SALES ANALYSIS AND FORECASTING

TEAM 05

MEMBERS :

AKASH S V	- 221501006
ANISH A	- 221501008
ARUNA S	- 221501011
ARUNACHALAM T	- 221501012
ASHWIN KUMAR C	- 221501014
BABU NIRANJAN G	- 221501016
BALAJI S	- 221501017
BAVISHYA K	- 221501018
BHARATH BALAN P	- 221501019

Dr. BALAMURUGAN
REC Mentor

MS. JYOTHI TRIPATHI
CTS Mentor

RAJALAKSHMI ENGINEERING COLLEGE

Department of Artificial Intelligence and Machine Learning

Rajalakshmi Engineering College, Thandalam

INDEX

ABSTRACT

This report details a data analytics project focused on the analysis and forecasting of pharmaceutical sales. Using a time-series dataset of over 600,000 point-of-sale records from 2014 to 2019, this study aims to identify historical sales trends and seasonal patterns to build accurate predictive models. The project evaluates various forecasting methodologies, from traditional statistical models to advanced machine learning techniques, to determine the most effective approach for handling sales data with diverse characteristics. The ultimate goal is to provide a data-driven framework that enables pharmaceutical companies to optimize sales strategies, improve inventory management, and enhance resource allocation, thereby reducing wastage and ensuring drug availability.

Keywords *— Inventory Management,Predictive Modeling,Seasonality Analysis, Holiday Impact*

INTRODUCTION

The pharmaceutical industry operates within a highly complex and dynamic global market, where the ability to accurately anticipate future demand is not merely a competitive advantage but a cornerstone of operational excellence and public health responsibility. The intricate supply chain, from manufacturing to distribution and final sale, is exceptionally sensitive to fluctuations in consumer demand. Inaccurate sales forecasting can trigger a cascade of negative consequences, leading to significant financial losses from overstocked inventory, expired products, and inefficient resource allocation. Conversely, underestimating demand can result in critical stockouts, which not only represent missed revenue opportunities but can also severely impact patient access to essential medicines, compromising healthcare outcomes. Therefore, the transition from traditional, often intuition-based forecasting methods to a more robust, data-driven framework is an imperative for modern pharmaceutical enterprises.

This project addresses this critical need by undertaking a comprehensive, data-centric analysis of historical pharmaceutical sales. The foundation of this study is a substantial time-series dataset, capturing over 600,000 point-of-sale (POS) records from pharmacies over a six-year period, from 2014 to 2019. This granular dataset provides a rich, detailed view of consumer behavior, with attributes including the precise date and time of sale, the specific drug code (based on the Anatomical Therapeutic Chemical classification system), and the quantity sold. The quantity is measured in Defined Daily Doses (DDD), a standardized unit that allows for meaningful comparisons of drug consumption across different products and formulations.
By leveraging this historical data, this project moves beyond simple retrospective analysis. The primary objective is to dissect these time-series records to uncover and quantify the underlying patterns that govern sales, including long-term growth or decline trends, predictable seasonal cycles, and the specific impact of external events like public holidays and seasonal illnesses. Through a systematic evaluation of various statistical and advanced machine learning models, this research aims to develop a highly accurate and reliable forecasting system. The ultimate goal is to transform raw sales data into actionable business intelligence, providing a validated methodology that empowers pharmaceutical companies to optimize their sales strategies, streamline inventory management, and ensure that the right medicines are available to the right patients at the right time.

EXISTING SYSTEMS

Traditionally, sales forecasting in many sectors, including pharmaceuticals, has relied on less sophisticated methods. These existing systems often involve manual analysis of past sales reports, reliance on anecdotal evidence from sales teams, or the use of simple statistical methods that fail to capture the complex, non-linear patterns present in real-world data. Such approaches are often reactive rather than proactive and struggle to account for dynamic factors like holidays, seasonal illnesses, or other market influences. This can lead to significant inaccuracies in forecasting, resulting in inefficient inventory management, missed sales opportunities, and suboptimal resource allocation. The limitations of these existing systems highlight the need for a more advanced, data-driven approach.

