2023) utilized hourly, weekly, monthly, and yearly sales values and various types of drugs as predictive variables. In Pall et al. (2023), the features were days of supply per patient, the number of patients, past shortages, and drug hierarchy within therapeutic classes to forecast drug shortages. Saena & Suttichaya (2023) used monthly, 3-month, and 6-month sales data, whereas Gurnani et al. (2017) utilized weekly sales data, promotional events, and holidays as the main features of sales forecasting. Saha and Rathore (2024) examined inventory policies, drug dependence, and pharmaceutical demand. These characteristics specify important determinants in predicting sales, inventory management, and demand forecasting performance and accuracy of machine learning models.

The seventh column lists the evaluation of the method, which gauges the efficacy of machine learning (ML) models. Pall et al. (2023) obtained 69% precision for shortage forecasts, while Ahmadi et al. (2022) showed cost-effectiveness over periodic review approaches in pharmaceutical inventory management. In order to optimize inventory policies, Saha & Rathore (2024) used a hospital case study to evaluate their approach. Boute et al. (2022) examined deep reinforcement learning (DRL) models using several training methods in supply chain optimization, and Angula & Dongo (2024) discovered that Random Forest had the lowest RMSE (0.74). Mehdiyev et al. (2018) demonstrated that deep learning increased prediction accuracy in business applications, Rane et al. (2024) examined AI-driven business strategies, Chaudhary et al. (2023) evaluated the influence of machine learning on automation, and Van Dun et al. (2023) contrasted GAN-based process optimization with human-designed enhancements. These analyses determine the strength and the stability of the methods employed where it indicates how the performance of each individual model is quantitatively assessed and contrasted to the rest.

The eighth column lists the machine learning (ML) models from each study, which assist in putting the results of the research into perspective since it considers ways in which the suggested approach compares to similar methodologies. For example, in Pall et al. (2023), the research team compared the XGBoost-based model with other performance assessment models in drug shortage prediction and demonstrated that the XGBoost predicted 59% of the major shortages a month earlier. The XGBoost model presented in Manna et al. (2023) has outperformed the Linear Regression, SVR, and Random Forest models in terms of predicting the accuracy of the drug categories, especially M01AB and N02BA. Likewise, Fourkiotis & Tsadiras (2024) compared XGBoost to ARIMA and LSTM and found that the former had the lowest values of MAPE in all categories of drugs, which indicated its better efficiency. These contrasts demonstrate why it is necessary to benchmark models with existing methods and ML's dominance in drug forecasting.

The final column lists each study's findings, which also illustrates how machine learning affects drug forecasting. As illustrated in Dutta et al. (2022) the Linear Regression model performed better than others and showed the best MAPE of 19.07% in predicting sales of drugs and Manna et al. (2023) reported that the XGBoost model showed the highest accuracy in predicting sales of drugs, such as M01AB and N02BA. The model based on XGBoost presented in Pall et al. (2023) allowed predicting 69% of drug shortages with a 0.44 kappa statistic, which is a huge leap in forecasting accuracy. Gurnani et al. (2017) have found that STL Decomposition has delivered the most promising results in terms of forecasting, as well as the RMSE value was the lowest compared with the other models. In the same manner, Zeng et al. (2025) concluded that their TS-LGBM model showed the best performance among all the models and predicted the results more accurately. Angula & Dongo (2024) verified that machine learning (ML) enhanced demand forecasting despite adoption issues. These findings support how ML methods can improve drug sales forecast, decision-making, efficiency, and cost savings.

## Main Results of the 2nd Section

This section's analysis demonstrates the different machine learning techniques applicable to forecast pharmaceutical sales and to optimize the business process in the management of the pharmacy. As a key conclusion, it can be emphasized that machine learning approaches, especially XGBoost and LSTM, are becoming essential to precise pharmaceutical sales prediction. These methods are effective as they outsmart conventional statistical models such as ARIMA and Holt-Winters as they can deal with non-linear trends and grasp intricate dynamics of the time-series data. It is also demonstrated that machine learning, especially XGBoost, approaches harshly outperform linear regression-based models in terms of drug sales prediction, as revealed by both Manna et al. (2023) and Fourkiotis & Tsadiras (2024).

The analyzed literature suggests that hybrid frameworks involving a combination of several machine learning methods, e.g., CNN-LSTM or XGBoost and LSTM, outperform in terms of the accuracy of the forecast. This model combines the advantages of the two models to be able to capture long-term dependencies in the time-series data as well as non-linear relationships in the sales pattern. Saena & Suttichaya (2023) and Gurnani et al. (2017) established that such hybrid models can be more accurate than the single-algorithm approaches, offering a more stable methodology for predicting pharmaceutical sales at various time scales.

Another essential outcome of this part is the significance of data preparation and feature engineering. This observation was confirmed by a number of studies, like Xie & Zhang (2021) and Pall et al. (2023), who found that time-based features, i.e. sales data aggregated by day, week, month, and seasonality, are crucial to the successful performance of machine learning models. Furthermore, other features, such as price changes, promotions, and stock-outs, were observed to improve the predictive performance of the models. This points to the fact that such machine learning applications which are used in forecasting require accurate well-structured data.

The reviewed studies also made performance assessment to be one of their core subjects. The performance of the various machine learning models was deduced using the various performance metrics, such as MAPE and RMSSE. The findings revealed that the XGBoost model recurrently got the most accuracy and computational efficiency when compared to other models such as ARIMA, Linear Regression, and Ridge Regression. Specifically, the study by Fourkiotis & Tsadiras (2024) demonstrated that the XGBoost model resulted in the lowest MAPE and MSE values in several drug groups, thus proving to be an efficient pharmaceutical sales prediction model.

Finally, the section highlights the significance of novel methods in solving more complicated forecasting issues, including drug shortage. As an illustration, Pall et al. (2023) have applied the models based on XGBoost to predict drug outages in Canadian pharmacies with an accuracy of 69%. That way, machine learning could be used to not only optimize the sales forecast but also improve inventory management and supply chain activities in the pharmaceutical sector. The application of machine learning in predicting a shortage of drugs is a major addition to the process of enhancing the effectiveness of pharmacy stores in their operations.

# **Proposed Approach**

## Introduction to the Proposed Method

The method makes the best possible drug sales forecasting in pharmacy stores with the help of a machine learning model -XGBoost. The procedure entails a number of important steps: data collection, preprocessing, feature engineering, model training and evaluation. The aim is to make the correct estimation of the drug sales according to the past data and the seasonal patterns help the pharmacies to make the best forecasting.

The approaches start with the historical sales data collection of the pharmacy store that contains critical variables that include drug IDs, sales quantities, prices, and seasonal data. Preprocessing is important because this data is usually inaccurate or has missing values. Missing values are accounted for in this step. All categorical variables are converted into numerical forms and any outliers or irrelevant observations were removed due to the making of data clean and to be used in the modeling process. After, the feature engineering step is done by generating features that make sense by generating lag feature (sales of previous days), statistics that roll (moving average) and season indicators (Winter, Spring, etc.). These features assist the model in recognizing valuable trends and seasonal changes in the sale of the drugs.

The resulting data is then fitted into the XGBoost model. This algorithm is selected because of its capacity to deal with complicated data sets and reflect non-linear connections. Through training, the model gets to learn how to reduce the errors between the predicted and actual sales. After, the model evaluation tests the performance using error measures like RMSE and MAPE. Hyperparameters (learning rate or the number of trees) of such a model can be optimised in the event that this is necessary to achieve accuracy.

At last, sales forecasts are generated by the model once it is trained and optimized to assist pharmacies in prediction. The model is implemented in a way such that new updates of forecasts can be made on the basis of new sales. Besides, the model will be seasonal, which makes it very relevant in line with businesses in pharmacy stores.

## BPMN Diagram: Drug Sale Prediction using XGBoost (Seasonal Focus)

The BPMN diagram below shows the process of predicting the sales of drugs in a drugstore by applying the XGBoost algorithm with seasonal emphasis. The workflow is designed along the two major jobs of a Data Engineer and an ML Engineer. The workflow identifies the process involved in the collection, processing, modelling and deployment of the data that gives the sales forecasts. The system applies machine learning to forecast the demand for drugs and so that pharmacies can optimize their inventories per season (Winter, Spring, Summer, and Autumn). Workflow is cycling and the forecasting is renovated periodically according to new information.

