NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 2: Siemens

Chloé Deschanel, number: 20240693

Diogo Carvalho, number: 20240694

Ingrid Gil Lopez, number: 20240692

Rúben Marques, number: 20240352

Group D

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March 2025

INDEX

1. EXECUTIVE SUMMARY

In a complex and data-driven world, accurate forecasting of sales is essential to ensure operational efficiency, resource allocation, and strategic decision-making. Being a global leader in technology and innovation, Siemens wants to introduce artificial intelligence into the heart of its business. To this end, the present project seeks to develop a sales forecasting model for Siemens' Smart Infrastructure business unit in Germany based on past sales data and macroeconomic variables to improve forecast accuracy and business performance.

The main objectives were to create a robust, automated forecasting pipeline capable of delivering monthly sales predictions per product category, include external economic variables, and evaluate model performance using RMSE, time efficiency, and responsiveness. The project aimed to reduce human bias and enhance Siemens' forecasting processes with scalable machine learning approaches.

Through a systematic CRISP-DM process, the team preprocessed and cleaned historical sales data and macroeconomic factors. SARIMA was applied to forecast economic factors, while various algorithms were tried for sales forecasting. Among these, XGBoost with hyperparameter optimization achieved the lowest average RMSE across categories and was therefore the better performing model overall. However, inconsistencies in preprocessing, especially missing values between the training and test sets, highlighted the need for further refinement before the model can be reliably deployed.

To move forward, it is recommended to finalize and standardize the preprocessing pipeline, automate data integration and retraining processes, and deploy the forecasting model into Siemens' internal environment using a batch deployment strategy.

In short, the project has immense potential for AI application in sales forecasting at Siemens. With certain improvements, the model can be a useful tool, facilitating data-driven decision-making, optimizing resource allocation, and assisting Siemens in realizing its broader digital transformation goals.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. THE CONTEXT – SALES FORECASTING IN TODAY'S WORLD

As a leading global technology company focused on industry, infrastructure, mobility, and healthcare, Siemens' purpose is to create technology to transform the everyday, for everyone (Siemens, 2024). In fiscal 2024, which ended on September 30, 2024, Siemens Group generated revenue of €75.9 billion and net income of €9.0 billion (Siemens, 2024). With a diverse and extensive product and service portfolio, Siemens provides solutions ranging from automation, energy distribution, rail transportation, industrial manufacturing, medical imaging, to digitization services. To maximize value for customers, Siemens is committed to integrate AI into its business operations. In fact, the CEO of Siemens AG, Dr. Roland Busch, stated that "we will build AI into all our offerings based on a coherent data strategy".

For businesses to be successful and their processes to be efficient, sales forecasting plays an important role, as it ensures efficient resource, marketing and finance allocation (Gustriansyah, et al., 2022). Sales forecasting can be defined as "how much of a product is likely to be sold in a specified future period in a specified market at specified price" (Deepa and Raghuram, 2021; p.3928). In order for forecasting to be beneficial, Lawrence et al. (2000) argue that accuracy is the most crucial criterion. In fact, research demonstrates that many businesses lose at least 10 percent of their net gross profit due to forecast inaccuracies (Wacker et al., 2002). Wacker et al. (2002) further argue that this can result in increased costs, as well as "production replanning that creates purchasing, financing, and scheduling difficulties" (p.1019). While some studies highlight a preference for judgmental forecasting practices (e.g. Winklhofer et al., 1996), other recent ones contend that managers find it more challenging to make decisions on manual and judgmental forecasting, especially in an continuously changing environment. For example, Lawrence et al. (2000) conducted a field study on judgmental sales forecasting across 13 manufacturing organizations. They reported that forecasts were not consistently more accurate, which may be due to inefficiencies, systematic errors or inherent biases. They also question whether better forecasts could have been made in advance using a computer-based forecasting model. Indeed, in today's environment, data is abundant and businesses need to learn how to handle this to make data driven decisions rather than decisions based on intuition. Therefore, the need to automate and use machine learning algorithms for processes, such as sales forecasting, is critical to remain competitive.