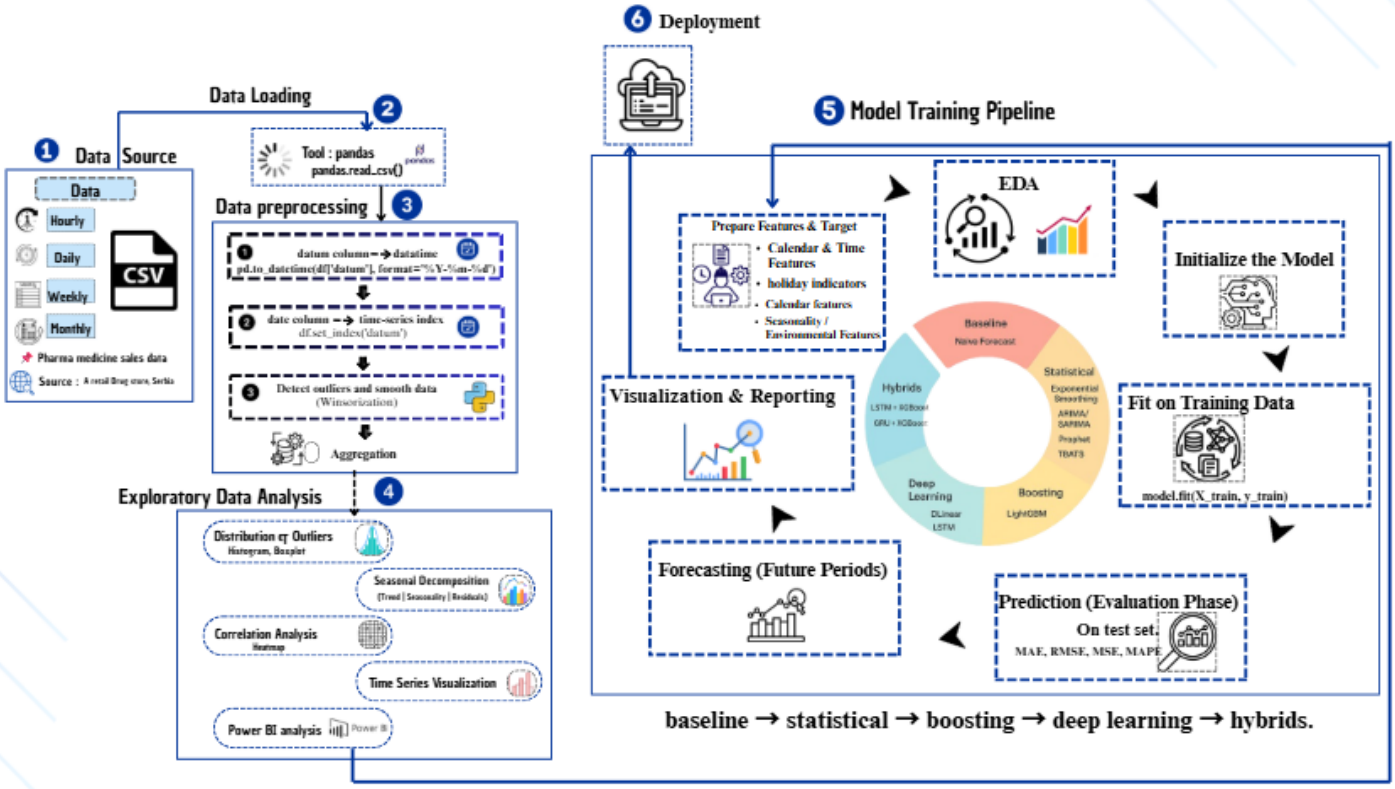
PROBLEM STATEMENT

The pharmaceutical industry is characterized by highly dynamic demand patterns influenced by seasonality, holidays, health trends, and unexpected events such as flu outbreaks or allergy seasons. Traditional forecasting methods often fail to capture these complex variations, leading to challenges such as stockouts, overstocking, and inefficient resource allocation. Pharmacies and pharmaceutical companies rely heavily on accurate demand forecasting to ensure timely drug availability, optimize inventory, minimize wastage, and maximize profitability. However, sales data in this domain is inherently time-series based, exhibiting diverse characteristics such as stationarity, seasonality, and varying sales volumes across different drug categories. This makes the problem of sales analysis and forecasting both challenging and critical. The problem lies in effectively preparing, analyzing, and forecasting large-scale pharma sales data using robust analytical and machine learning approaches that can handle trends, periodicity, variance, and sudden spikes in demand. The dataset under consideration, collected from point-of-sale (POS) systems of pharmacies in Serbia, contains over 600,000 sales records spanning six years (2014–2019) across multiple drug categories such as anti-inflammatory drugs, analgesics, psycholeptics, and antihistamines. While exploratory data analysis reveals long-term growth, seasonal peaks, and holiday-driven sales drops, translating these insights into accurate predictive models is a non-trivial task. Existing methods such as ARIMA, SARIMA, and exponential smoothing provide useful baselines but often underperform when sales patterns are highly nonlinear. Advanced methods such as boosting algorithms (LightGBM, XGBoost), deep learning models (LSTM, GRU), and hybrid approaches promise better accuracy by capturing both long-term trends and short-term fluctuations. The core problem, therefore, is to validate different forecasting approaches, identify the most effective techniques for pharma sales prediction, and recommend data-driven sales and marketing strategies that improve operational efficiency, inventory management, and customer satisfaction while addressing the volatility inherent in pharmaceutical demand.