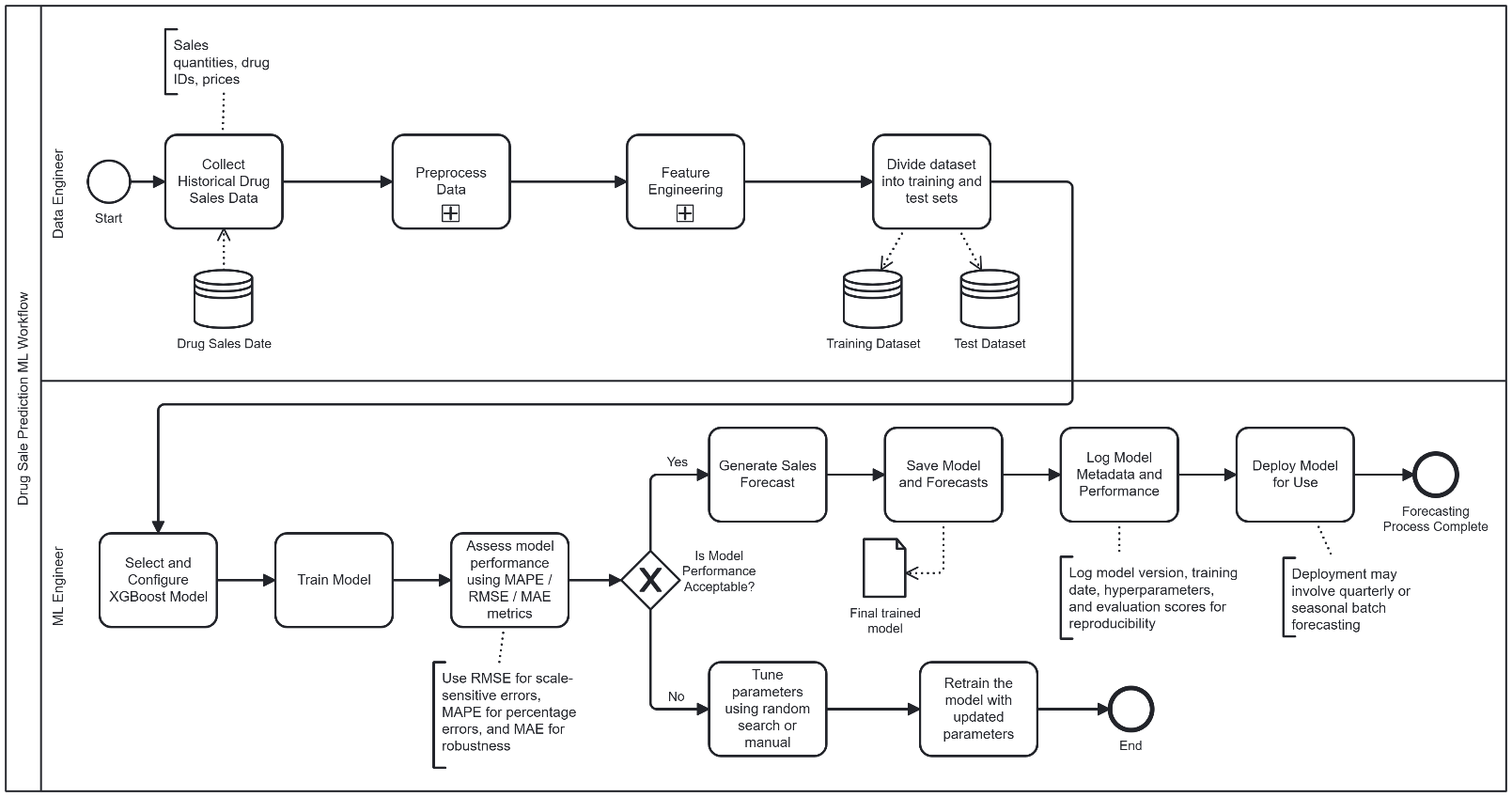


Figure . BPMN Table of Drug Sale Prediction using XGBoost

The diagram consists of 2 swimlanes:

1. Data Engineer: Handles data collection, preprocessing, and feature engineering.
2. ML Engineer: Focuses on model selection, training, evaluation, and deployment.

**Swimlane 1: Data Engineer:**

1. **Start Event: Forecast Cycle Started**

The process start with a forecast cycle activation.

1. **Task: Collect Historical Drug Sales Data**

The Data Engineer collects historical drug sales data from internal systems or external databases. This data is the main input of the forecasting system. It is stored in a central database and can be used as part of further forecasts. The data normally consists of:

* Sales quantities: Quantity of drugs sold at given times.
* Drug IDs: Unique identifiers of the drugs, or groups of drugs.
* Prices: Prices on which drugs were sold.

1. **Sub-Process: Preprocess Data**

It is a subprocess that consists of various major steps of data-cleaning and preparation in order to make data fit the model training purposes.

1. **Sub-Process: Feature Engineering**

In this subprocess, the Data Engineer generates features to enhance the predictive performance of the model.

1. **Task: Divide Dataset into Training and Test Sets**

The Data Engineer then divides the data into training and test data. An example of such proportions could be 80/20 where 80% of the data is used to train the model and the other 20% will be used to test the model performance. This ensures that the model performs on a set of data which it has not been trained with, getting an unbiased score of the performance of the model.

**Swimlane 2: ML Engineer:**

This is the lane that shows the ML Engineer who is in charge of configuring, training, evaluating and deployment of the machine learning model.

1. **Task: Select and Configure XGBoost Model**

In this step, the ML Engineer selects the XGBoost model and configures its key parameters, such as:

* Learning rate (regulates the contribution of each tree).
* Maximum depth (defines the depth of trees).
* Number of estimators (number of trees).
* Objective function (recommended regression in this case to predict the sales of drugs).

1. **Task: Train Model**

XGBoost model is one that is fitted to the training data and minimizes the error function using repeated additions of decision trees. Each of the trees tries to correct the mistakes of the prior tree through gradient boosting.

1. **Task: Assess Model Performance**

The trained model is tested with some important metrics:

* RMSE (Root Mean Squared Error): Measures large errors.
* MAPE (Mean Absolute Percentage Error): Offers percentage-based accuracy.
* MAE (Mean Absolute Error): Measures the average absolute error.

These are metrics with which the ML Engineer can thoroughly assess the performance of the model.

1. **Exclusive Gateway: Is Model Performance Acceptable?**

This is a decision point and in this point, the ML Engineer is to determine that the performance of a model is enough to achieve the standards set.

* Yes: in case the performance of the model is acceptable, the procedure proceeds to the generation of sales forecasts and saves the model.
* No: Here, the task of the ML Engineer is to tune the hyperparameters of the model by applying a strategy such as grid search, random search or brute force tuning (trial and error). This is aimed at obtaining the optimal set of parameters that lead to minimal error and model performance. After parameters have been tuned the model is retrained on the training data using these new hyperparameters. Such an iteration will go on until the performance of the model reaches the required standards.

1. **Task: Generate Sales Forecast**

When the model is trained and appraised satisfactorily, it produces sales estimates of the drugs in the coming time. For example: next season. Such forecasts may consist of the amount of predicted number of each drug each season (Winter, Spring, etc.).

1. **Task: Save Model and Forecasts**

The trained model of XGBoost and the results of the forecasts are stored as a file after they are generated.

1. **Task: Log Model Metadata and Performance**

The model’s metadata is recorded because it is required to be reproducible and traceable. This is crucial in ensuring the preservation of the history of model performance and in order to become transparent during an audit and retraining. The most important details that are recorded are:

* Model version.
* Training date.
* Hyperparameters used.
* Evaluation scores from the training process.

1. **Task: Deploy Model for Use**

At last, the model is being trained to be applied in operations. As the drug sales forecast relies on seasonal forecast, batch forecasting which is common in each new season is usually performed in this process. The model is carried out on a cyclical basis (seasonally) in order to draw new forecasts.

1. **End Event: Forecasting Process Complete**

The process would come to an end when the forecast is implemented. The forecasting cycle can now be repeated into the next cycle and it will be at this stage that updated data will be gathered and fed into the cycle to begin the other stage of forecasting.

