International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Pharmaceutical Sales Data Prediction Using Time Series Forecasting

Ankita M., Ananya Srinivas, Anurag Soni, Garima Prajapati & Dr. Manjunath P. S.

Submitted: 25/11/2023 **Revised**: 05/01/2024 **Accepted**: 15/01/2024

Abstract- The pharmaceutical world is often considered evergreen, holding timeless importance in today's world. It plays a major role in developing modern drugs and driving healthcare advancements. The pharmaceutical industry has come a long way, it now has the most advanced technologies and some of the world's leading scientists continually addressing global health challenges.

This booming industry mainly relies on efficient resource allocation, streamlined inventory management, cost-effectiveness, data-driven decision making and competitive strategies for its success. This paper discusses various methodologies in handling sales time series data within the pharmaceutical industry, aiming to enable data-informed decision making for industry professionals.

The objective aims at facilitating recommendation of sales and production of drugs by understanding trends and seasonality behind the data and accurately predicting its sales. The forecasting dashboard discussed later on in the paper, enables visualizing the sales and performance of a certain drug over a specified period of time. The research provides a simplified way to track and interpret drugs' sales data for better decision making.

Keywords: *LSTM, ARIMA, Dashboard, Time-series forecasting*

1. Introduction

Rising Trends in the Pharmaceutical Industry

The pharmaceutical industry plays a vital role in today's world. It significantly impacts the healthcare realm and plays a major part in improving the well-being of communities. The rapid advancements in the development of drugs and biotechnology has paved the way for groundbreaking innovations which aid the creation of advanced pharmaceuticals tailored to various medical needs.**Role of data science in empowering pharmaceutical sales**The integration of data science has become extensively indispensable in the pharmaceutical world, especially in the field of sales prediction. This synergy of data science and sales offers a multifaceted advantage to workforces related to this domain.

Making use of historical sales data and market trends, predictive models accurately predict future demand, this inturn provides an edge in inventory management. Pharmaceutical entities can align their supply chains, mitigate stock imbalances and ensure a steady production of medications by fine tuning forecasts. These algorithms furnish critical information by dissecting market dynamics, aiding in decision-making processes. These data-driven decisions help formulate tailored campaigns that can effectively help reach specific demographics or healthcare providers. These models also serve as tools for risk assessment and identifying potential shifts in market trends that could impact sales trajectories. Combining data science with sales forecasting only bolsters market responsiveness for better patient access to essential medications.

A Comprehensive Forecasting Engine for Pharmaceutical Sales Prediction

Dept. of Electronics & Telecommunication Engineering, BMS College of Engineering
E-mail : manjunathps.tce@bmsce.ac.in

The integration of time series forecasting with a specialized engine represents a vital advancement for industry professionals. This approach involves an examination of sales time series data in depth, aiming to unveil intricate trends and seasonal nuances that influence drug sales. By harnessing the power of these forecasting methodologies, this method constructs predictive models that draw from historical data, accurately predicting the sales of the drugs and empowering strategic decision-making.At the core of this methodology lies the development of a sophisticated forecasting engine made to decode the complexities that lie in the data provided. The engine harnesses a wide-array of advanced algorithms, meticulously identifying and extracting patterns to offer valuable foresight into future trends. Moreover, the development of an interactive forecasting dashboard acts as the central element of this framework. It serves as a visual aid, graphically depicting the performance of drugs over a specified time. This enables intuitive tracking and interpretation of sales dynamics. Through this predictive approach, industry leaders gain important insights which enables them to devise informed strategies regarding sales projections, resource allocation and market adoption. This approach seeks to empower stakeholders to foster more efficient and informed decision making strategies within the pharmaceutical landscape.

2. Related Work

[1] The research paper revolves around making use of multiple models, including ARIMA and LSTM for sales forecasting. It involves analyzing the time series data to evaluate the effectiveness of each model. The paper reiterates constructing and testing the ARIMA model, revealing that the ARIMA-LSTM combination displayed the best forecasting accuracy among others. Moreover, the study also includes building and training the LSTM model and also creating and training the ARIMA-LSTM hybrid model. In order to assess the models, various statistical analyses are presented, assessing how well the model predicts the sales data. Ultimately, the research concludes the superiority of the combined ARIMA-LSTM model over individual models. The paper also discusses the implications and applications of these results. Towards the ends, it identifies limitations and suggests scope of improvement. It sheds light on the effectiveness of various forecasting models and the advantages of combining different models for better prediction.

[2] The paper delves into the important stage of performance estimating in machine learning, specifically focusing on its application to time series forecasting. It sheds light on the challenges faced while assessing model performance in time series data due to temporal dependencies, requiring a comparison of various estimation methods. This includes three categories: out-of-sample, prequential, and cross validation. The study conducts comprehensive empirical analyses on real-world and synthetic time series data, revealing unique performancetrends based on data stationarity. According to the study, cross-validation suits stationary series, non-stationary data benefits more from the holdout method. The paper extends a previous work, broadening the experimental scope and discussing stationarity impacts. [3] In this paper, automated machine learning in the context of deep learning applications is extensively explored. The paper begins by shedding light on deep learning's accomplishments in image recognition, object detection, and language modeling but underscores the tedious manual effort required to create high-performing models. AutoML emerges as a solution to streamline this process, automating multiple ML stages, specifically focusing on neural architecture search within deep learning. It categorizes the AutoML pipeline into data preparation, feature extraction, hyperparameter tuning and NAS, discussing methods within each phase. It emphasizes NAS methods' performance on datasets like CIFAR-10 and ImageNet, discussing subtopics like one-shot NAS and joint optimization methods. The paper concludes by highlighting the existing challenges in the realm of AutoML and suggesting areas for future research to advance this field. [4] This study puts forward a pioneering approach to forecast pharmaceutical distribution companies' (PDCs) sales by combining network analysis and time series forecasting. The method takes advantage of cliques in a network of medicine and co-members' sales data for prediction due to paucity of data. The study shows that incorporating a drug;s historical data and those of its related group significantly improves the prediction of sales. The network based analysis identifies drugs with similar sales trends which enables more accurate forecasts despite the lack of data. Employing a hybrid approach that combines linear and non-linear models outperforms traditional linear models and standalone neural networks. This exhibits this method's efficacy in optimizing sales forecasts for PDC's. [5] The research focused on AutoML tools' usage in Machine learning pipelines. It initially analyzes the most used AutoML tools via Github projects, and the top 10 most used tools are revealed. The tools are mainly used for hyperparameter tuning, model training, feature engineering and model management. Furthermore, the study discusses how these tools are used together. The paper also highlights that most of the projects employ a single AutoML tool, with very few projects combining multiple tools like Hyperopt

with TPOT or Optuna. This analysis helps ML practitioners seek suitable tools and developers aiming to enhance their integration capabilities. [6] The article revolves around the complexities of predicting economic and financial time series data. It also talks about the challenges arising from evolving trends and insufficient data. The paper scrutinizes traditional approaches like ARIMA, highlighting its limitations in capturing complex relations within non-stationary data. In response to this analysis, the paper analyzes LSTM, a deep learning model, showcasing its capacity to handle huge data and its ability to capture intricate patterns and nonlinearities within time series forecasting. Both the models, ARIMA and LSTM are compared and it is noticed that LSTM outperforms ARIMA significantly. It showed an average error rate reduction of 84-87%. The paper also brings to light that iteration after a certain number of epochs, doesn't make the model more accurate. The paper emphasizes on LSTM's superiority and reiterates its potential for revolutionizing forecasting in many domains especially the economic and finance sectors. [7] The research delves into how technology, specifically in finance after COVID-19, is becoming more automated. It discusses the importance of predicting stock prices for investing and planning out business strategies. The study compares two models, ARIMA, a statistical model and LSTM, a deep learning model, to forecast NASDAQ stock exchange data. In order to avoid making any model too specific, monthly and daily average stock prices from different sectors were used. ARIMA makes predictions using historical data whereas LSTM looks at long-term patterns using gates. The results favores ARIMA in predicting stock prices over various specified time frames except for very-short term predictions. The paper also discusses the need to use logarithms data for both models and suggests improving LSTM by adding more data and analyzing trends. The paper concludes that a mix of these models could improve prediction of data.