OBJECTIVES

The primary objective of this project is to develop, validate, and compare different forecasting models to analyze pharmaceutical sales data and accurately predict future demand. By leveraging both statistical and machine learning approaches, the project aims to address the challenges posed by seasonality, trend variations, holidays, and sudden fluctuations in sales patterns. The first step is to perform comprehensive data preprocessing and exploratory analysis to identify key patterns, correlations, and anomalies within the dataset. This includes detecting outliers, analyzing long-term sales trends, decomposing seasonal effects, and integrating external factors such as holidays and environmental influences like cold/flu and pollen seasons. The next objective is to systematically evaluate a wide range of forecasting techniques, starting from traditional baselines such as naïve forecasting, exponential smoothing, ARIMA, and SARIMA, and extending to advanced methods like Random Forest, LightGBM, XGBoost, and deep learning architectures including LSTM and GRU. Furthermore, the project aims to experiment with hybrid models that combine the strengths of both deep learning and boosting methods to enhance accuracy and robustness. Evaluation metrics such as MAE, RMSE, MSE, MAPE, and R^2 will be employed to benchmark the models across various drug categories, ensuring fair and comprehensive assessment. Beyond forecasting, the project also seeks to derive actionable business insights by linking sales predictions to inventory optimization, production planning, and targeted marketing strategies. For instance, understanding seasonal spikes in analgesics or flu-related drugs can help pharmacies and pharmaceutical companies plan promotional campaigns and adjust inventory levels in advance. Ultimately, the objective is not only to identify the best-performing models for sales forecasting but also to provide data-driven recommendations that improve supply chain efficiency, reduce wastage, ensure timely drug availability, and enhance decision-making for stakeholders in the pharmaceutical industry.

ARCHITECTURE DIAGRAM



EXPLORATORY DATA ANALYSIS

The Exploratory Data Analysis (EDA) is a crucial step in understanding the structure, patterns, and behavior of pharmaceutical sales data before applying forecasting models. The dataset used in this study consists of over 600,000 sales records collected from POS systems of pharmacies in Serbia between 2014 and 2019. Each record contains attributes such as date, time, drug code (ATC classification), and quantity sold, covering eight key drug categories, including analgesics, anti-inflammatory drugs, psycholeptics, and antihistamines. The first step in EDA was to clean and preprocess the data by converting the sales date into a time-series index, aggregating transactions at daily, weekly, and monthly levels, and identifying missing values or inconsistencies. Outliers, such as extreme peaks and dips caused by unusual sales events, were detected and treated using smoothing techniques like winsorization. Trend analysis revealed long-term upward growth in sales, particularly in categories like analgesics (N02BE), which dominate during both cold/flu and pollen seasons. Seasonal decomposition highlighted strong periodic patterns, with noticeable peaks during winter months when flu outbreaks are common and smaller peaks during spring allergy seasons. Correlation analysis using heatmaps confirmed relationships between certain drug categories, such as increased antihistamine (R06) demand alongside respiratory drugs (R03) during pollen seasons. Stationarity tests, including Augmented Dickey-Fuller (ADF) and KPSS, were performed to assess whether sales series were stable over time or required differencing for forecasting models. Additionally, holiday analysis using the Serbian holiday calendar revealed significant sales drops during public holidays and sudden increases just before these periods due to stockpiling behavior. Visualizations through histograms, boxplots, and time-series line plots provided deeper insights into sales distribution, seasonality, and variance across categories. Overall, EDA not only uncovered meaningful business patterns but also guided feature engineering and model selection, ensuring that forecasting models account for seasonality, holidays, and environmental effects influencing pharmaceutical demand.