2.2. BUSINESS OBJECTIVES AND SUCCESS CRITERIA

Aligning with academic and field research, as well as Siemens' vision, an AI-driven sales forecast model is essential for mitigating against challenges and for enhancing resource efficiency, reducing human bias, integrating scattered data resources, and reducing the opportunity cost associated with inaccurate forecasting (i.e. working capital and customer satisfaction). Thus, Siemens aims to develop a monthly sales forecasting model for its Smart Infrastructure division in Germany, to 1) improve forecasting accuracy, 2) enhance scalability and automation of the process, and ultimately, 3) optimize business performance.

To measure the business success, the model should 1) show accuracy with a reduction in RMSE (Root Mean Squared Error) and forecast bias (targeting ≥25% improvement and ≤±5% bias) compared to other methods (e.g. manual forecasting), 2) save time by reducing human intervention by over 75%

and shortening forecast cycle time, 3) improve working capital efficiencies through an increase in inventory turnover and reduction in excess inventory or stockouts, and 4) be adaptable across product categories with ≤10% performance variation and retraining time under 1 hour, even under volatile market conditions.

2.3. SITUATION ASSESSMENT

To achieve this, a dedicated Siemens team of four data scientists is responsible for the project. They are supported by domain knowledge, computational resources and access to two key datasets. The first is historical sales data from Siemens Smart Infrastructure division, Germany. This data includes daily sales data per product group (GCK) in EUR from October 2018 to April 2022. The second is macroeconomic indicators, spanning from February 2004 to April 2022. It includes information on production indices for machinery and electricals, shipments index, raw material prices, and producer price indices. Using these datasets, the objective of the project is to develop a model capable of accurately forecasting sales quantities for each product category over a 10 month period from May 2022 to February 2023. Model performance will be assessed using RMSE to ensure accurate and reliable predictions. The project follows the CRISP-DM methodology, an established framework recognised for its effectiveness in the industry and academia [Fig.1].

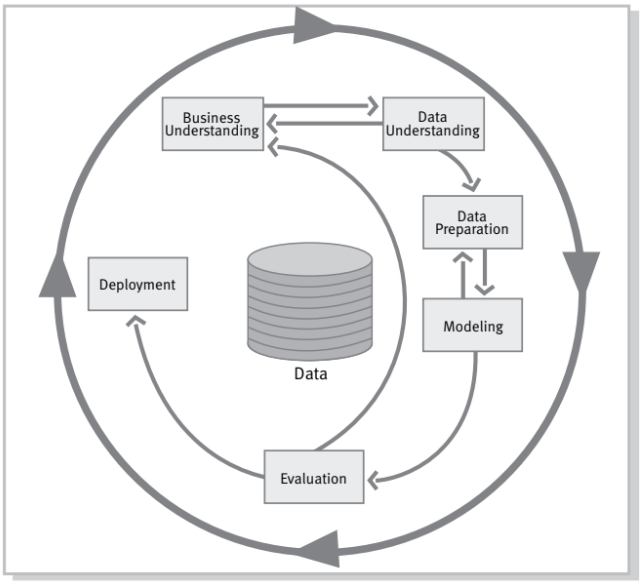

Fig.1: CRISP-DM Methodology

For the project to succeed and the model to be robust and practically valuable, there are considerations and challenges to take into account.

On the technical side, the datasets need to be properly pre-processed to avoid inconsistencies, errors or gaps that could affect performance. In fact, one key challenge is missing market data from January 2022 to April 2022, which represents a gap in the dataset just before the forecast period begins. Similarly, there is an additional uncertainty because the model needs to forecast sales past the available data. Thus, this requires to also predict features used in the model for the 10 month forecast

period. This will require filling missing data with estimations and predictions, but this could introduce bias, as well as overfitting.

On the business side, stakeholders across different levels of the organisation must be onboard, to ensure the model is properly adopted and benefit from the practical impact of the project. Stakeholders must understand the value and see it as an asset and tool that can provide valuable business insights and intelligence that can contribute to strategic decision-making.

Finally, external factors are another challenge, because they are not explicitly captured in the datasets. For example, consumer behaviour and economic stability can be affected by Covid-19 pandemic or changes in geopolitical policies. However, these factors are difficult to measure and incorporate into historical models. Forecasts can be guided by historical trends, but the model's applicability in real-world scenarios may be limited by its incapacity to take unexpected external disruptions into account.