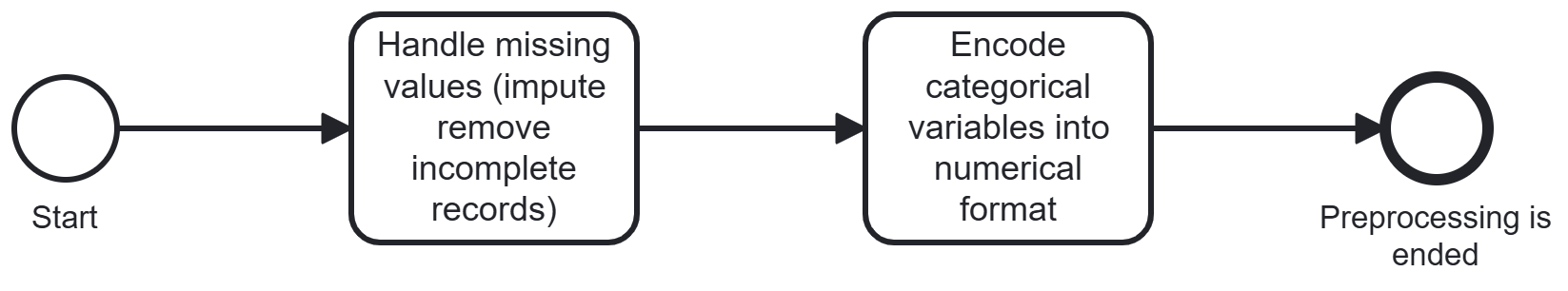


Figure . Sub-process: Preprocess Data

The "Preprocess Data" sub-process makes collected raw data ready to be modelled so that it is clean, consistent, and in a form in which it will be accepted by the XGBoost model.

1. **Task: Handle Missing Values**

This stage deals with the treatment of missing data either by imputing or by deleting rows or columns that contain too many missing values. This makes the data set full and trustworthy.

1. **Task: Encode Categorical Variables into Numerical Format**

Categorical data get a transformation into numbers (label encoding or one-hot encoding) and can be used as inputs to the XGBoost model.

1. **Task: Encode Categorical Variables into Numerical Format**

The data is then prepared to be dealt with by the next phases of the pipeline (feature engineering and model training) after the calculation of missing values and encoding of categorical variables.

This sub-process is essential in preparing the data so that it makes it clean and correctly structured to best perform the XGBoost model.

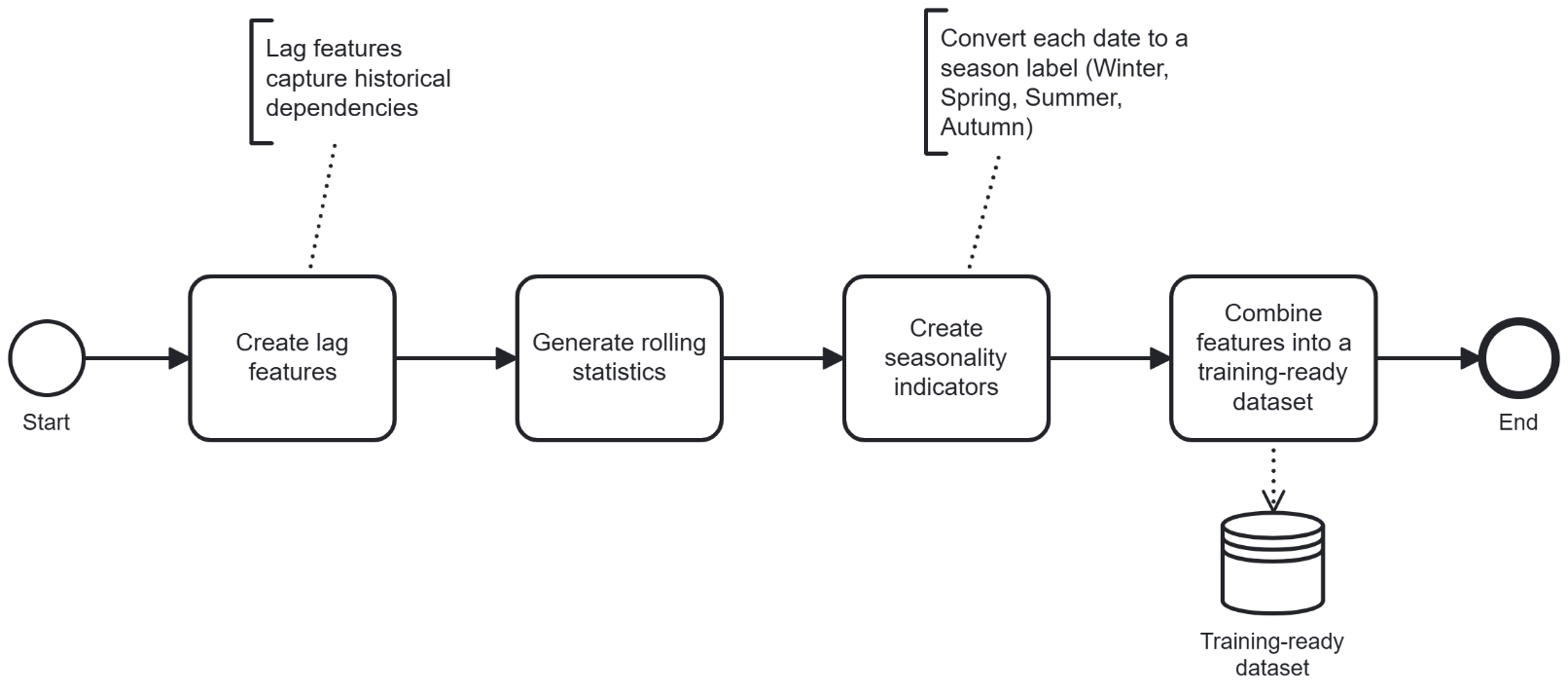


Figure . Sub-process: Feature Engineering

The sub-process of Feature Engineering will convert the raw data to meaningful characteristics that increase the predictive capabilities of the machine learning model. This procedure helps to include the most important patterns in the dataset, including seasonality and the history that is critical when predicting drug sales.

1. **Task: Create Lag Features**

These properties will capture historical sales data (sales from 1, 7, or 30 days ago) to enable the model to learn that there are time dependencies and be able to predict them at a later date using the previous historical behaviour.

1. **Task: Generate Rolling Statistics**

This involves computing the moving averages (e.g. 7-day) and standard deviations to assist the model in detecting local trends and deviations in the sales.

1. **Task: Create Seasonality Indicators**

This data is encoded into the season label (Winter, Spring, Summer, Autumn) to capture seasonal variations in sales so that the effects of the season on the demand are known by the model.

1. **Task: Combine Features into a Training-Ready Dataset**

The set of all newly constructed features is referred to as a final dataset and it is now eligible to be trained on models.

## BPMN Diagram: Drug Sale Prediction Process in Pharmacy Stores

The present BPMN diagram maps an overall business process of applying machine learning (ML) methods, namely XGBoost to forecast drug sales as an example in a pharmacy setting. The aim of the process is to facilitate and enable seamless management of inventory and optimisation of stock through forecasting the future sales of drugs per season (Winter, Spring, Summer and Autumn). The workflow has been broken down into three main swimlanes, and each swimlane depicts a different role or system that is involved in the forecasting business: Pharmacy Staff, Forecasting System, and Pharmacist. The entire exercise has a sequence that appears logical starting with the collection of the data to the ultimate application of the forecasts in replenishing the inventories.