FEATURE ENGINEERING

Feature engineering plays a pivotal role in improving the accuracy and robustness of sales forecasting models by transforming raw time-series data into meaningful inputs that capture seasonality, trends, holidays, and external influences. In this project, sales records from Serbian pharmacies were enriched with multiple time-based and event-driven features to ensure models could effectively learn complex demand patterns. Calendar features such as day of the week, week of the year, month, quarter, and year were extracted from the date column to capture recurring temporal cycles. Rolling statistics, including moving averages and standard deviations over 3, 7, and 14 days, were computed to smooth short-term fluctuations and provide context on recent sales trends. Lag features were added to incorporate past sales values (e.g., 1-day, 2-day, 7-day lags), allowing models to detect autocorrelation and continuity in demand. Public and religious holidays were integrated into the dataset using the Python "Holidays" library, enabling the identification of demand spikes or drops associated with holiday effects. Additionally, custom holiday windows (days before and after holidays) were created to capture spillover effects where customers stockpile or delay purchases. Seasonal indicators, such as cold/flu season and pollen season, were introduced to reflect health-driven demand variations, with flu seasons showing sharp increases in analgesics and respiratory drugs, while pollen seasons boosted antihistamine sales. Interaction features were also engineered, such as combining holiday effects with drug categories, to highlight product-specific seasonal demand. Furthermore, normalization and scaling techniques were applied to stabilize variance across drug categories with different sales magnitudes. Feature engineering not only enhanced the dataset with domain-specific insights but also enabled advanced models like LightGBM, XGBoost, and hybrid deep learning approaches to perform significantly better than baseline statistical methods. By capturing both temporal dependencies and external influences, these engineered features ensured that the forecasting models produced more accurate, reliable, and business-relevant predictions.

MODELS

CatBoost

CatBoost is a gradient boosting algorithm that is optimized for handling categorical variables efficiently. It requires minimal preprocessing, avoids overfitting through advanced regularization, and is well-suited for datasets with both numerical and categorical attributes. In sales forecasting, CatBoost is effective at capturing complex relationships between features while maintaining high accuracy.

XGBoost

XGBoost is a high-performance gradient boosting algorithm widely used in predictive modeling. It builds decision trees sequentially, where each new tree corrects the errors of the previous ones. Known for its speed, accuracy, and regularization techniques, XGBoost is particularly powerful for capturing complex non-linear sales patterns and preventing overfitting.

Random Forest Regressor

XRandom Forest is an ensemble method that combines predictions from multiple decision trees to produce more accurate and stable results. Each tree is trained on random subsets of data and features, and the final prediction is an average of all trees. This reduces variance, handles non-linear data effectively, and prevents overfitting, making it reliable for diverse sales categories.

Naïve Forecasting

Naïve forecasting is the simplest time-series prediction method where the forecast for the next period is assumed to be equal to the most recent actual value. It serves as a baseline model to evaluate the performance of more advanced forecasting techniques. While it lacks sophistication, it is useful for datasets with little variation or as a benchmark for comparison.

Exponential Smoothing

Exponential smoothing is a statistical forecasting technique that assigns exponentially decreasing weights to past observations, giving more importance to recent data. It has three main forms: single smoothing for level patterns, double smoothing for trend-adjusted series, and triple smoothing (Holt-Winters) for seasonal data. It is widely used for short-term forecasts in demand and inventory planning.

ARIMA (AutoRegressive Integrated Moving Average)

ARIMA is a powerful statistical model designed for time-series forecasting. It combines autoregression (AR), differencing (I), and moving averages (MA) to capture trends and correlations in stationary data. ARIMA works best for short-term forecasting and is effective at modeling temporal dependencies, though it struggles with highly seasonal or non-linear patterns.

LightGBM

LightGBM (Light Gradient Boosting Machine) is a fast, scalable gradient boosting algorithm designed for large datasets and complex features. It uses leaf-wise tree growth, enabling better handling of variance and seasonality compared to traditional methods. Its efficiency and accuracy make it well-suited for forecasting pharmaceutical sales with large volumes of time-series data.

N-BEATS + LightGBM (Hybrid)

The hybrid approach combines the strengths of deep learning and boosting methods. N-BEATS, a neural network model, learns long-term trends and seasonal patterns directly from raw time-series data. LightGBM is then applied to correct short-term residuals and fluctuations, resulting in improved stability and accuracy. This hybrid approach is particularly effective for irregular and volatile sales data.

Prophet

Developed by Meta (Facebook), Prophet is a time-series forecasting model that decomposes data into trend, seasonality, and holiday effects. It is highly interpretable and user-friendly, making it suitable for business applications. Prophet works well with datasets influenced by events such as holidays and seasonal changes, though it may underperform when dealing with highly non-linear patterns.