2.4. MACHINE LEARNING GOALS

The objectives of machine learning in this project are 1) to preprocess optimally by building a data pipeline that cleans, imputes missing data, removes outliers, and engineers features for maximum predictive accuracy, 2) to predict market data through statistical methods ensuring consistency with sales forecasting and economic realism, and 3) to predict sales by comparing time series and machine learning models, and using macroeconomic predictions, hyperparameter tuning, and selection of the most accurate model based on RMSE.

3. METHODOLOGY

3.1. DATA UNDERSTANDING

3.1.1. Sales Data

The sales dataset consists of 9,802 rows and three columns. These include 1) date records which are daily transactions in dd.mm.yyyy format, 2) the product category for each transaction, and 3) sales revenue in euro per transaction. There are no duplicates or missing values. Initially, all columns are stored as objects; these are converted to appropriate types: date becomes datetime64[ns] in yyyy-mm-dd format, and sales becomes float to account for decimals. When checking for potential anomalies, there are 276 instances of negative sales. Every product category has at least one case of negative sales. Product 3 has the most (56), while product 13 has the fewest (3). These may indicate returned items, failed transactions, or products that never left the factory due to defects. Although the term "sales volume" typically refers to units sold, in this case it seems to reflect the frequency of product appearances. Across 14 distinct product categories, Product 1 appears most frequently with 1,179 instances. Regarding the total revenue across the dataset, it amounts to around €2,673,845,301. However, the distribution is highly skewed, the top three products [Fig.2], which are 1, 3 and 5, contribute to 93.77% of the total revenue. Indeed, product 1 is the leading revenue generator. It doesn't dominate in frequency, which suggests higher unit prices or larger order sizes. Products 3 and 5 have high transaction counts but lower revenue, likely lower-priced items.

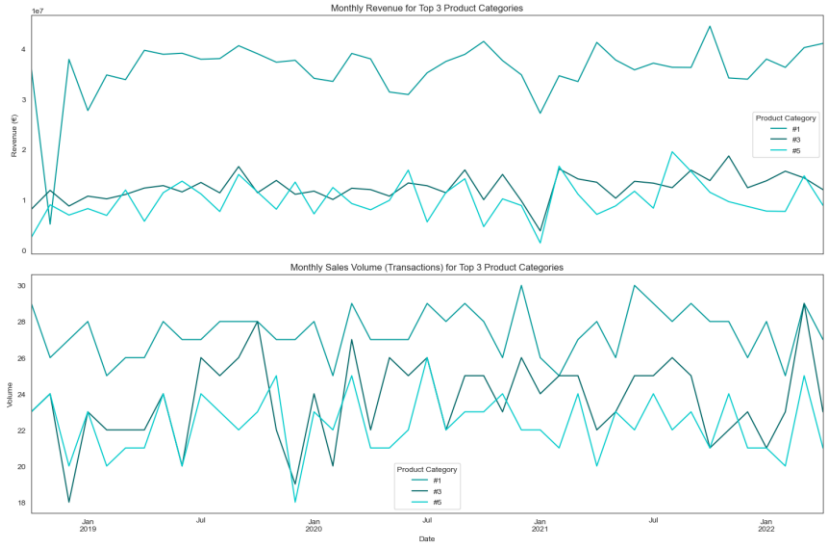

Fig.2: Sales revenue and frequency of top 3 products

When checking seasonality, there are noticeable month-to-month and year-to-year fluctuations in both frequency and revenue. March sees the highest transaction count (969) and revenue (€264.7M), while May records the lowest frequency (666) and €181.97M in revenue. On average, monthly frequency stands at 816.83, and monthly revenue at €222.82M. The mean annual revenue is around €534.77M. 2020 has the highest transaction volume (2,779), 2021 tops revenue charts with €773.2M, and 2018 shows the weakest performance, both in frequency (639) and revenue (€141.54M). However, it is important to note that there are only 3 months in 2018, as the dataset starts from October 2018, and 2022 only has 4 months.