3. Prerequisites

I. Python: Python serves as the primary language for this project. It acts as a bridge between theory and practice. It is not just a tool but a cornerstone and requires a profound command. This entails data manipulation using libraries like Pandas and NumPy, mastering the art of control structures including conditional statements and loops and the capability to create and manage functions efficiently. The importance of object-oriented-programming (OOP) in python comes into play when translating abstract concepts into concrete algorithms. Python is an ideal programming language for implementing and experimenting with deep learning models due to its versatility and readability.

II. TensorFlow: Tensorflow being a pervasive deep learning framework, assumes a central role in the project. Proficiency is not an option but a prerequisite in order to harness its powers effectively. The proficiency extends to core concepts such as tensors, the elemental data structures that underpin computations, and the intricacies of computation graphs, they dictate the organization of operations. Beyond the foundational knowledge, the construction of deep learning models using TensorFlow is pivotal. This entails designing neural network architectures, creating custom layers, and configuring the intricate web of interconnected components that drive model behavior. The training process, with all its nuances and intricacies, demands a comprehensive understanding of the concept.

III. Keras: Keras is a user-friendly deep learning framework. It acts as a valuable ally in this project and offers a streamlined path from theory to practice. Due to its pythonic syntax and modular design, it simplifies the development of neural networks reducing the gap between idea and implementation. With its pre-trained models, Keras accelerates research and application development, enabling users to leverage existing knowledge and expertise. This framework caters to both industrial and academic researchers. It offers a smooth and productive research experience.

IV. Time Series Forecasting: Time series forecasting involves various techniques to predict future data points. It revolves around leveraging fundamental methods like Average, Naïve, and Seasonal Naïve, these serve as benchmarks in this process. The Average method predicts future data by averaging historical data, often making use of the mean of the training set. Naive forecasting assumes that the next value will be similar to that of the recent observation, suitable for short-term predictions. Seasonal Naïve is useful when the data exhibits seasonality. It helps calculate forecasts by averaging values from relevant times in past seasonal cycles. Seasonal Decomposition involves building models on data with the residual component removed.

V. Stationary and Non-Stationary Data: In the time series analysis, to understand the difference between stationary and non-stationary data is important. Where stationary data is characterized by a consistent statistical property over time, which means that factors like its mean, variance, and autocorrelation structure remain constant. This simplifies the modeling process and thus helps in accurate forecasting. Non-stationary data consists of spatial time variations in statistical properties, which displays the seasonality. Such variations often complicate the model and make predictions less reliable. Transformation techniques, such as differencing, are commonly employed to convert non-stationary data into a stationary form, to optimize the effectiveness of time series models

VI. Seasonality: Seasonality in time series data refers to the recurring patterns that follow a regular and predictable trend within a specific time frame, such as daily, weekly, or yearly cycles. These patterns produce outcomes as periodic increases or decreases in the data, where external factors like weather, holidays, or economic cycles contribute to the predictions and analysis.. Identifying seasonality is crucial for accurate forecasting, as models account for these regular variations. Time series decomposition techniques, such as additive or multiplicative decomposition, help isolate and analyze seasonality components, which helps in more effective modeling.

VII. ARIMA: ARIMA is short for Auto-Regressive Integrated Moving Average which are widely used for forecasting univariate stationary time-series. It makes use of the interdependence between an observation and its lagged counterparts (AR), incorporates differencing (I) to achieve stationarity, and accounts for residual errors from a moving average model (MA). The model’s hyperparameters include, ‘p,d,q’ values that stand for, lad order, differencing degree and moving average order respectively. Identifying these parameters involve analyzing the PACG and ACF plots. The point at which the PACF cuts off is considered at the p value, while q value is identified based on ACF cutoff. [1] and [7] discuss the functioning of the ARIMA model in depth.

VIII. LSTM: LSTM stands for Long-Short-Term Memory and is a type of recurrent neural network(RNN). It is known for handling sequences and time series data because of its ability to retain information over long periods. LSTM has a unique architecture with specialized memory cells and gates that regulate the flow of information. These gates enable LSTMs to selectively retain, forget and update information, making it proficient in capturing temporal dependencies and handling vanishing or exploding gradient problems often encountered in RNNs. LSTM models excel in modeling complex sequential patterns, making them an ideal tool for several applications like time-series forecasting, speech recognition and natural language processing. LSTM is discussed in greater detail in [1] and [7]

IX. Flask: a Python microframework, powers our project's web functions efficiently. Handling routing, it directs requests to designated functions, managing diverse HTTP methods. Flask seamlessly integrates predictive models like ARIMA and LSTM through distinct endpoints like '/predict', delivering data insights. Utilizing Jinja2 templates, it creates dynamic HTML content for an intuitive user interface. Its adaptability allows easy integration with extensions for added features like form validation and authentication. Flask's deployment on platforms like Heroku ensures accessibility, making it a robust framework for our web-based solution.

X. Dashboard and Forecasting: The dashboard acts as an interface that provides a clear visualization of the predictions of analysis of each methodology. Through this dynamic dashboard, it becomes easy to navigate and comprehend the comparative performance of Machine-learning models in real-time. The integration of automation and visualization not only helps in the model selection process but also optimizes the process of interpretability and efficiency of the forecasting outcomes. In the time series data, forecasting involves prediction of future values based on the patterns and trends observed in historical data. [2] and [6] delve into various forecasting methods. Time series forecasting methods consider the dependencies present in the data and the goal here is to analyze these patterns, train the model which in turn helps in prediction of future outcomes. Forecasting engine is the core of these predictions as it utilizes the computation f algorithms.