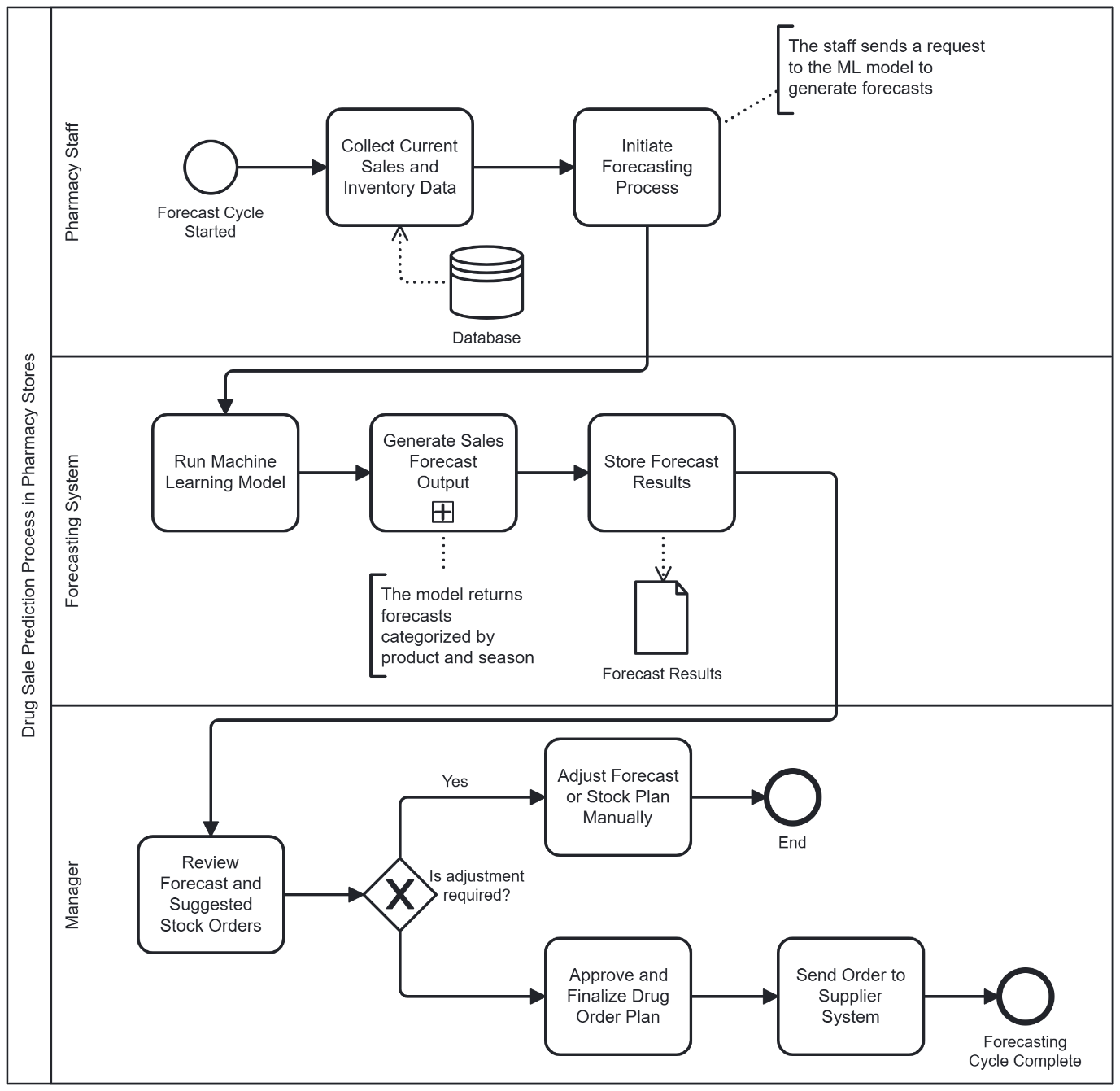


Figure . BPMN Table of Business Process

The diagram consists of 3 swimlanes:

1. Pharmacy Staff.
2. Forecasting System.
3. Manager.

**Swimlane 1: Pharmacy Staff:**

1. **Start Event: Forecast Cycle Started**

The process starts with a forecast cycle activation.

1. **Task: Collect Current Sales and Inventory Data**

The pharmacy personnel obtains the details of sales and inventory directly through information systems or databases of the pharmacy. This data will normally contain:

* Sales quantities.
* Drugs ID.
* Price of drugs.

An annotation tells that high-quality predictions are critical for accurate and up-to-date data. All this information gets stored in a central database that can be retrieved by the forecasting system.

1. **Task: Initiate Forecasting Process**

After gathering the necessary data, the members of the pharmacy initiate the forecasting process. It is achieved through sending a demand to the forecasting system to commence the drug sales forecasting process in the coming time (the next season). The dashboard or system interface could automate the request. The task is an opening to a machine learning model and with the collected data, future sales can be predicted.

**Swimlane 2: Forecasting System:**

1. **Task: Run Machine Learning Model**

The data collected is run by the forecasting system in the pre-trained XGBoost model. The process is computationally intensive and may be time-consuming based on the quantity of data and the intricacy of a model. The objective is that the system calculates sales forecasts for the next operation period (upcoming season).

1. **Task: Generate Sales Forecast Output**

After running the input data through the machine learning model, it gives the output of sales forecast that predicts:

* The amount of each drug could assume selling in the next period.
* These are seasonal (Winter, Spring, etc.) and drug-based forecasts and the pharmacy can make a seasonal demand schedule based on the forecasts.

The output is a structured data file (e.g. CSV, JSON or database records) containing the predicted quantities of sales of each drug.

1. **Task: Store Forecast Results**

The results that are forecasted are stored in a report file that can be easily accessed and used subsequently. This enables reviewing, analyzing and utilising the forecasts in the planning of inventory.

**Swimlane 3: Pharmacist:**

1. **Task: Review Forecast and Suggested Stock Orders**

The forecasting system produces data on forecasted sales and this is reviewed by the manager. Such a review can be conducted via a dashboard, report or even a spreadsheet which shows the anticipated quantities on the sales of the various drugs at the various seasons.

1. **Exclusive Gateway: Is Adjustment Required?**

It is at this level that the manager determines whether or not manual changes need to be done to the forecast. It guarantees that the end order plan is correct and viable taking into account any land local conditions or irregularities that cannot be reflected in the model. They include:

* Yes: In the event that the manager realizes that some modification is needed (order quantities of high-demand products have to be increased), they will change the forecast or stock plan manually.
* No: In case the forecast appears to be sensible and does not need any revision, there will be no additional changes in the plan.

1. **Task: Approve and Finalize Drug Order Plan**

After the manager makes a drug order plan after any required changes, it is approved. This is where the estimates get to be the official order quantities for the next period.

1. **Task: Send Order to Supplier System**

Once approved, then order details are fed into the supplier system to be furnished.

1. **End Event: Forecasting Cycle Complete**

The process is ended.

## Advantages of the Proposed Method

### Comparison of Models

Table . The differences in main aspects between algorithms

|  |  |  |
| --- | --- | --- |
| **Aspect** | **XGBoost** | **ARIMA** |
| Prediction Accuracy | High accuracy, especially for complex and non-linear patterns. | Limited accuracy for complex, non-linear patterns. |
| Handling Non-linear Relationships | Can model non-linear relationships effectively. | Primarily designed for linear relationships. |
| Seasonality Handling | Can incorporate seasonal indicators as features for better predictions. | Handles seasonality with seasonal ARIMA (SARIMA), but may not be as effective in complex cases. |
| Feature Engineering | Requires feature engineering (lag features, rolling stats, etc.). | No feature engineering required, works directly with time series data. |
| Data Requirements | Works well with large datasets and complex data types. | Typically works best with smaller, simpler datasets. |
| Real-time Forecasting | Can adapt to real-time forecasting with continuous updates. | Typically used for static forecasts with historical data. |
| Prediction Accuracy | High accuracy, especially for complex and non-linear patterns. | Limited accuracy for complex, non-linear patterns. |

XGBoost is a developed machine learning algorithm, which has proven to be very useful in predicting non-linear, complicated relationships amongst data. It is particularly beneficial in drug sales forecasting because there have been many factors that may affect the data related to this concept whether it is pricing, promotion, market conditions, and seasonality. Among the greatest benefits of XGBoost is its capability to work with big data sets and to derive complex patterns within the data set. Due to the fact that seasonality plays an important role in the prediction of drug sales, XGBoost has the potential to increase the accuracy of prediction by including seasonality features by using season (winter or summer) indicators. Further, XGBoost needs feature engineering due to feature optimization purposes. This comprises the designing of lag features (sales of the past weeks) and rolling averages (a moving average of sales) that will aid the model in comprehending the past sales patterns which enhances the prediction of the model.