After reviewing overall time-based patterns, seasonal decomposition is applied to better understand the time series, including trend, seasonal, and residual components [Fig.3]. There is a gradual upward trend from 2018 through 2019, then a slight dip in early 2020 (likely Covid-related), and finally strong recovery in 2021 and early 2022. Seasonally, the data shows consistent monthly effects, where some months (e.g., year-end or mid-year) regularly spike in sales, while others (e.g., January or summer months) experience dips. These seasonal patterns are not symmetrical, the size of upswings and downswings varies.

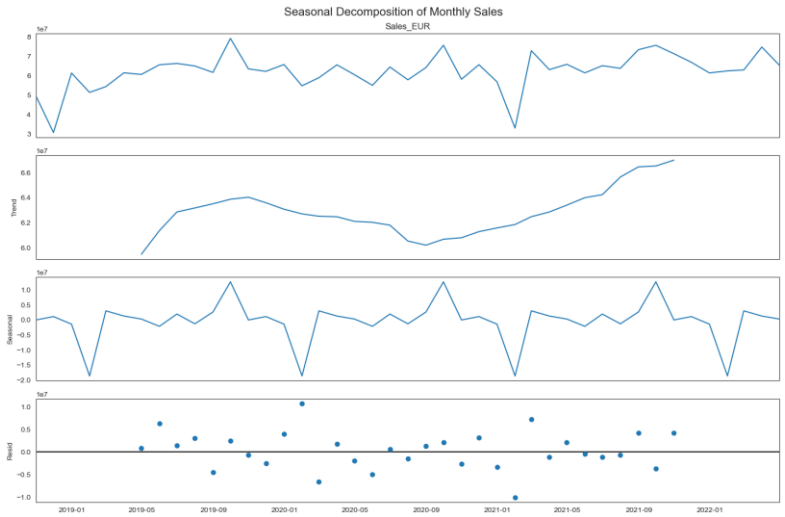

Fig.3: Seasonal Decomposition of Monthly Sales

At the product level, 2020 shows a revenue drop across nearly all categories, especially for product 13, which has yet to fully recover. Product 16, in particular, has experienced a steady decline since 2019. However, revenue increases in other products have helped balance out these losses, keeping the overall annual revenue trend relatively stable. Looking at average monthly sales [Fig.4], March is generally strong across many products (especially 1, 3, 5, 12, 13), with the exception of product 36. The September–December window also performs well across products. This can potentially be driven by holidays, end-of-year campaigns, or Q4 pushes. Conversely, January and the summer months tend to be slower, likely due to vacation or off-season effects.

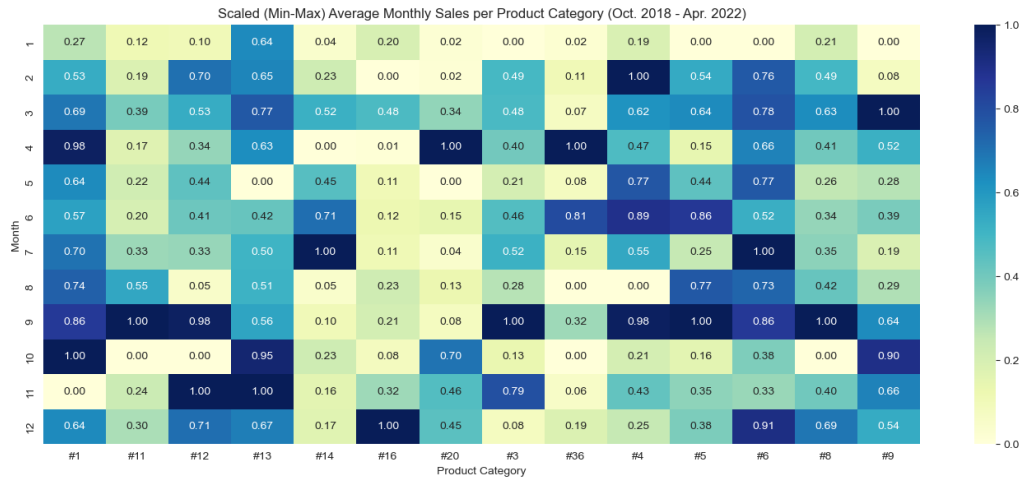

Fig.4: Scaled Average Monthly Sales per Product

Finally, as for inter-product relationships, no strong linear correlations are found overall. The highest positive correlation is between Product 8 and Product 12 with a value of 0.6.