4. Proposed Methodology

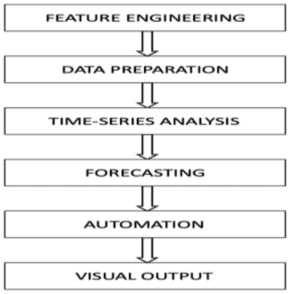


Fig 1: Block diagram for implementation

i. Feature Engineering : Feature engineering is the most common way of changing and making new features from raw data to improve the performance of machine learning models. It includes choosing, changing, or creating features that capture relevant information, upgrade model interpretability, and add to better predictive outcomes.A structured feature engineering methodology was employed to improve the dataset's illustrative power. This included creating intermediary time series, collecting sales within weekly time spans across various classes of drug items, including anti-inflammatory, analgesics, psycholeptics, drugs for obstructive airway diseases, and antihistamines.

ii. Data Preparation: Involves cleaning and transforming raw data into a format suitable for analysis or modeling, addressing issues such as missing values, outliers, and ensuring data consistency. It includes steps like normalization, scaling, and handling categorical variables to prepare the data for effective use in machine learning algorithms.We started by defining the problem and objectives related to pharmaceutical sales, utilizing information from the sales dataset. We found the dataset on kaggle platform, details of which will be explained later. The obtained information went through fastidious cleaning to address missing qualities and abnormalities. Maintaining proper temporal order, anomalies identified in consultation with pharmacy staff were rectified. The selected group of drugs, consisting of 57 drugs, was classified into 8 Anatomical Therapeutic Chemical (ATC) Classification System categories. After that, the dataset was turned into hourly time series followed by weekly, which separated pharmaceutical products into distinct classes.

iii. Time series analysis: Time series analysis is the exploration and statistical examination of sequential data points over time to uncover patterns, trends, and dependencies within the temporal structure of the dataset. Here, we sought after a double goal. It, first and foremost, involved a comprehensive examination of annual, weekly, and daily sales data to extract insights for refining sales and marketing strategies. Secondly, statistical analyses, encompassing stationarity, autocorrelation, and predictability assessments, were conducted on individual drug categories. These analyses provided the underlying boundaries needed for subsequent forecasting methods.

iv. Forecasting : Forecasting is the process of estimating future trends, values, or outcomes based on historical data and statistical models, providing insights for decision-making and planning. Sales forecasting was executed at a week by week scale, taking on two unmistakable methodologies. The rolling forecast method involved predicting sales for the next week using a model trained on all historical data. This model proved valuable for short-term resource planning and stock procurement. Additionally, a long-term forecasting approach aimed at predicting sales for an entire year using a model trained on historical data. This approach was geared towards strategic decision-making and business planning.The forecasting models evaluated included ARIMA/SARIMA and Long-Short Term Memory (LSTM) neural network architectures. Hyper-parameters were optimized using manual analysis, Python's statsmodels function, and grid search optimization. Model execution was assessed utilizing a train-test split approval, with the last year of information held for testing. The primary performance metric was Mean Squared Error (MSE), with Mean Absolute Percentage Error (MAPE) serving as an illustration.

v. Automation : Automating the selection of the optimal machine learning model is achieved by evaluating performance metrics, specifically Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). In this research study, we employed two distinct forecasting methodologies, ARIMA and LSTM, and implemented an automated process to determine the optimal model. This automation played a pivotal role in systematically comparing the forecasting results generated by ARIMA and LSTM, streamlining the model selection process based on key evaluation metrics. The results and insights obtained from both methodologies were meticulously analyzed, leading to the identification of the best-performing mode.

vi. Visual output (Dashboard) : The visual output or dashboard developed, users can choose from a dropdown menu highlighting eight distinct drug classes (derived from the classification of fifty seven drugs into eight groups). One more dropdown permits clients to pick either ARIMA and LSTM, with an extra choice to choose the AI model with the least error. The result of the dashboard is a powerfully produced predicted graph based on the chosen parameters, providing a visual representation of the forecasted pharmaceutical sales.

5. Implementation

This section of the research paper discusses in detail the functioning of the project. It will cover the following topics:

1) About dataset

2) Exploratory data analysis

3) Forecasting

4) Automation and Dashboard

1) About Dataset:

The initial dataset (obtained from kaggle) comprised 600,000 transactional records spanning six years (2014-2019), featuring date and time of sale, pharmaceutical drug brand names, and sold quantities. Following discussions with pharmacists, the analytical focus and forecasting shifted from individual drugs to broader drug categories. Consequently, a selected set of 57 drugs was categorized into eight Anatomical Therapeutic Chemical (ATC) Classification System categories, each representing a distinct therapeutic class. These categories include :

1) M01AB for anti-inflammatory and antirheumatic products, non-steroids, Acetic acid derivatives.

2) M01AE for anti-inflammatory and antirheumatic products, non-steroids, Propionic acid derivatives.

3) N02BA for other analgesics and antipyretics, Salicylic acid and derivatives.

4) N02BE/B for analgesics and antipyretics, Pyrazolones, and Anilides.

5) N05B for psycholeptics drugs, Anxiolytic drugs.

6) N05C for psycholeptics drugs, Hypnotics and sedatives drugs.

7) R03 for drugs for obstructive airway diseases.

8) R06 for antihistamines for systemic use.

These ATC codes were integrated as features in the dataset, transforming the model structure, and the data were subsequently resampled into hourly time-series followed by weekly and cleaned.

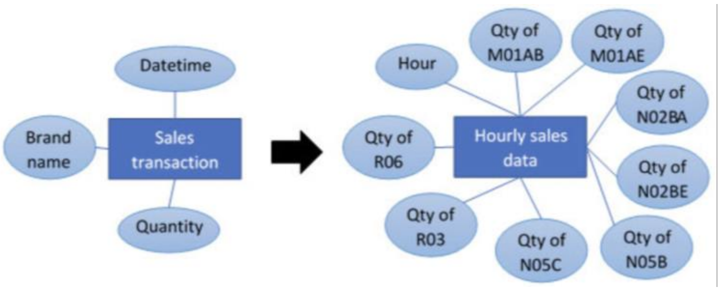


Fig 2: Dataset representation

Outliers were detected through consultation with pharmacy staff. The treatment of outliers followed a two-step process. First, missing data was imputed using appropriate methods. Subsequently, representative data was imputed, employing various methods to ensure a comprehensive and accurate representation of the pharmaceutical sales dataset. Post-cleaning, the data was transformed into hourly time series followed by weekly time series, categorizing pharmaceutical products into eight specific classes. This intermediary time series was then rescaled to a weekly time-series format, facilitating a more granular and manageable temporal resolution. The rescaled data was systematically stored for subsequent analyses.

2) Exploratory data analysis:

Time Series Analysis or Exploratory data analysis included seasonality, stationarity, autocorrelation, regularity and data distribution analysis. It consists of parts:

I) Seasonality analysis

II) Stationarity analysis

III) Regularity analysis

IV) Autocorrelation and Partial autocorrelation analysis

Seasonality Analysis:

It is the process of identifying and quantifying repetitive patterns or fluctuations in a time series that occur at regular intervals, often associated with certain times of the day, week, month, or year. In the context of time series forecasting, seasonality refers to systematic, calendar-related variations in the data that can impact predictive accuracy. For the purpose of this research, seasonality analysis helps us to analyze:

1) Temporal Patterns: Helps recognize recurring patterns and trends within the pharmaceutical sales data. Understanding these temporal variations is crucial for accurate forecasting.

2) Pattern Recognition: Allows the model to capture and incorporate cyclic behaviors related to specific times, such as increased sales during flu seasons or certain times of the year when demand for certain drugs may be higher.