ARIMA (Autoregressive Integrated Moving Average) is a classical statistical model, similar to XGBoost, which is mainly used in managing time series data that demonstrate linear relationships. It works well when there is a predictable pattern in data underlying it and it is stationary to some extent. ARIMA is able to deal with seasonality by using seasonal variables (SARIMA or Seasonal ARIMA), though it becomes less useful once required to deal with more complicated, non-linear relationships, with these appearing frequently in real sales data. XGBoost is much more scalable and has flexibility which suits large datasets and has different types of features like numerical and categorical. ARIMA is more specifically appropriate in smaller simple data and it necessitates the stationary data and this frequently needs further preprocessing.

Real-time forecasting is one of the other most important distinctions between the two models. XGBoost will fit continuous and real-time updates, such as pharmacy stores where sales variables keep on fluctuating. This means that the model is flexible so that new information can be entered regularly to enable the pharmacies to change their predictions in real time.

Finally, in comparison, ARIMA is generally applicable in static forecasting and needs redress when new information gets issued. This means that XGBoost is more applicable in conditions of dynamic environment where the data is changed regularly, but not in case that the data would follow a pattern that could not be changed regularly.

## Main Results of the 3rd Section

The 3rd part of the document is devoted to the BPMN chart on a prediction of drug sales based on the XGBoost model with a seasonal concern. The process of predicting drug sales described in the diagram represents the workflow in which tasks are organized and distributed among the work of the Data Engineer and ML Engineer. It is designed in a cyclical manner where old data is continually updated by gathering new information and re-running the models to come up with new forecasts of sales made on drugs therefore enabling the pharmacies to plan their inventory accordingly according to the seasons (winter, spring, summer and autumn).

The BPMN diagram involves two swimlanes and they describe the roles of the Data Engineer and the ML Engineer. The Data Engineer is assigned with a collection of Historical data on drug sales, preprocessing of the data and engineering features. Such procedures are meant to clean up the data and make it ready to be trained. The data Engineer then splits the data into two parts training and test sets, and this process is the concern of the unbiased assessment of the model. On the other hand, the ML Engineer will be tasked with the process of choosing and parameterizing the XGBoost model, training the model with regard to the pretested data, testing the model on the prepared data and evaluating the model's performance parameters thanks to such metrics as RMSE, MAPE, and MAE.

Model evaluation and tuning are also stressed in the section. In case the performance of the model is below acceptable, the ML Engineer adjusts the hyperparameters by approaches like grid search or random search. After the abidance of performance requirements of the model, the system will make sales estimations of each of the seasons, and these estimations are retained together with the metadata of the model to be used in the future as well as for audit purposes. The last step of this workflow is the deployment of the model whereby new forecasts are produced on a regular basis using new data.

In addition, a sub-process of data preprocessing is provided in the section including such operations as filling gap values and encoding categorical variables to make the data ready to use in the XGBoost model. It also describes the process of feature engineering, in which the Data Engineer would build features like lag features, rolling statistics and seasonality indicators to make the model more successful in making accurate predictions.

Finally, the 3rd section contains the descriptions of the workflow of drug sale prediction, which includes the stages of gathering the data, its preprocessing, model selection, training, testing and deployment. This cyclical business process makes the forecasting model current and accurate and makes pharmacies able to maintain their stocks and address the seasonal changes in demand.

## Future Work

Although the developed drug sales forecasting model based on XGBoost is rather useful, it can be enhanced by taking into consideration several approaches that will make it more correct and flexible. There are other possible machine learning algorithms that would be used in the future, like neural networks or ensemble methods to reflect more complicated relationships. Further, the use of external data besides weather patterns, public health data and market trends may also be used to make the model more predictive in future. The switchover to real-time forecasting through the incorporation of live sales figures would enable dynamic movement in inventory handling as well as sensitive decision-making within the company.

# **Initial Analysis of Drug Sale Dataset in Pharmacy Store**

## General information about datasets

### 1st dataset (Turkish Pharmacy Store’s Data):

The following features are some general understandings of what kinds of data presented in that dataset (some of them were translated from Turkish):

* 1. **IlacReceteTutari (Drug Prescription Amount):** Amount of money concerned with the drug prescription. This is the financial element of the drug(s) prescribed to the patient.
  2. **DrugSerNo (Drug Serial Number):** The unique identifier of any drug in the drug database or drug inventory of the pharmacy.
  3. **ReceteBrutTotal (Gross Prescription Total):** The entire prescription amount without deductions or modifications.
  4. **Pharmacy:** Name of the pharmacy store through which the prescription was processed.
  5. **Medicine (Sale Quantity):** Quantity of drugs sold according to the receipt.
  6. **ERecete (E-Prescription):** A binary variable (e.g. 1 or 0) that indicates whether the prescription was carried out electronically (e-prescription) or in a paper-based prescription.
  7. **EReceteNo (E-Prescription Number):** The individual identifier or number to an electronic prescription.
  8. **The Patient’s Debt:** The funds which the patient is being charged due to the drugs or services rendered by the pharmacy.
  9. **PatientTour (Patient Turnover):** This is the frequency or number of prescriptions given to a particular patient for a period of time.
  10. **Hkp (Hospital-acquired Pneumonia):** A medical term indicating pneumonia acquired in hospital.
  11. **ReceteIlacFarki (Prescription Drug Difference):** The cost of price variance of the prescription drug and any other choice or generics.
  12. **Field of Medicine:** The specific medical department of medicine (e.g. cardiology, oncology) to which the drug prescribed is concerned.
  13. **ReceteDrug DomainName (Prescription Drug Category):** The type of the prescribed drug (e.g. antibiotics, pain killers).
  14. **Recipe:** This is the prescription of the medication which means a particular formulation or brand of a drug which was prescribed.
  15. **ReceteIndOnTop (Prescription Indicator on Top):** This means that there is extra treatment under prescription or a special requirement.
  16. **Recetion (Prescription Submission):** States whether the prescription is taken to be processed or is pending.
  17. **Public Discount:** The discount on public health prescriptions is usually used in the public health systems.
  18. **Salary:** It means the wages of the pharmacist or the medical personnel participating in the medicine dispensation.
  19. **ReceteNetAmount (Price of Drugs):** The remaining value after discounts or deductions has been deducted from the prescription total.
  20. **ReceteKP (Prescription KP):** This is an administrative code regarding the prescription.
  21. **RecipeDate:** The date of prescription or filling of the prescription.
  22. **Reporting:** This aspect is the kind of report that is created off the prescription.
  23. **NobetciMi (On-Call Status):** A binary indicator that indicates whether the pharmacy was doing an on-call shift or night shift when the transaction was done.
  24. **Season (Season of the year):** The season when the drug was sold (Winter, Spring, Summer, Autumn).
  25. **ID:** The distinctive code of the transaction.
  26. **Drug Group:** The type or the class of drug.
  27. **Receipt Date (Date of Sale):** The date on which the purchase of the drug was done.

### 2nd dataset (Polish and German Pharmacy Stores’ Data):

The following features are some general understandings of what kinds of data presented in that dataset:

* 1. **Distributor:** Name of the supplier to whom the products are being distributed.
  2. **Customer Name:** This is the name of the customer buying the products.
  3. **City:** The locality of a customer.
  4. **Country:** The country of location of the customer.
  5. **Latitude:** Geographic latitude of the location.
  6. **Longitude:** The geographical longitude of the location.
  7. **Channel:** The primary sales channel in which the product is marketed.
  8. **Sub-channel:** More specific type of sales on the channel.
  9. **Product Name:** The product being sold is called.
  10. **Product Class:** The classification of the product or the type of product.
  11. **Quantity:** The number of the sold goods.
  12. **Price:** The sale value of the product.
  13. **Sales:** The amount of money which will be made in total due to the sale.
  14. **Month:** The month in which a sale was undertaken.
  15. **Year:** The year where the sale was made.
  16. **Name of Sales Rep:** The name of the person in the sales department that conducted the sale.
  17. **Manager:** The manager effective in the sales process.
  18. **Sales Team:** The team responsible for the sales transaction.