3.1.2. Market Data

The market dataset contains 219 entries and 48 columns, with dates representing monthly observations across several years. The features include product and shipment indexes for machinery and electricals, as well as production indexes for machinery and electrical equipment from a range of countries, like Germany, the United Kingdom, the United States, and China. Additionally, the dataset incorporates global commodity price indicators. There are no duplicate records. However, ten features contain missing values, two of which fall within the time frame shared with the sales dataset, spanning from October 2018 to April 2022. Below are interesting insights found in this dataset.

The production and shipment indices for machinery and electricals reveals that China stands out in both measures. From around 2008 onward, China exhibits a steep and consistent upward trajectory. In contrast, the indices for other countries remain relatively stable over time, suggesting more limited long-term growth in the sector compared to China. Temporary dips in the indices are observable around 2008/ 2009 and again in early 2020, aligning with the global financial crisis and the initial impact of the COVID-19 pandemic. Most countries show signs of recovery following these disruptions.

Global commodity prices show a pronounced spike between 2021 and 2022, reflecting broader inflationary pressures, supply chain disruptions, and the economic rebound following the pandemic. Natural gas and crude oil prices are particularly volatile, often responding rapidly to global crises. In contrast, metals and base materials track more closely with industrial production cycles and global

manufacturing demand. Major downturns across all commodities coincide with the 2008 financial crisis and the 2020 pandemic shock, both of which caused synchronized, cross-commodity price declines.

3.2. DATA PREPARATION

3.2.1. Market Data Preprocessing and Forecasting

There is a wide range of macroeconomic data, reflecting different economic variables influencing sales. By applying tailored preprocessing techniques, we expect to improve modeling accuracy and interpretability. Therefore, we decided to separate the dataset into four categories:

1. Product & Shipment Indices (industrial output),
2. Material Prices (commodity cost structures),
3. Producer Prices (upstream inflation pressures),
4. Production Index (macroeconomic activity).

To detect statistical outliers in all macroeconomic segments, an IQR-based method was applied, adapting the thresholds to reflect the economic nature of each variable. In some cases, particularly in price data, outliers represent real market phenomena, such as supply shocks or inflationary spikes, rather than noise. In order not to disrupt these significant signals, a contextualised approach was chosen in which stable indicators were treated more strictly, leaving volatility series intact or selectively limiting them. This strategy preserved the predictive power of market dynamics, ensuring a balance between statistical rigor and economic realism, while improving the robustness of the model and the reliability of forecasts.

Regarding missing values, a layered imputation strategy adapted to the temporal and economic behaviour of each indicator was used to deal with them. Forward fill and linear interpolation were mainly used to maintain continuity and consistency of trends in time series where lags were small and recent values represented reliable approximations. In cases of longer or irregular lags, especially in output prices, moving averages were used to preserve smoother transitions, followed by backfilling to remove any residual NaN. This approach ensured that all series remained complete and consistent with the trend.

Finally, to obtain forecasts of market data for the next 10 months, each macroeconomic dataset was modelled individually using the SARIMA (Seasonal Autoregressive Seasonal Integrated Moving Average) technique, a method that captures both time trends and seasonal fluctuations, which are particularly prevalent in industrial production series, shipping indices and commodity prices. The flexible treatment of non-stationary data, together with automatic parameter adjustment by auto_arima, allowed the models to be statistically robust and computationally efficient.

3.2.2. Sales Data Preprocessing

The original dataset collects daily sales records per product, but as the forecasting task is designed to operate at a monthly granularity, the sales values were aggregated accordingly. In addition, the dataset was split into 14 individual datasets to allow for independent modelling per product category. This approach allows for greater flexibility in capturing seasonality, growth patterns and product-specific noise characteristics, which ultimately improves forecast accuracy and model interpretability.

Prior to feature engineering, the datasets were then split into training and test sets using a hold-out approach adapted to time series data. The split was performed at the end of December 2021, so that all data prior to this date were used for training, while data from January 2022 onwards were retained for testing. This chronological separation preserves the time structure of the series and avoids data leakage that might be caused by random shuffling. As a result, training covers approximately 90% of the time period (39 months) and test covers 10% (4 months), across the 14 product categories.