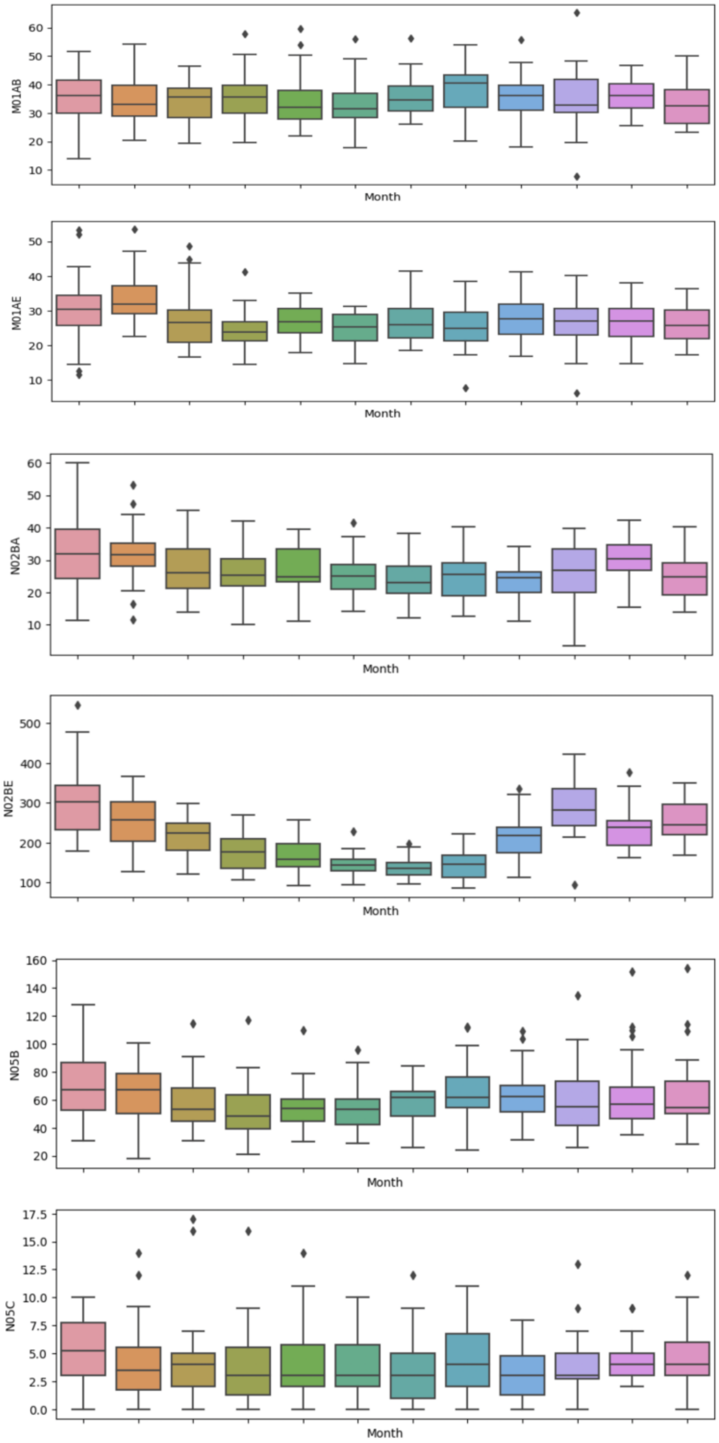
3) Model Adjustments: The forecasting model can adjust its predictions based on historical trends during specific periods. This adjustment enhances the accuracy of predictions, especially when dealing with time-dependent phenomena.

4) Decision Support: Provide valuable information for strategic decision-making, allowing stakeholders to

proactively address fluctuations in demand, plan inventory, and optimize resource allocation.

In this paper, detailed exploration of seasonality patterns is conducted through the utilization of boxplots. Boxplots visually represent the distribution of a dataset, displaying the median, quartiles, and potential outliers. The central box encapsulates the interquartile range (IQR), providing a concise summary of the data's spread and central tendency. The presence of seasonality is notably evident in the categories of R03, R06, and N02BE. Further observations reveal that R03 and N05C exhibit a higher number of outliers compared to other categories, suggesting that predicting sales for these categories is more challenging.

The provided boxplots depict each category on the y-axis and months on the x-axis, with each box corresponding to the months of the year in sequential order. This visual representation effectively illustrates seasonality, enabling a clear visualization of monthly patterns within each category.



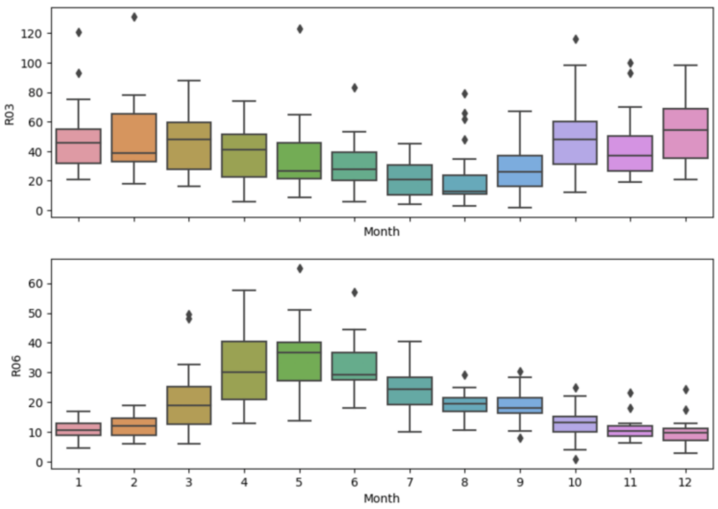


Fig 3: Boxplot representation per category

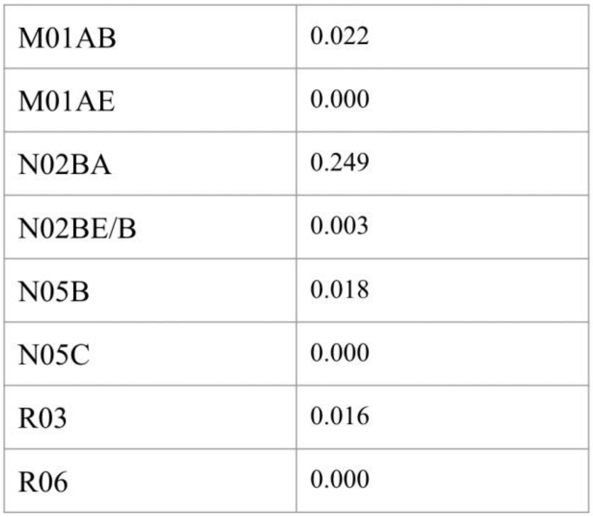
Stationarity analysis:

Stationarity in a time series refers to the property of maintaining consistent statistical characteristics over time, such as mean, variance, and autocorrelation. This can be visually assessed through rolling statistics like means and variances. In a stationary time series, the mean, variance, and covariance between the i-th term and the (i + m)-th term should not be time-dependent.

i) The Augmented Dickey-Fuller (ADF) test, with parameters including 'c' for constant only (default), 'ct' for constant and trend, 'ctt' for constant, and linear and quadratic trend, and 'nc' for no constant and no trend, can be employed to verify the stationarity of the data.

The Augmented Dickey-Fuller (ADF) test is a statistical method used to assess the stationarity of a time series. It evaluates whether a unit root is present in the data, indicating the non-stationary nature of the series. The test compares the null hypothesis that the data possesses a unit root against the alternative hypothesis that the data is stationary after differencing. Different versions of the test exist, allowing for the inclusion of trends or drifts in the model. The test outputs a p-value, and if this value is below a chosen significance level, the null hypothesis of non-stationarity is rejected, indicating that the time series is likely stationary.