## Dataset Analysis

Table . Table of Dataset Comparison

|  |  |  |
| --- | --- | --- |
| **Aspect** | **Turkish Pharmacy Store’s Data** | **Polish and German Pharmacy Stores’ Data** |
| Data Focus | Has extensive data on individual drug sales, and patient and dose details and thus at a pharmacy level accurate forecasts can be made. | It offers a detailed perspective of distributors, sellers and finance processes that pertain to drugs, hence it can be applied in a broad-based forecasting of sales. |
| Structure | Organized in a format of patient, doctor, and transaction-wise information, which might need more preprocessing to be able to work with the fine data to make the predictions. | Structured with distributor, customer and transactions. Easy to pre-process to macro-level predictions. |
| Use Case | Less concerned with regional forecast and focusing on individual drug sales hence it does allow one to have a detailed forecast but in cases where regional sales are required an aggregation might be required. | More applicable in predicting the sale of drugs on the regional level, taking into consideration the performance of the distributors and the market behaviour. |
| Granularity | Very detailed, excellent for conducting both specific and accurate sales forecasts at the per-drug or per-pharmacy level, but not so well for bigger-scale sales forecasts. | Less detailed but useful for overall trends and patterns that affect the scaling of sales. |
| Flexibility | Very flexible in fine-casting a forecast on a smaller scale and can need further processing to suit the large models. | Making predictions at a macro-level is more flexible, which is particularly helpful with the integration of economically or demographically oriented databases in XGBoost applications. |
| Complexity | Complexity may pose a problem that also needs feature engineering to effectively perform in machine learning models. | Less complex to structure, and because of that it becomes easier to control as input data to machine learning algorithms such as XGBoost. |
| Strengths | Optimal at such modelling and predictions of some drug selling tendency, it best to use XGBoost when it is necessary to have more detailed predictions per one drug or one pharmacy. | Perfect to predict the sale of drugs in a region through the habits of distributors and clients, so that the model of XGBoost can be trained effectively. |
| Weaknesses | It can also be unsuitable for general predictions and will result in overfitting without proper processing due to high granularity. | Does not provide the level of specific knowledge about drugs. May aggregate forecasts on a drug-by-drug basis. |
| Data Coverage | Represents the sales of certain drugs in pharmacy store which is very useful when making focused predictions. | Good coverage of the market trends in Poland and Germany, which is helpful in making sales forecasts. |
| Integration Potential | More difficult to build into other larger datasets but outstanding characteristics to XGBoost in predicting at the individual drug level. | Easier to merge with other data sets to have a complete modeling of forecasts which is a good fit with XGBoost at a higher level. |

The second data presents a wide outlook on the market trends, and it concentrates on distributors, customers and finances. This qualifies it as a perfect tool where it can be combined with statistics. It is less elaborate and simpler to preprocess and use in machine learning algorithms such as XGBoost. In addition, it gives detailed information at the drug level hence raising the accuracy of an individual drug prediction.

The General Drug Sale Data, on the other hand, contains detailed information about drugs sold, patient-based data and accompanying transactional details. Such detail level makes it ideal for forecasting at a pharmacy level or at an individual drug level is possible and provides accurate insights. Nonetheless, it takes underlying big levels of preprocessing and feature engineering to be applied meaningfully in the machine learning models.

Concerning integration, the second data is convenient to merge with other datasets to make class-level analyses, whereas the General Drug Sale Data is the best standard when it comes to learning intricate facts about a particular drug and its pharmacy behaviour. Each of the two datasets is distinctly advantageous, but especially because of the size of the data second one is better than the first one, which was shown a larger number of sales over a long period of time, which is more suitable for its use in machine learning.

## The Spearman rank-order correlation coefficient

The Spearman rank-order correlation is a measure of the statistics to assess the strength and direction of the monotonic relationship, between two variables expressed vis-a-vis the point on the scale. Unlike Pearson correlation, which measures the linearity of relationships, Spearman rho deals with monotonic relationships, i.e. it measures whether one variable tends to increase or to decrease monotonically, as the other variable increases, without demanding a particular functional form. Spearman rank correlation is determined in accordance with the order of the ranking of the values as opposed to actual values. It basically takes the raw data ranks them and then calculates the Pearson correlation coefficient of the ranks (Van den Heuvel & Zhan, 2022).

The range of Spearman correlation is between -1 and +1:

* +1: An ideal positive association (when one variable goes up, the other variable rises in a perfectly monotonic style).
* -1: Ideal negative correlation (when one variable rises the other decreases in a perfectly monotonic way).
* 0: No correlation (no monotonic relationship).

Spearman rho formula:

* d\_i = the difference of the ranks of each two values.
* n = the number of data points.

Spearman rank-order correlation is critical in understanding the relationships between the various variables (drug IDs, sale quantities, date of sale, seasonality, and price of drugs). This project aims to predict the sales of drugs, and it is important to know the interaction of the features with each other to enhance the performance of the model in predicting the sales of drugs.

**Identifying Strong Relationships:** There are certain interesting tendencies, for example, a small positive relationship between "Sale Quantity" and "Month of Sale", "Sale Quantity" and "Year of Sale". Such correlations imply that gradually, there might exist a slight growth in sales with time. Also, there is a negative relationship between "Price of Drugs" and "Sale Quantity", hence higher priced drugs could have some negative effect on the quantity of sales, however, the relationship is not significant. Despite weak correlations, they give an idea of how time and price can generate a "Sale Quantity".

**Feature Selection:** Correlation in the variables “Season of the Year“ and „Sale Quantity“ produces a positive relationship, implying that seasonality plays a role in sales, but not a strong one. This association implies that in some seasons, sales are a bit higher, especially in the Winter when there is more demand for products such as cold medicines. Although the correlation is low, „Season of the Year“ is an adequate feature to take into consideration in the model since it represents significant seasonal trends. The correlation indicates that seasonality has an effect on influencing sales and its inclusion can be used to get better predictions, especially in the case of predicting drug sales that are seasonal in nature.

**Handling Non-linear Relationships:** Compared to Pearson correlation which assumes the existence of a linear relationship between the variables, Spearman's rho can identify non-linear monotonic relationships. This is especially essential in the case of drug demand forecasting, where the demand patterns may not have a linear trend but can have a certain monotonous behavior over the time (seasonal changes in demand). Spearman correlation enables one to interpret these trends better hence it is an appropriate tool for complex and non-linear relationship data.

### Description of the Spearman Rank-Order Correlation Matrix

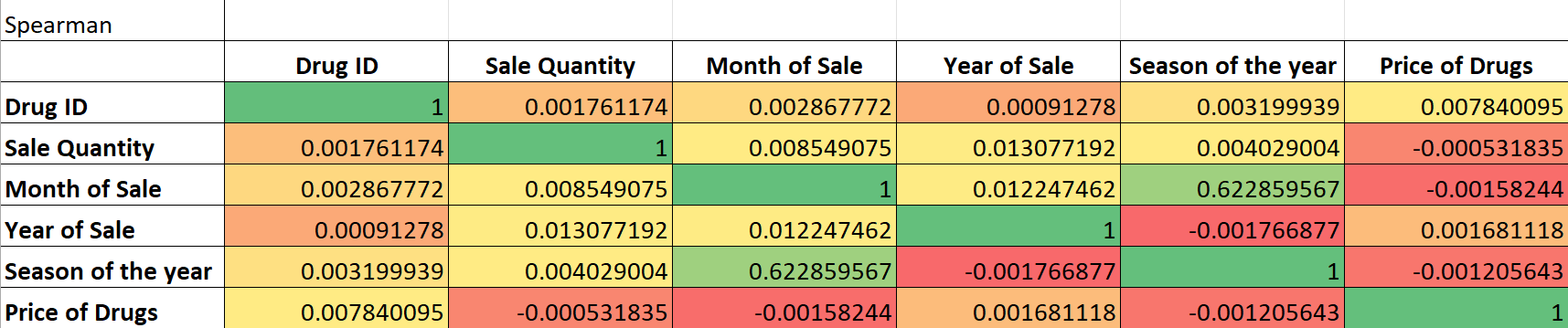


Figure . Spearman rank-order correlation matrix

**Drug ID and Sale Quantity:** Its correlation coefficient is very near weak, and that is why the variable Drug ID is very weakly positively correlated to Sales Quantity. This is simple logic since drug IDs are different products, and the relationship between the ID and quantity of a drug is likely to be poor since different drugs do not sell similar quantities.