DLinear

DLinear is a lightweight deep learning model that decomposes time series into trend and seasonal components. By using simple linear layers, it achieves competitive performance while being computationally efficient. Its simplicity makes it suitable for large-scale forecasting tasks across multiple categories without requiring heavy resources.

TBATS

TBATS is a state-space model designed to handle complex and multiple seasonalities in time-series data. It incorporates Box-Cox transformation, trigonometric seasonal terms, and ARMA errors to capture autocorrelation. TBATS performs well for datasets with irregular or overlapping seasonal cycles, making it suitable for pharmaceutical sales that experience multiple demand peaks.

LSTM + XGBoost (Hybrid)

This hybrid model combines Long Short-Term Memory (LSTM) networks with XGBoost. LSTM captures long-term temporal dependencies in sales data, while XGBoost is applied to correct residual errors and account for short-term fluctuations. Together, they handle non-stationarity, holiday-driven spikes, and abrupt changes, resulting in more accurate forecasts compared to standalone models.

GRU + XGBoost (Hybrid)

The GRU + XGBoost model integrates Gated Recurrent Units (GRU), a simplified

version of LSTM, with XGBoost for residual correction. GRUs capture temporal patterns efficiently with fewer parameters and faster training, while XGBoost enhances accuracy around holidays and irregular spikes. This combination balances performance and efficiency, making it effective for volatile sales datasets.

RESULT AND DISCUSSION

In The pharmaceutical sales forecasting project explored a wide spectrum of forecasting models, ranging from simple statistical techniques to advanced machine learning and deep learning architectures. Each model was applied to the same dataset of six years of pharmacy sales records, with attributes including drug code, date, and quantity sold. The goal was not only to forecast sales accurately but also to evaluate the comparative strengths and weaknesses of each method and determine which model was most effective for practical business decision-making. The models tested included Naïve Forecasting, Exponential Smoothing, ARIMA, SARIMA, CatBoost, XGBoost, Random Forest Regressor, LightGBM, N-BEATS + LightGBM hybrid, Prophet, DLinear, TBATS, LSTM + XGBoost hybrid, and GRU + XGBoost hybrid.

1. Baseline Models: Naïve Forecasting and Exponential Smoothing

The Naïve Forecast served as the simplest baseline, where each future value was predicted as the most recent observed sales. As expected, it produced a flat-line prediction, useful for benchmarking but inadequate for capturing seasonality, holidays, or sudden spikes. Similarly, Exponential Smoothing provided slightly better performance by assigning higher weight to recent sales values. The Holt-Winters method, which accounts for both trend and seasonality, was particularly useful in detecting seasonal cycles in cold/flu and pollen periods. However, these models lacked flexibility when dealing with sudden demand fluctuations or multiple seasonalities, making them insufficient for pharmaceutical sales forecasting in isolation.

2. ARIMA and SARIMA: Statistical Time-Series Models

ARIMA was employed to capture autocorrelation, differencing for stationarity, and moving average components. While ARIMA produced stable short-term forecasts, its performance was limited in highly seasonal categories. SARIMA extended ARIMA by introducing seasonal terms, allowing the model to account for recurring annual cycles. For example, SARIMA performed reasonably well in forecasting winter peaks in flu-related drugs. Nevertheless, both ARIMA and SARIMA struggled when sales displayed non-linear behaviors or multiple overlapping seasonalities. Their accuracy was lower compared to machine learning models, though they served as valuable statistical baselines and benchmarks.

3. Ensemble Tree-Based Models: CatBoost, XGBoost, Random Forest, and LightGBM

The ensemble models demonstrated strong forecasting ability, outperforming traditional statistical methods. CatBoost handled categorical features efficiently and required minimal preprocessing. It performed consistently across drug categories but lagged behind other boosting models in capturing extreme demand spikes. XGBoost stood out with its ability to model complex non-linear relationships and avoid overfitting using regularization and early stopping. It achieved excellent accuracy, particularly for categories like N05C and M01AE, where RMSE values were very low. Random Forest Regressor, another tree-based model, showed high stability and resistance to overfitting by averaging predictions from multiple trees. It performed exceptionally well on low-volatility categories such as R06, though it was not as strong in handling abrupt changes. LightGBM proved to be the most efficient among tree-based models, both in terms of computational speed and accuracy. By leveraging leaf-wise tree growth and incorporating rolling statistics, lag features, and holiday effects, LightGBM consistently produced low error metrics across most drug categories. In fact, it achieved R^2 values above 0.84 for many categories, making it one of the most reliable single models tested.