The output of ADF test is as follows:

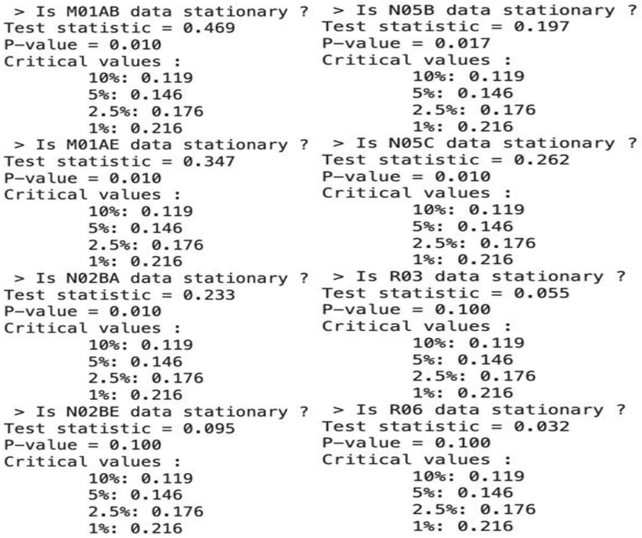


M01AB	0.022
M01AE	0.000
N02BA	0.249
N02BE/B	0.003
N05B	0.018
N05C	0.000
R03	0.016
R06	0.000

Table 1: Categorical p-values of ADF test

The results of the Augmented Dickey-Fuller (ADF) test indicate that all series in the dataset, except for N02BA (P-value=0.249), demonstrated stationarity with high confidence.

ii) The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is a statistical method used to assess the stationarity of a time series. Unlike the Augmented Dickey-Fuller (ADF) test that tests for the presence of a unit root, the KPSS test examines whether the data is trend-stationary, meaning it has a constant mean and variance over time, or difference-stationary, indicating stationarity after differencing. The test has null and alternative hypotheses, and



```
 > Is M01AB data stationary ?   > Is N05B data stationary ?
Test statistic = 0.469         Test statistic = 0.197
P-value = 0.010                P-value = 0.017
Critical values :              Critical values :
        10%: 0.119                     10%: 0.119
        5%: 0.146                      5%: 0.146
        2.5%: 0.176                    2.5%: 0.176
        1%: 0.216                      1%: 0.216
 > Is M01AE data stationary ?   > Is N05C data stationary ?
Test statistic = 0.347         Test statistic = 0.262
P-value = 0.010                P-value = 0.010
Critical values :              Critical values :
        10%: 0.119                     10%: 0.119
        5%: 0.146                      5%: 0.146
        2.5%: 0.176                    2.5%: 0.176
        1%: 0.216                      1%: 0.216
 > Is N02BA data stationary ?   > Is R03 data stationary ?
Test statistic = 0.233         Test statistic = 0.055
P-value = 0.010                P-value = 0.100
Critical values :              Critical values :
        10%: 0.119                     10%: 0.119
        5%: 0.146                      5%: 0.146
        2.5%: 0.176                    2.5%: 0.176
        1%: 0.216                      1%: 0.216
 > Is N02BE data stationary ?   > Is R06 data stationary ?
Test statistic = 0.095         Test statistic = 0.032
P-value = 0.100                P-value = 0.100
Critical values :              Critical values :
        10%: 0.119                     10%: 0.119
        5%: 0.146                      5%: 0.146
        2.5%: 0.176                    2.5%: 0.176
        1%: 0.216                      1%: 0.216
```

Fig 1: Output of KPSS

The results of the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test indicate non-stationarity in the trend for N02BE, R03, and R06.

Regularity analysis:

It refers to the extent to which a pattern or behavior follows a consistent and predictable structure over time. In the context of time series data, regularity implies a certain level of order or repetitiveness in the observed patterns. A time series is considered more regular when there is a systematic and predictable relationship between data points, allowing for the identification of consistent trends or cycles. Regularity is often assessed to understand the stability and predictability of a time-dependent phenomenon. For this research, time series were assessed using the Approximate Entropy test. The Approximate Entropy (ApEn) test quantifies the irregularity or complexity of a time series by assessing the likelihood that similar patterns persist as more data points are added. It measures the conditional probability that sequences of close observations remain close, providing insights into the regularity or predictability of the time series.

The output of Approximate Entropy (ApEn) test is as follows:

M01AB	1.141130089570642
M01AE	1.166363924596575
N02BA	1.1370638730125302
N02BE/B	1.058024809082593
N05B	1.074437415034502
N05C	1.0361887401424648
R03	1.1847216239035152
R06	1.031759595747876

Table 2 : Categorical entropy values

Entropy values exceeding 1 were observed for all series, suggesting low predictability. The highest entropy values were recorded for M01AE, M01AB, and N02BA categories.

Autocorrelation and Partial autocorrelation analysis:

It assesses the strength and patterns of relationships between observations at different time points in a time series. Analyzing autocorrelation provides insights into the predictive potential of time series data. Autocorrelation plots, visualized through the AutoCorrelation Function (ACF), graphically represent the strength of relationships between observations at different time steps. The plot displays lag values on the x-axis and correlations on the y-axis, with confidence intervals drawn to identify statistically significant correlations. The default setting is a 95% confidence interval, indicating significant correlations beyond this range. This analysis assumes a normal Gaussian distribution of the data.

Partial autocorrelation (PACF) measures the unique correlation between two variables, removing the influence of a specified set of other variables. In the context of regression, if we are regressing a variable Y on variables X1, X2, and X3, the partial correlation between Y and X3 isolates the correlation between Y and X3 not explained by their common correlations with X1 and X2.

The plot obtained from ACF is as follows:

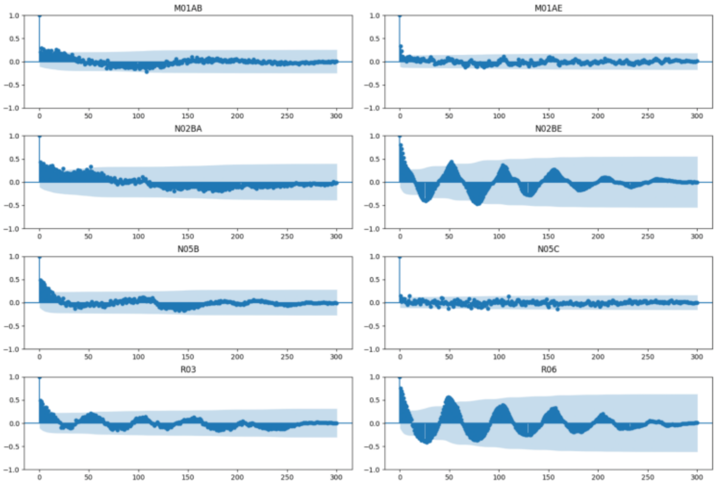


Fig 3: Categorical ACF plots

The plot obtained from PACF is as follows:

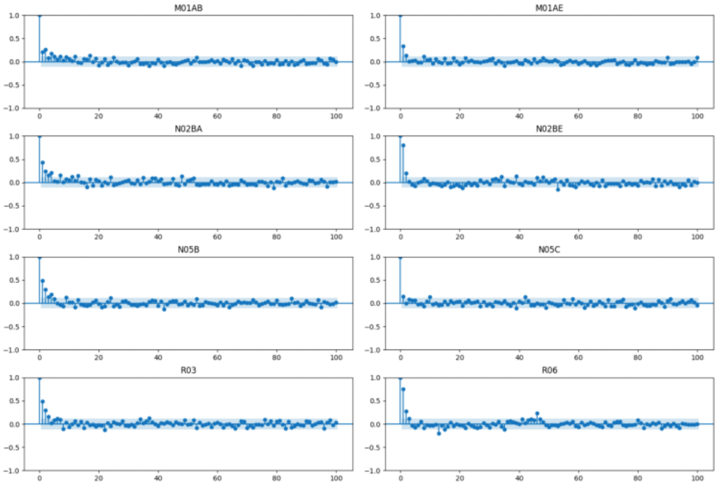


Fig 3: Categorical PACF plots

Slight autocorrelation is evident in Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots for all series, except for N05C sales. Notably, the N02BE, R03, and R06 series display discernible annual seasonality patterns.