**Sale Quantity and Month of Sale:** Sale Quantity and Month of Sale are positively related, which is an indication that sales are marginally increasing in the course of the month. This minimal correlation does not refute that the effect of seasonality persists in a way that will influence sales, particularly the products of seasonal nature.

**Drug ID and Season of the Year:**

A scanty positive relationship is likely to be true because the correlation is not very strong. This means that Drug ID is not substantially, correlated with the season of the year and this makes sense since the Drug ID does not depend on the season at which drugs are demanded.

**Drug ID and Price of Drugs:** The correlation is a very weak positive correlation between the Drug ID and the Price of Drugs. This indicates that the relationship between the price of the specific drug and the specific drug is minimal in this dataset and hence it may indicate that the prices of different drugs do not differ considerably in this dataset.

**Sale Quantity and Season of the Year:** The degree of correlation between such variables as Sale Quantity and Season of the Year is low, yet not insignificant. The correlation shows a positive relationship, indicating that there is a correlation in autumn that is associated with a slight increment in sales relative to other seasons. Nevertheless, the seasonality factor is also present even though it is imperfectly correlated in the case of seasonal drugs (cold medications during Winter). Thus, the factor Season of the Year is not to be left out of the model, even though the correlation is minor.

**Sale Quantity and Price of Drugs:** The coefficient in the relationship between Sale Quantity and the Price of Drugs denotes a negative correlation that supports the view that the selling price will vary correlated to the quantity sold. This may be a result of pricing policies or an overall demand for high-priced medicine. This can indicate that other aspects, like the demand for certain drugs or the seasonal tendency, can be more influential to sales.

**Season of the Year and Price of Drugs:** The correlation value of is very weak and positive and this indicates that the Season of the Year does not play an important role in determining the Price of Drugs in the given set of data. The prices are probably left uninfluenced by the change in demand caused by the seasons.

## Seasonal Trends of Drug Sales

These charts represent the selling change in the quantity of some drugs during the several years (2017 to 2020) during different seasons:

* 2017-S1 = Winter 2017
* 2017-S2 = Spring 2017
* 2017-S3 = Summer 2017
* 2017-S4 = Autumn 2017
* 2018-S1 = Winter 2018
* 2018-S2 = Spring 2018
* 2018-S3 = Summer 2018
* 2018-S4 = Autumn 2018
* 2019-S1 = Winter 2019
* 2019-S2 = Spring 2019
* 2019-S3 = Summer 2019
* 2019-S4 = Autumn 2019
* 2020-S1 = Winter 2020
* 2020-S2 = Spring 2020
* 2020-S3 = Summer 2020
* 2020-S4 = Autumn 2020

**A graph with numbers and lines

AI-generated content may be incorrect.**

Figure . Statistics on the number of sales of "Dantocept Ferurone" in the Autumn.

Based on the chart, it becomes apparent that the sales of this medicine are the highest in the months of Autumn, especially in the Autumn of 2019 and 2020. This means that Autumn is always associated with the increased sales of this drug as opposed to other seasons, especially the Winter and Spring when the sales here are recorded to be significantly low. This trend indicates that there is probably more demand for the product during Autumn. Those external reasons could include such input as seasonal medical issues (flu season) or market trends in these months. Thus, the months of Autumn are the months of highest performance of the "Dantocept Ferurone" and hence those are the important months to focus on in terms of sales and stocking in the pharmacy stores.

A graph showing the number of coronavirus

AI-generated content may be incorrect.

Figure . Statistics on the number of sales of "Amavirase" in the Summer.

This chart shows that the sale of Amavirase experiences a particular peak in the Summer months, especially in the Summer of 2018 and 2019. In Summer were recorded, above average sales of this drug as opposed to other seasons (Spring and Winter) where the sales are recorded to be low. The trend indicates that the demand for Amavirase might be higher in Summer. The increased sales may be caused by external efforts like seasonal health conditions (e.g., allergies, summer illnesses), or market trends at this period. Thus, the Summer months may be described as the most successful season in terms of the work of Amavirase that is why these months are the time when sales and keeping in stock are especially crucial in the pharmacy stores.

A graph showing the number of different types of doxivorin

AI-generated content may be incorrect.

Figure . Statistics on the number of sales of "Doxivorin" in the Spring.

According to the chart, the sale of Doxivorin displays a slight variation on the basis of seasons. It increases in the months of Spring at least in the Spring of 2017 and 2018. It implies that Spring constantly displays average sales of this drug when compared to the other seasons such as Winter and Summer which turn more unstable. The trend indicates that there could be a moderate demand for the product in Spring possibly because of a particular health aspect that dominates during that season, or just an increase in the number of prescriptions being issued at that time. Thus, the Spring months could be chosen as the medium-performing period of selling Doxivorin, so the period is of interest to control the content of the stocks and predict the sales in the pharmacy stores.

## Main Results of the 4th Section

The 4th section of the document was performed on two sets of drug sales: the pharmacy stores drug sales data of Turkey and the pharmacy store drug sales data of Poland and Germany. The Turkish dataset contains detailed data about drug selling in units of individuals, as well as the prescriptions and the transaction histories of patients, which would allow performing fine-scale forecasting, but the Polish and German datasets are more beneficial because it is larger and include longer periods of sales. This data contains information about customers and sales channels, which makes it more inappropriate in the case of extensive sales forecasting. The Turkish collection of highly detailed drug predictions and The Polish and German data is in a better position to make major market predictions as it is highly flexible and provides long-term knowledge, which would provide essential information about the long-term sale prediction of drugs.

The application of the Spearman rank-order correlation coefficient in the analysis of the connections between the important features in the drug sales data is provided. In the section, the advantage of Spearman correlation against Pearson correlation has been noted especially in the circumstances where the correlations between the variables are non-linear. There was also discussed the Spearman rank-order correlation in association with different variations of variables: "Drug ID", "Sale Quantity", "Seasonality", and "Price of Drugs". The Spearman correlation was chosen because it detects monotonic relationships, despite the fact that the data may not be linear. It showed some relationships, one of which is the positive relationship between "Sale Quantity" and "Month of Sale, and the other is the negative relationship of the variable "Price of Drugs" with the variable "Sale Quantity". Such correlations indicate the existence of certain trends, which proves the fact that more sophisticated machine learning models such as XGBoost are required to make more accurate predictions.

It is marked in the section as well that the correlation between the Season of the Year and Sale Quantity is weak in the negative sense. The correlation shows that Winter registered record sales than in other seasons, especially the Summer season when sales registered normally are low. This trend indicates that no matter how bad sales are at any time of the year, during winter it becomes a little bit better, as the demand for some drugs grows. Although seasonality has a low correlation, it is a significant aspect in forecasting models since it is used to predict any variation in terms of demand for a particular type of drug, say, when people have a cold and flu during winter, or buy allergy medicines in spring. Through Spearman correlation, the analysis takes into consideration these nuances and thus seasonality along with other aspects that affect the sales are well factored and a better forecast can be made of the future demand.

The Feature Selection analysis further indicated that the seasonality (especially the season of the year) has an influence on sales trends with the Winter and Autumn seasons having sales of some drugs that are higher. Even though the correlations are weak, seasonality is vital in the prediction model especially when disposing of drugs which tend to have high demand in certain seasons such as cold medications. The analysis also highlighted that incorporating seasonal indicators in the model would increase the levels of the prediction particularly when making the forecasts of sales of drugs with seasonal fluctuations.

Finally, the charts that are found in this section graphically demonstrate the advantage of sales of different drugs within the season (Winter, Spring, Summer, and Autumn). For example, sales of Dantocept Ferurone, which are maximum during Autumn, and Amavirase are most sold during Summer. These seasonality patterns demonstrate the non-linearity in drug sales because the influence of seasonal factors on the quantity of sales could not be explained by linear models only. Based on the results of the charts above, it can be concluded that to enhance further performance of the forecasting and predict more accurate values on the number of drug sales at various times of the year, more intensive machine learning methods, such as XGBoost, capable of catching such non-linear and seasonal trends, should be used. It makes pharmacies ready to handle quotas efficiently.