4. Deep Learning and Hybrid Models: N-BEATS + LightGBM, LSTM + XGBoost, GRU + XGBoost

The deep learning and hybrid approaches provided further improvements by combining the strengths of neural networks and boosting methods. N-BEATS, designed for time-series forecasting, captured long-term trends and seasonal patterns directly from raw sales data. When combined with LightGBM for residual correction, the hybrid model delivered highly stable forecasts, especially in categories with irregular seasonal patterns. Similarly, the LSTM + XGBoost hybrid leveraged LSTM's ability to capture long-term temporal dependencies and XGBoost's residual correction to handle abrupt fluctuations like holiday-driven spikes. This model significantly reduced MAE compared to Prophet and ARIMA, achieving improvements of 5–10% in accuracy. The GRU + XGBoost hybrid followed a similar logic, with GRUs providing faster training than LSTMs while still learning temporal dependencies effectively. XGBoost enhanced its accuracy around holiday-driven

sales surges. Both LSTM + XGBoost and GRU + XGBoost proved to be highly competitive, especially for drug categories with volatile sales patterns.

5. Prophet, DLinear, and TBATS: Specialized Forecasting Approaches

Prophet, developed by Meta, was applied to decompose sales into trend, seasonality, and holiday components. While Prophet provided interpretable forecasts and captured holiday effects well, its accuracy was not as strong as boosting or hybrid models, particularly in highly non-linear categories. DLinear, a lightweight deep learning model that decomposes time series into trend and seasonal parts, was efficient and scalable, but it underperformed compared to more complex architectures like N-BEATS. TBATS, designed for complex and multiple seasonalities, performed well in categories like M01AB and N05C where overlapping seasonal patterns existed. However, it had higher error variance in volatile categories, making it less reliable overall.

6. Comparative Performance and Insights

When comparing the results across all models, several insights emerged. Statistical models like ARIMA and SARIMA were useful baselines but lacked flexibility in handling multiple seasonalities and non-linear patterns. Prophet, while interpretable, was not accurate enough for operational decision-making. Ensemble tree-based models like Random Forest, XGBoost, and LightGBM showed strong performance, with LightGBM excelling due to its computational efficiency and ability to handle complex features. Deep learning and hybrid models consistently outperformed standalone methods. N-BEATS + LightGBM and LSTM + XGBoost demonstrated significant advantages in balancing long-term trend capture with short-term fluctuation correction. GRU + XGBoost also provided strong results but was slightly less accurate than LSTM-based hybrids.

7. Best Model Recommendation

Based on the evaluation metrics such as MAE, RMSE, MAPE, and R^2, as well as stability across categories, the **N-BEATS + LightGBM hybrid** emerged as the best-performing model. It consistently provided high accuracy, handled irregular and volatile sales data effectively, and combined the interpretability of boosting with the trend-seasonality learning capacity of deep neural networks. While LightGBM alone was highly accurate, the hybrid with N-BEATS achieved marginal yet meaningful

improvements in error reduction and stability. The LSTM + XGBoost hybrid was also a strong contender, particularly for volatile series, but N-BEATS + LightGBM offered a more balanced solution across all categories

In this study, a wide range of forecasting models were evaluated on pharmaceutical sales data to determine their relative effectiveness in predicting future demand. The dataset, spanning six years and covering multiple drug categories, presented various challenges such as seasonality, stationarity, volatility, and holiday-driven fluctuations. The objective was to explore statistical, machine learning, deep learning, and hybrid approaches, compare their strengths and weaknesses, and ultimately identify the most suitable model for pharmaceutical sales forecasting. The models tested included Naïve Forecasting, Exponential Smoothing, ARIMA, SARIMA, CatBoost, XGBoost, Random Forest Regressor, LightGBM, N-BEATS + LightGBM hybrid, Prophet, DLinear, TBATS, LSTM + XGBoost hybrid, and GRU + XGBoost hybrid.