To improve the predictive power of the sales forecasting model, several variables were engineered to capture temporal patterns, seasonality and external economic influences. These include:

- Lag characteristics: These represent past sales values at key intervals, such as the previous month, the previous quarter and the same month of the previous year. Thus, the model recognises recurring short-, medium- and long-term trends in sales behaviour.
- Moving averages: Moving averages were calculated over 3 and 6 months (excluding the current month) to smooth out short-term fluctuations and highlight local trends, giving the model a context of how sales have evolved over time.
- Temporal features: Monthly and quarterly indicators have been incorporated to reflect calendar-based sales cycles, allowing the model to detect and learn from seasonal patterns that repeat over time.
- Lag and delta characteristics of macroeconomic variables: For each external variable, the previous value and its month-to-month variation have been included. In this way, the model not only learns from current macroeconomic conditions, but also from their direction and momentum, which is essential to capture the influence of economic changes on sales performance.

Finally, a structured feature selection process was conducted to determine the most influential macroeconomic indicators driving sales. Prior to model training, several features were discarded because they presented a risk of data leakage, introduced noise or lacked sufficient predictive power. After this initial filtering, a separate XGBoost regressor was trained for each product group. From each model, the ten most important features were extracted based on their contribution to error reduction in the boosted decision trees. This model-based system allowed the identification of nonlinear interactions and featured importance in a robust, data-driven manner.

3.3. MODELING

Sales can be considered as a time series. This means that observations are recorded sequentially over time at regular intervals. It is often used to analyze trends, seasonality, and patterns to make future predictions. Historically, sales forecasting has been influenced by classical statistical methods of forecasting (Gustriansyah et al., 2022). This includes autoregression (AR) methods like ARIMA or ETS. However, Pavlyshenko (2019) suggests that there are some limitations of time series approaches for sales forecasting, including not facing access to enough historical data, sales data having a lot of outliers and missing data, as well as exogenous factors. In fact, the researcher argues that sales prediction is rather a regression problem rather than a time series one. The use of regression approaches for sales forecasting can often give better results compared to time series methods (Pavlyshenko, 2019). More recently, other methods are employed, such as Prophet, which combines

additive regression with seasonality modeling, or generative AI, which leverages deep learning to uncover complex patterns and enhance predictive accuracy.

For our forecasting approach, we tested both a time series model and a machine learning model. We chose Prophet (Meta's open-source algorithm) for its speed, ability to capture seasonal trends, and automation, making it efficient with minimal manual tuning. For the machine learning approach, we selected XGBoost due to its high accuracy, scalability, and ability to handle complex relationships in data. It also includes regularization to prevent overfitting and effectively manages outliers.

3.4. EVALUATION

To assess model performance, we trained and predicted sales for each product individually using Prophet and XGBoost. Prophet without regressors served as a benchmark, while we also tested Prophet with regressors. For XGBoost, we compared a version without hyperparameter tuning to one optimized with GridSearch. We evaluated each model using RMSE for every product and then calculated the average RMSE across all products to determine the best-performing model. Since no single model was consistently superior for all products, this approach helped us identify the most accurate overall. Based on the results, XGBoost with hyperparameter tuning had the second lowest average RMSE but we chose it because the XGBoost without hyper parameters tended to overfit them in some products making it our final chosen model. See table below for scores and Fig.5 for predicted sales.

Model	Average RMSE
Prophet Vanilla	931,954.3549
Prophet with Features	2,698,845.8484
XGBoost	709,405.5846
XGBoost Tunned	703,202.7735