Hence, time series analysis, including ACF and PACF plots, stationarity tests, Approximate Entropy, and partial correlation, is crucial for refining forecasting models and achieving accurate pharmaceutical sales predictions.

3) Forecasting:

In this research study, we employed two distinct forecasting methodologies, namely ARIMA and LSTM, and implemented an automated process to determine the optimal model. The automation was integral in selecting the best-performing model by systematically comparing the forecasting results generated by ARIMA and LSTM. Leveraging this automated approach, the research aimed to streamline the model selection process based on key evaluation metrics as we discuss below. The results and insights obtained from both ARIMA and LSTM were meticulously analyzed, and the chosen model was subsequently showcased in a comprehensive dashboard. This dashboard serves as a visual representation of the selected forecasting model, providing a user-friendly interface for users to comprehend and interpret the predictive capabilities of each methodology. The integration of automation and visualization not only expedites the model selection process but also enhances the accessibility and transparency of the forecasting outcomes.

I) Arima forecasting:

ARIMA method was chosen for short-term rolling forecast and long-term forecast. To carry out the forecast, we first go through a process of computing the hyperparameter optimization for the ARIMA model that consists of p,q and d as the parameters.

After calculating the optimal sets of parameters, both short-term and long-term forecasting was implemented. These parameters are computed since it is important and very helpful in defining the structure of ARIMA's model.The goal is to create a model that effectively predicts future values based on historical observations.

To choose parameters for the ARIMA model, the arma_order_select_ic method was utilized. This method calculates Akaike's Information Criterion (AIC) for numerous ARIMA models and selects the best configuration.

To test for the accuracy we have used, mean squared error. Hence. MSE becomes the criteria for optimization. In this method, we try several different combinations of hyperparameters and then compare the combination for the lowest MSE which becomes the optimal value.For the rolling forecast, the grid search optimization identified the following best combinations of hyperparameters,Grid search optimization is a systematic approach to fine-tune model parameters,helping to improve forecasting accuracy by selecting the most effective combination of hyperparameters.

The grid search optimization involves an exhaustive search through a predefined set of hyperparameter values to find the combination that minimizes a specified evaluation metric, such as Mean Squared Error (MSE).The mathematical formulation for grid search optimization can be expressed as follows:

Let H represent the set of hyperparameter combinations to be tested, where

H={(p1,q1,d1),(p2,q2,d2),...,(pn,qn,dn)}

Each tuple (p,d,qi) represents a candidate set of hyperparameters.

Define a function MSE(p,d,q) that calculates the Mean Squared Error for the given hyperparameters.Then apply grid search optimization algorithm, After the grid search is complete, optimal hyperparameters will contain the combination that minimizes the MSE over the specified set of hyperparameters.

Grid search optimization was carried out for both short-term and long-term forecasts to get the best hyperparameters. From an optimization perspective, certain time series are categorized as "white noise," representing random and uncorrelated data.

These are series for which the optimal modeling configuration is achieved with

p=0, d=0, and q=0 in the context of an ARIMA model.

The absence of non-zero values for these parameters implies that the best forecasting results are obtained without relying on past observations, differencing, or considering the influence of past forecast errors. Identifying series characterized as white noise is valuable, suggesting that traditional time series patterns may be minimal, and randomness dominates the data, influencing the optimal configuration for the forecasting model.

The below fig shows the architecture of ARIMA model:

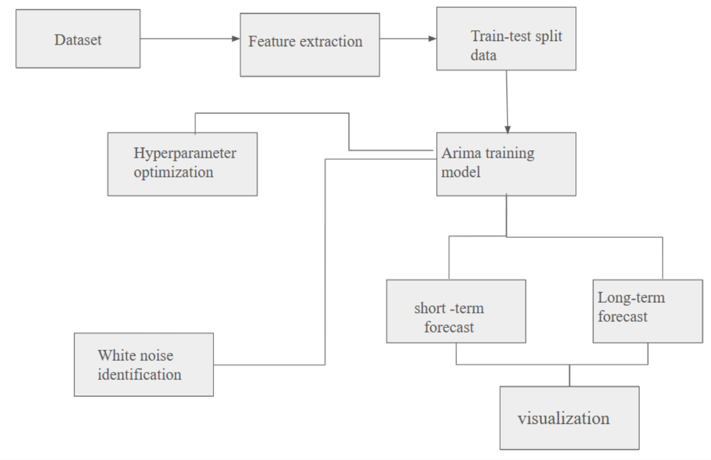


Fig 4: ARIMA Architecture

II) LSTM Forecasting

The study focuses on long-term forecasting using three distinct configurations of Long Short-Term Memory (LSTM) neural networks: Vanilla LSTM, Stacked LSTM, and Bidirectional LSTMas discussed in the previous sections. This research explores the impact of vanilla LSTM architectures on forecasting accuracy, employing standardized and supervised time-series data. The activation function is Rectified Linear Unit (ReLU), the optimizer is Adam, and the loss function is Mean Squared Error (MSE).

In the data preprocessing phase, standardization and transformation were applied to ensure uniformity across the dataset. The data was rescaled to fit within the interval [-1, 1], facilitating consistent and comparable inputs for subsequent modeling. Additionally, the split_sequence function was employed to structure the data into a supervised machine learning format, enabling the creation of input-output pairs crucial for training the LSTM models. The time series data was further processed in previous sections, where the split_sequence function was utilized to generate input-output pairs for model training. Notably, two configurations were explored for the number of past observations (n_steps): a value of 10 for series characterized by high variances and randomness (specifically, N05B and N05C), and a value of 5 for all other cases. This parameter selection was informed by its discernible impact on forecasting accuracy.

Moving to the model configuration phase, vanilla LSTM architectures were considered. The choice of architecture played a pivotal role in determining the network's proficiency in capturing long-term dependencies within the time-series data. As we know, the use of the ReLU activation function and the Adam optimizer, renowned for their effectiveness in diverse deep learning applications. Additionally, all models underwent training for 400 epochs, a value determined through careful experimentation to achieve optimal forecasting results. To ensure the reproducibility of results, a meticulous approach was taken. Random seeds for Python, NumPy, and TensorFlow were fixed, maintaining consistency across different runs. The 'PYTHONHASHSEED' environment variable was set, and a new global TensorFlow session was configured with a fixed seed, collectively providing a robust framework for reproducible and optimal forecasting experiments. Model implementation involved utilizing the Keras library to construct LSTM models. Each model was designed with one or more LSTM layers, followed by Dense layers, embodying the chosen architectural configurations. In the experimentation and evaluation phase, models were trained on a dedicated training dataset, validated on a separate validation set, and ultimately evaluated on an independent test set. Mean Squared Error (MSE) was used as the primary evaluation metric which is the quantitative method of computing forecasting accuracy.