The starting point of the analysis was the Naïve Forecasting method, which served as a baseline. In this approach, the forecast for the next period is simply the last observed value. Although extremely simple to implement, it resulted in high errors and produced flat-line predictions, failing to account for trends, seasonality, or sudden changes in sales. Nevertheless, its role as a benchmark was important, as it allowed the performance of more advanced models to be compared against a basic baseline. Exponential Smoothing, another baseline statistical method, was then applied. By assigning higher weights to recent sales observations, this method improved slightly upon the Naïve Forecast. The Holt-Winters variant, which incorporates trend and seasonality, was able to capture recurring patterns such as flu-related seasonal peaks. However, these smoothing-based approaches struggled with abrupt demand spikes or irregular fluctuations, making them insufficient as standalone forecasting solutions for pharmaceutical data.

To further explore statistical approaches, ARIMA (AutoRegressive Integrated Moving Average) and its seasonal extension SARIMA were employed. ARIMA uses a combination of autoregressive terms, differencing for stationarity, and moving averages of past errors to generate forecasts. It provided stable short-term

predictions, particularly in drug categories with relatively smooth demand. However, ARIMA struggled in situations where seasonality was strong or where non-linear relationships dominated. SARIMA improved upon this by incorporating additional seasonal components, allowing it to model annual and monthly cycles. For example, SARIMA was able to reasonably capture recurring winter peaks in flu-related drugs. Despite these improvements, both ARIMA and SARIMA demonstrated limitations in handling multiple seasonalities or irregular demand shifts. Their accuracy lagged behind more advanced machine learning methods, but they served as strong statistical baselines.

The performance significantly improved with the use of ensemble tree-based machine learning models. CatBoost, a gradient boosting algorithm designed to handle categorical features efficiently, performed well across different drug categories. Its ability to handle categorical data without heavy preprocessing made it user-friendly. However, while CatBoost was robust, it did not always match the accuracy levels of other boosting methods, particularly in categories with highly volatile sales. XGBoost, another gradient boosting algorithm, stood out due to its ability to capture non-linear patterns and avoid overfitting through regularization. It achieved very strong performance in categories like N05C and M01AE, with RMSE values as low as 1–2, demonstrating excellent predictive power. Random Forest Regressor, an ensemble of multiple decision trees, also performed consistently well. By averaging predictions across many trees trained on random subsets of the data, Random Forest achieved high stability and robustness against overfitting. It was particularly effective for categories with low volatility, such as R06, but showed some limitations in capturing abrupt demand spikes. Among the ensemble methods, LightGBM emerged as one of the strongest contenders. Designed for speed and scalability, LightGBM used leaf-wise tree growth to efficiently model complex relationships. By leveraging lag features, rolling averages, and holiday effects, LightGBM produced low error metrics across most drug categories, achieving R^2 values above 0.84. Its accuracy, combined with computational efficiency, made it a strong candidate for large-scale forecasting applications.

The introduction of deep learning and hybrid models brought further improvements. N-BEATS, a neural network architecture specifically built for time-series

forecasting, excelled at learning long-term trends and seasonal patterns directly from raw data. When combined with LightGBM in a hybrid approach, the two models complemented each other effectively: N-BEATS captured smooth, long-term patterns, while LightGBM corrected short-term fluctuations and irregular spikes. This hybrid consistently outperformed single models, delivering highly stable and accurate forecasts across categories, especially in volatile drug series. Similarly, the LSTM + XGBoost hybrid model leveraged the memory capabilities of Long Short-Term Memory (LSTM) networks to capture temporal dependencies while using XGBoost to correct residuals. This combination significantly improved accuracy, reducing MAE by 5–10% compared to Prophet and ARIMA. The hybrid was especially useful in handling non-stationary data and holiday-driven demand fluctuations. The GRU + XGBoost hybrid followed the same logic but replaced LSTMs with Gated Recurrent Units (GRUs), which are computationally simpler and faster to train. GRU + XGBoost captured temporal patterns well and provided good accuracy, though it was slightly less effective than LSTM + XGBoost in highly volatile categories. Nonetheless, it offered a strong trade-off between efficiency and accuracy.

Specialized forecasting approaches were also considered. Prophet, developed by Meta, decomposed sales into trend, seasonality, and holiday effects, providing interpretable forecasts that were easy to communicate. Prophet was particularly good at incorporating holiday-driven effects, but it underperformed compared to boosting and hybrid models in terms of raw accuracy, especially in highly non-linear series. DLinear, a lightweight deep learning model, decomposed time series into trend and seasonal components using simple linear layers. While computationally efficient and scalable, it could not match the predictive power of more complex architectures like N-BEATS. TBATS, a state-space model designed for handling multiple and complex seasonalities, performed reasonably well in categories with overlapping seasonal cycles such as M01AB and N05C. However, it displayed high error variance in volatile drug categories, making it less reliable overall.