# **Conclusion**

Here the general conclusions are presented:

1. Based on the performed research and comparison, by using optimization techniques and predictive analytics, Machine Learning makes it possible to allocate resources in a dynamic, scalable, and efficient manner. This improves operational performance, flexibility, and decision-making in intricate business processes.
2. Based on the performed competitive analysis of drug sales data of the Turkish pharmacy stores and, Polish and German pharmacy stores, the obtained results of the Polish and German samples show that it is more useful in terms of its size and time range providing additional value concerning the regional-level prognosis. The obtained results indicate that seasonality has an effect on sales, however, to make the forecasting more accurate, it is reasonable to consider the case when advanced machine learning models such as XGBoost are integrated and external data sources are incorporated to provide real-time forecasting, which will help to optimise drug sale.

# References

Dutta, S. R., Das, S., & Chatterjee, P. (2022). Smart sales prediction of pharmaceutical products. *2022 8th International Conference on Smart Structures and Systems (ICSSS)*, 1–6. <https://doi.org/10.1109/ICSSS54381.2022.9782271>

Manna, A., Kolpe, K., Mhalungekar, A., Pattewar, S., Kaloge, P., & Patil, R. (2023). Comparative study of various machine learning algorithms for pharmaceutical drug sales prediction. *2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)*, 1–6. <https://doi.org/10.1109/IConSCEPT57958.2023.10169964>

Saena, W., & Suttichaya, V. (2019). Predicting drug sale quantity using machine learning. *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 1–6. <https://doi.org/10.1109/iSAI-NLP48611.2019.9045222>

Al-Gunaid, M. A., et al. (2018). Analysis of drug sales data based on machine learning methods. *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*, 32–38. <https://doi.org/10.1109/SYSMART.2018.8746968>

Fourkiotis, K. P., & Tsadiras, A. (2024). Applying machine learning and statistical forecasting methods for enhancing pharmaceutical sales predictions. *Forecasting, 6*(1), 170–186. <https://doi.org/10.3390/forecast6010010>

Gurnani, M., Korke, Y., Shah, P., Udmale, S., Sambhe, V., & Bhirud, S. (2017). Forecasting of sales by using fusion of machine learning techniques. *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, 93–101. <https://doi.org/10.1109/ICDMAI.2017.8073492>

Zeng, Z., Yang, W., Wang, M., et al. (2025). Chain drugstore sales prediction method based on the fusion of self-attention mechanism and LightGBM. *Journal of Systems Science and Systems Engineering*. <https://doi.org/10.1007/s11518-025-5652-1>

Dairu, X., & Shilong, Z. (2021). Machine learning model for sales forecasting by using XGBoost. *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 480–483. <https://doi.org/10.1109/ICCECE51280.2021.9342304>

Ahmadi, E., Mosadegh, H., Maihami, R., Ghalehkhondabi, I., Sun, M., & Süer, G. A. (2022). Intelligent inventory management approaches for perishable pharmaceutical products in a healthcare supply chain. *Computers & Operations Research, 147*, 105968. <https://doi.org/10.1016/j.cor.2022.105968>

Mousa, B. A., & Al-Khateeb, B. (2023). Predicting medicine demand using deep learning techniques: A review. *Journal of Intelligent Systems, 32*, 20220297. <https://doi.org/10.1515/jisys-2022-0297>

Mostofi, A., & Jain, V. (2021). Inventory management and control of deteriorating pharmaceutical products using Industry 4.0. *IEEE 8th International Conference on Industrial Engineering and Applications*. <https://doi.org/10.1109/ICIEA52957.2021.9436744>

Saha, E., & Rathore, P. (2024). A smart inventory management system with medication demand dependencies in a hospital supply chain: A multi-agent reinforcement learning approach. *Computers & Industrial Engineering, 191*, 110165. <https://doi.org/10.1016/j.cie.2024.110165>

Pall, R., Gauthier, Y., Auer, S., & Mowaswes, W. (2023). Predicting drug shortages using pharmacy data and machine learning. *Health Care Management Science, 26*(2), 395–411. <https://doi.org/10.1007/s10729-022-09627-y>

Bhat, S. S., Satheesh, S. S., Srihari, V. R., Bidrohi, A. B., & Prabhune, A. (2024). Optimizing medication access in public healthcare centers: A machine learning stochastic model for inventory management and demand forecasting in primary health services. *IEEE Conference on Advancing Digital Healthcare*. <https://doi.org/10.1109/IITCEE59897.2024.10467229>

Lam, L. D., Le Luong, B. P., Mai Linh, H. T., & Hung, P. M. (2023). *Application of machine learning in predicting the amount of pharmaceutical drugs ordered for the manufacturer*. In 2023 1st International Conference on Health Science and Technology (ICHST) (pp. 1–6). IEEE. <https://doi.org/10.1109/ICHST59286.2023.10565367>

Kalaichelvan, K., Ramalingam, S., Dhandapani, P. B., Leiva, V., & Castro, C. (2024). Optimizing the economic order quantity using fuzzy theory and machine learning applied to a pharmaceutical framework. *Mathematics, 12*(819), 1–22. <https://doi.org/10.3390/math12060819>

Rekabi, S., Sazvar, Z., & Goodarzian, F. (2023). A machine learning model with linear and quadratic regression for designing pharmaceutical supply chains with soft time windows and perishable products. *Decision Analytics Journal, 9*, 100325. <https://doi.org/10.1016/j.dajour.2023.100325>

Boute, R. N., Gijsbrechts, J., van Jaarsveld, W., & Vanvuchelen, N. (2022). Deep reinforcement learning for inventory control: A roadmap. *European Journal of Operational Research, 298*(2), 401–412. <https://doi.org/10.1016/j.ejor.2021.07.016>

Almentero, B. K., Li, J., & Besse, C. (2024). Forecasting pharmacy purchase orders. *IEEE Transactions on Industrial Engineering and Applications*. <https://doi.org/10.23919/FUSION49465.2021.9627017>

Zhu, X., Ninh, A., Zhao, H., & Liu, Z. (2021). Demand forecasting with supply-chain information and machine learning: Evidence in the pharmaceutical industry. *Production and Operations Management, 30*(9), 3231–3252. <https://doi.org/10.1111/poms.13426>

Chamekh, M., El Asmi, S., Hamdi, M., & Kim, T. H. (2017). Context-aware middleware for RFID-based pharmaceutical supply chain. *IEEE International Conference on Internet of Things and Smart Systems*. <https://doi.org/10.1109/IWCMC.2017.7986576>

Mehdiyev, N., Evermann, J., & Fettke, P. (2018). A novel business process prediction model using a deep learning method. *Business & Information Systems Engineering, 62*(2), 143–157. <https://doi.org/10.1007/s12599-018-0551-3>

Fernandes, E. C., Fitzgerald, B., Brown, L., & Borsato, M. (2019). Machine learning and process mining applied to process optimization: Bibliometric and systemic analysis. *Procedia Manufacturing, 38*, 84–91. <https://doi.org/10.1016/j.promfg.2020.01.012>

Di Francescomarino, C., Dumas, M., Federici, M., Ghidini, C., Maggi, F. M., Rizzi, W., & Simonetto, L. (2018). Genetic algorithms for hyperparameter optimization in predictive business process monitoring. *Information Systems, 74*, 67–83. <https://doi.org/10.1016/j.is.2018.01.003>

Abbasi, M., Nishat, R. I., Bond, C., Graham-Knight, J. B., Lasserre, P., & Lucet, Y. (2024). A review of AI and machine learning contribution in business process management (process enhancement and process improvement approaches). *Business Process Management Journal*. <https://doi.org/10.1108/BPMJ-07-2024-0555>

Aydemir, E., & Yavuz, M. (2020). *Drug Sales Dataset* [Data set]. Kaggle. <https://doi.org/10.34740/KAGGLE/DSV/1628858>

Van den Heuvel, E., & Zhan, Z. (2022). Myths about linear and monotonic associations: Pearson’s *r*, Spearman’s *ρ*, and Kendall’s *τ*. *The American Statistician, 76*(1), 44–52. <https://doi.org/10.1080/00031305.2021.2004922>

Inemesit Umoh. (n.d.). *Pharmaceutical‑sales‑analysis* [Data analysis project]. GitHub. Retrieved June 23, 2025, from [https://github.com/InemesitUmoh/Pharmaceutical‑sales‑analysis](https://github.com/InemesitUmoh/Pharmaceuticalsalesanalysis)