In the later section, we deep dive into the results, where the forecasting accuracy and performance of each LSTM configuration were rigorously scrutinized using the MSE metric. A comparative analysis was conducted to discern the strengths and weaknesses inherent in each architecture of the LSTM model.

4) Automation and Dashboard:

The aim is to automate the selection of the optimal machine learning model for drug category forecasting between ARIMA and LSTM based on accuracy as the primary metric. Additionally, a user-friendly dashboard was created to facilitate seamless selection of drug categories and machine learning models.

The study mainly focuses on the machine learning model automation process (Section 4). Both ARIMA and LSTM models are systematically trained and evaluated on the dataset, utilizing accuracy as the primary metric for performance assessment. The automation script, in Section 4.2, navigates through the various drug categories, conducting training and evaluation for ARIMA and LSTM models. The optimal model for each category is then determined based on the highest accuracy, computed using a well-defined mathematical formula where accuracy is the ratio of the number of correct predictions to the total number of predictions, multiplied by 100.

The script then records the optimal model for each drug category, and this automated process is seamlessly integrated into the interactive dashboard. Users can effortlessly trigger the automation from the dashboard, initiating the model selection process based on their chosen drug category. This sophisticated automation framework ensures efficiency and consistency in model selection, enhancing the practical applicability of the forecasting system.

6. Result

The outcomes of the research shed light on the effectiveness of time-series analyses and forecasts for pharmacy sales. Notably, the ARIMA method, particularly when supplemented with Auto-ARIMA for series displaying seasonal characteristics, demonstrated superior performance in rolling forecasts compared to Prophet as discussed in the previous sections, emerging as the preferred choice for short-term sales predictions. Our analyses of daily, weekly, and annual seasonality patterns proved valuable for pinpointing optimal periods for implementing special sales and marketing campaigns.

Let's take a look at ARIMA forecasting:

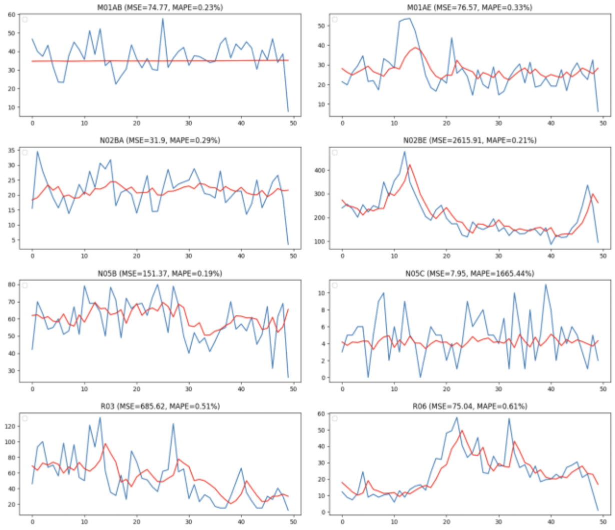


Fig 5: Short-term forecasting

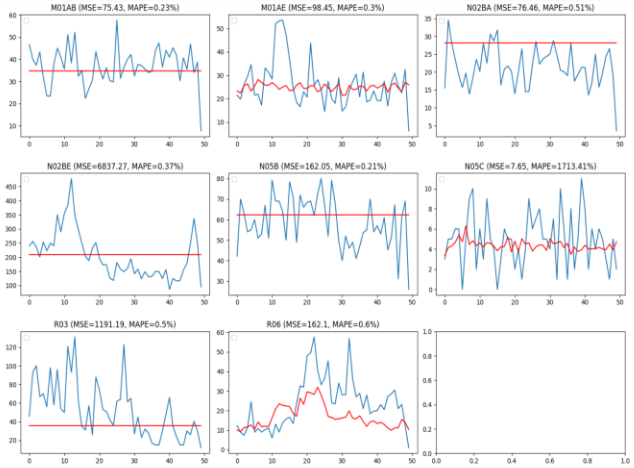


Fig 6: Long-term forecasting

The LSTM produced the results as follows:

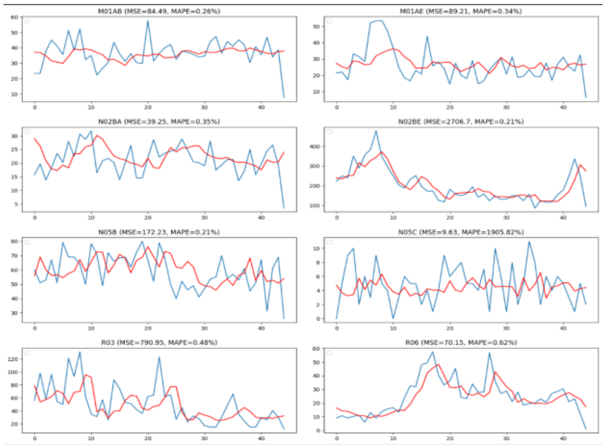
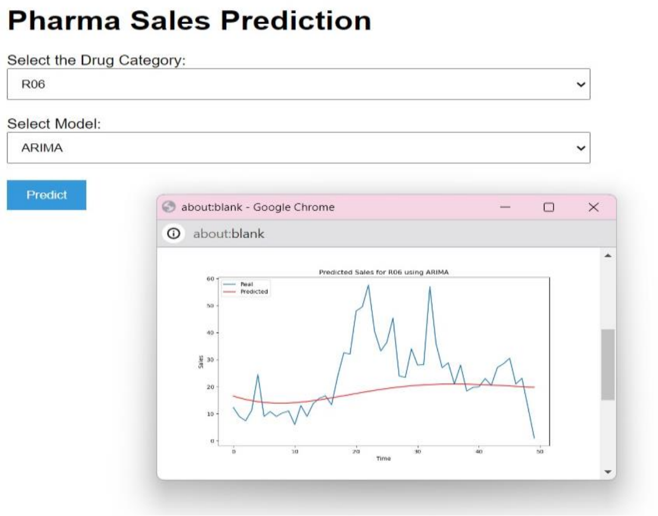


Fig 7: LSTM Forecasting

Overall, it contributes to the evolution of time-series forecasting methodologies for practical use.