When comparing the overall performance of all models, clear trends emerged. Naïve Forecasting and Exponential Smoothing served as simple baselines but lacked practical accuracy. ARIMA and SARIMA were useful statistical benchmarks but

struggled in the face of strong seasonality and non-linearity. Prophet offered interpretability but lacked accuracy. Among ensemble models, Random Forest Regressor provided stable forecasts, CatBoost was reliable with categorical data, and XGBoost delivered very strong results. LightGBM, however, stood out as the best-performing single model due to its combination of high accuracy, scalability, and efficiency. The hybrid models, particularly N-BEATS + LightGBM and LSTM + XGBoost, demonstrated the highest performance. N-BEATS + LightGBM provided the most consistent improvements, excelling in both trend capture and residual correction. LSTM + XGBoost was especially strong in volatile and holiday-driven categories, while GRU + XGBoost offered faster training with slightly lower accuracy.

After a comprehensive analysis of metrics such as MAE, RMSE, MAPE, and R^2, the **best overall model was determined to be the N-BEATS + LightGBM hybrid**. This model consistently outperformed others in balancing accuracy, stability, and scalability. LightGBM alone was highly accurate and efficient, but the hybrid approach provided marginal yet meaningful improvements in error reduction, particularly for irregular sales categories. The LSTM + XGBoost hybrid was a strong runner-up, especially for categories with abrupt fluctuations, but it required more computational resources compared to the N-BEATS hybrid. Prophet and TBATS were valuable for interpretability and specialized seasonal modeling but did not achieve the same level of accuracy as boosting or hybrid models.

In conclusion, the results demonstrate that hybrid models combining deep learning and boosting methods are the most effective for pharmaceutical sales forecasting. The N-BEATS + LightGBM model emerged as the best choice, offering robust and accurate predictions that capture both long-term trends and short-term fluctuations. Its application can significantly enhance inventory management, reduce wastage, optimize supply chain planning, and support data-driven marketing strategies in the pharmaceutical sector. The findings emphasize that while traditional statistical methods and simpler models provide useful baselines, the complexity of pharmaceutical sales data demands advanced machine learning and hybrid solutions for practical, real-world forecasting needs.

CONCLUSION

The pharmaceutical industry operates in a highly dynamic environment where demand is influenced by seasonality, holidays, health outbreaks, and consumer behavior. Accurate forecasting of sales is therefore critical to ensure timely drug availability, minimize stockouts, reduce wastage, and optimize supply chain efficiency. In this study, a wide range of forecasting models was applied to six years of pharmacy sales data, including statistical methods, machine learning algorithms, deep learning architectures, and hybrid approaches. Baseline methods such as Naïve Forecasting and Exponential Smoothing provided useful benchmarks but lacked the ability to capture complex seasonalities and sudden fluctuations. Statistical models like ARIMA and SARIMA performed reasonably well in stable or seasonal datasets but struggled with non-linear patterns. Ensemble models such as Random Forest, CatBoost, and XGBoost demonstrated stronger performance, while LightGBM emerged as the most effective single algorithm due to its scalability and accuracy.

The most significant improvements, however, came from hybrid models that combined the strengths of deep learning and boosting methods. The N-BEATS + LightGBM hybrid consistently outperformed other approaches, delivering high accuracy across drug categories and effectively balancing long-term trend detection with short-term fluctuation correction. LSTM + XGBoost also provided excellent results, especially for volatile and holiday-driven demand, though it required higher computational resources. Prophet and TBATS were valuable for interpretability and complex seasonality, respectively, but did not match the predictive power of hybrid approaches.

Overall, the study concludes that the **N-BEATS + LightGBM hybrid** is the best-performing model for pharmaceutical sales forecasting. Its ability to capture irregular demand, integrate seasonal patterns, and maintain stable performance makes it the most reliable solution for real-world applications. By adopting such models, pharmaceutical companies and pharmacies can enhance operational planning, improve marketing strategies, and ultimately ensure better healthcare delivery through timely drug availability.

REFERENCE

[1] Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). *Time Series Analysis: Forecasting and Control*. Wiley.

[2] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.

[3] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of KDD*, 785–794.

[5] Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

[6] Oreshkin, B. N., et al. (2019). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *International Conference on Learning Representations (ICLR)*.

[7] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation, 9*(8), 1735–1780.

[8] Cho, K., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP*.

[9] Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician, 72*(1), 37–45.

[10] De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association, 106*(496), 1513–1527.

[11]