The outcomes generated by the dashboard are:



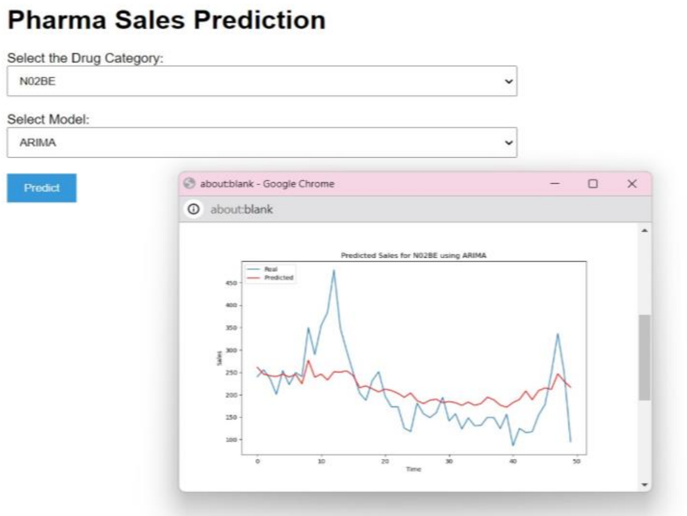
Pharma Sales Prediction

Select the Drug Category: R06

Select Model: ARIMA

Predict
about:blank - Google Chrome
about:blank
Predicted Sales for R06 using ARIMA
Real
Predicted
Sales
Time

Fig 8: Dashboard - R06 prediction



Pharma Sales Prediction

Select the Drug Category: N02BE

Select Model: ARIMA

Predict
about:blank - Google Chrome
about:blank
Predicted Sales for N02BE using ARIMA
Real
Predicted
Sales
Time

Fig 9: Dashboard - N02BE prediction

7. Applications

Sales prediction model has significant practical implications for businesses across various industries.The application of this model can provide businesses with a competitive edge by enabling them to make more informed decisions about production schedules, marketing strategies, and pricing, ultimately leading to increased profitability and growth. Some of its currently used applications are:

i. Supply Chain Planning:

a. Production Planning: Accurate sales prediction aids pharmaceutical manufacturers in planning production schedules to meet anticipated demand, optimizing resources and reducing wastage.

b. Distribution Planning: Efficiently plan distribution strategies and logistics to fulfill predicted demand across regions or outlets.

ii. Marketing and Promotion Strategies:

a. Targeted Marketing Campaigns: Insights from sales forecasts help in designing targeted marketing campaigns for specific drug categories or regions, optimizing promotional efforts.

b. Strategic Pricing: Adjust pricing strategies based on anticipated demand, ensuring competitive pricing while maximizing revenue.

iii.Financial Planning and Budgeting:

a. Budget Allocation: Sales forecasts assist in financial planning by providing estimates for revenue generation, aiding in budget allocation and resource planning for pharmaceutical companies.

iv. Healthcare Resource Allocation:

a. Healthcare Facility Planning: Anticipating drug demand aids hospitals and healthcare facilities in planning their pharmaceutical inventory to meet patient needs efficiently.

v. Regulatory Compliance and Safety:

a. Compliance Management: Ensuring compliance with regulations by maintaining necessary stock levels of pharmaceuticals as per regulatory requirements.

b. Patient Safety: Guaranteeing continuous availability of essential drugs contributes to patient safety and well-being.

vi. Research and Development:

a. R&D Investment: Insight into future demand trends can guide investment decisions in research and development of new drugs or improvement of existing ones.

vii.Business Decision-Making:

a. Strategic Decision Support: Sales forecasts serve as a critical input for strategic decision-making processes, guiding long-term business strategies.

viii. Inventory Management Optimization:

a. Stock Level Adjustment: Forecasting pharmaceutical sales enables optimized inventory management by ensuring adequate stock levels for different drug categories based on predicted demand.

b. Reduced Overstocking or Stockouts: Preventing overstocking minimizes unnecessary costs, while avoiding stock outs ensures continuous availability of drugs.

These applications demonstrate the versatility and significance of a Pharma sales prediction model.

8. Future Scope

I.Enhanced Model Accuracy:

a. Fine-Tuning Algorithms: Exploring more sophisticated algorithms or hybrid models to improve prediction accuracy.

b. Incorporating External Insights: Integrating external factors like economic indicators or weather patterns for refined and comprehensive predictions.

II. Real-Time Prediction Capabilities:

a. Live Data Integration: Developing frameworks to process real-time data for immediate and accurate sales forecasts.

b. Interactive Predictive Dashboard: Creating dashboards offering real-time insights for quick decision-making.

III. Industry-Specific Customization:

a. Tailored Solutions: Customizing models for specific pharmaceutical sub-sectors or markets, accounting for unique demand patterns.

b. Geographical Adaptation: Adapting models for different regions, considering regional variations in demand and market dynamics.

VI. Long-Term Forecasting and Trend Analysis:

a. Extended Forecast Horizon: Exploring methodologies for longer-term forecasts to aid strategic planning.

b. Pattern Recognition: Utilizing advanced analytical tools to identify evolving trends in sales data for proactive decision-making.

V. Explanatory Model Outputs:

a. Interpretability Focus: Developing techniques to make model predictions more interpretable for better understanding of influencing factors.

VI. Integration with Emerging Technologies:

a. AI-Driven Predictive Analytics: Leveraging advancements in AI and machine learning to enhance predictive capabilities.

b. Blockchain for Transparency: Exploring blockchain technology to ensure transparency within the pharmaceutical supply chain.

VII. User-Centric Interfaces and Accessibility:

a. Intuitive Interfaces: Designing user-friendly interfaces for easy interaction and comprehension of model outputs.

b. Mobile and cloud integration: Ensuring accessibility by integrating models into mobile platforms and cloud services.

VIII. Ethical and Privacy Considerations:

a. Privacy-Preserving Techniques: Implementing robust privacy measures to protect sensitive sales and patient data.

b. Ethical Guidelines and Compliance: Establishing ethical guidelines to govern data usage and model development in line with industry regulations.

9. Conclusion

In summary, this study underscores the pivotal role of data science in pharmaceutical sales forecasting, showcasing the integration of advanced techniques like ARIMA and LSTM. The industry's demand for precise sales predictions to optimize resource allocation and inventory management is met through the development of a comprehensive forecasting engine. The application of ARIMA for short-term and long-term forecasts, with grid search for hyperparameter optimization, emphasizes the importance of traditional time series analysis. Simultaneously, the exploration of various LSTM architectures highlights the efficacy of deep learning in capturing intricate patterns for long-term forecasting.

The automated model selection process, centered on accuracy, streamlines decision-making, and its seamless integration into an intuitive dashboard enhances accessibility. This research provides industry professionals with systematic insights into trends, seasonality, and predictive modeling, facilitating efficient and informed decision-making. The forecasting engine and interactive dashboard establish a robust foundation for navigating market dynamics, ensuring consistent medication production, and meeting diverse healthcare needs. As the pharmaceutical landscape evolves, the fusion of traditional time series methods and advanced deep learning techniques positions the industry for adaptive and precise sales predictions, contributing to its ongoing success in serving global healthcare requirements.

References

[1] Y. Han, “A forecasting method of pharmaceutical sales based on ARIMA-LSTM model,” 2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCT).

[2] Cerqueira, V., Torgo, L. & Mozetič, I. “*Evaluating time series forecasting models: an empirical study on performance estimation methods*” (2020)

[3] He, Xin & Zhao, Kaiyong & Chu, Xiaowen. (2021). “*AutoML: A survey of the state-of-the-art. Knowledge-Based Systems.*”

[4] Khalil Zadeh, Neda & Sepehri, Mohammad Mehdi & Farvaresh, Hamid. (2014). “ *Intelligent Sales Prediction for Pharmaceutical Distribution Companies: A Data Mining Based Approach.*”

[5] F. Majidi, M. Openja, F. Khomh and H. Li, “*An Empirical Study on the Usage of Automated Machine Learning Tools*,” in 2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)

[6] Sima Siami Namin, Akbar Siami Namin, “*Forecasting Economic and Financial Time Series: ARIMA vs LSTM*” (2018)

[7] Dariusz Kobielaa , Dawid Kreftaa , Weronika Krol , Paweł Weichbrothb, “*ARIMA vs LSTM on NASDAQ stock exchange data*” in 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)