Table 1

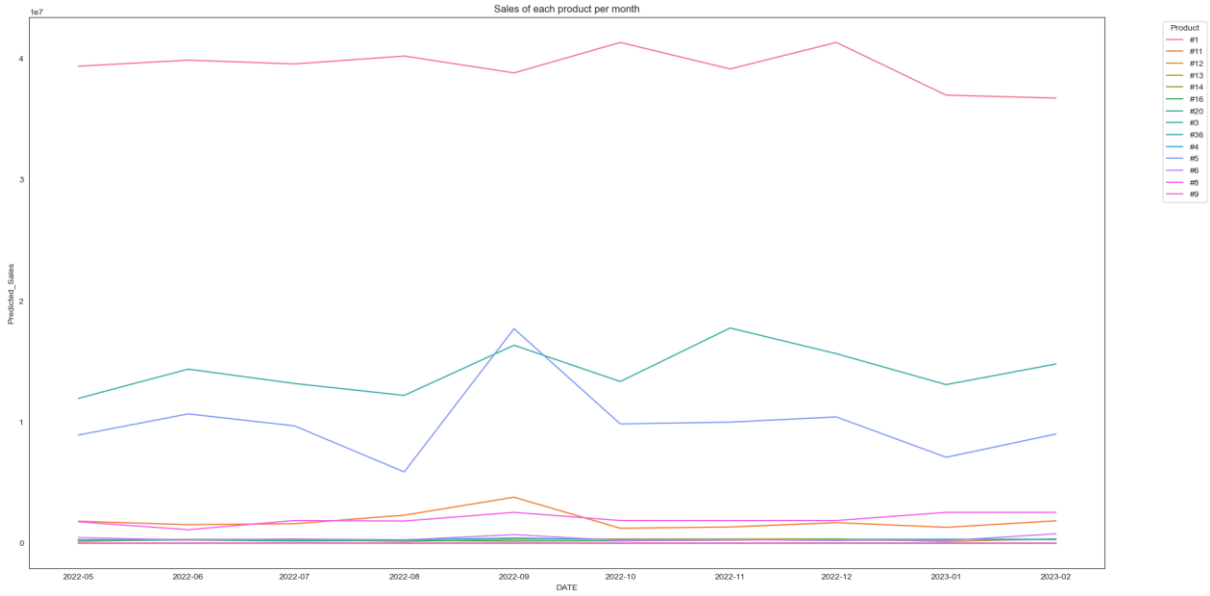

Fig.5: Predicted Sales of each Product per Month using XGBoost Tunned

4. RESULTS EVALUATION

Here, we match the machine learning results with the business success metrics defined above. We could successfully forecast macroeconomic indicators using SARIMA models for each dataset, covering both seasonality and trends. We incorporated these forecasts into the sales forecasting pipeline. In predicting sales, we tried two algorithms for 14 product categories. These are Prophet (with and without regressors) and XGBoost (with and without parameters). Of the RMSE calculated for all products and models, XGBoost with hyperparameter tuning was the best overall model with the lowest average RMSE for all product predictions.

However, the preprocessing pipeline can be further improved. We encountered missing values in the input features of the test set that were absent in the train data. We employed a backfill/forward fill strategy, complemented with a fixed window of 100 values to preserve sequence length to account for this gap. This was a pragmatic solution to maintain continuity and prevent model failure, but it identifies the need to fully optimize and synchronize preprocessing between training and testing phases. Until this is resolved, final model deployment cannot be approved for production. Having resolved them, the model can be transferred to production with confidence.

5. DEPLOYMENT AND MAINTENANCE PLANS

Model deployment for time series models can be challenging. According to Kis (2004), volatility, model drift, the need for robust real-time processing, and maintaining accuracy over time are all significant obstacles. To address these in the context of this project, the sales forecasting model needs to be deployed within Siemens' internal data infrastructure in a way that it is compatible with the firm's enterprise-level standards for security, scalability, and governance. Given the task's nature, predicting monthly sales revenue per product category, a batch deployment strategy is recommended. In this method, data is processed in batches at periodic intervals so that the model can make predictions for a batch at a time. The forecasts will be automatically delivered to Siemens' business intelligence tools, where planners and analysts will be able to view and interpret them. Forecast results

will make possible activities such as sales planning, inventory forecasting, and strategic planning, and will be accessible for comparison with actual results, trend analysis, and export for further analysis.

As the model is already trained and validated, deployment would be feasible within a month. Developing the deployment pipeline and integrating it with Siemens' infrastructure would take two weeks. There would be a week's requirement for internal security and compliance review and a week for the final dashboard integration and stakeholder testing. Because Siemens' needed infrastructure is probably already present, there won't be huge infrastructure costs. But the majority of the cost will be developer time, which is integration and setup related. This can be fully justified by the long-term business value of a stable, reliable forecasting tool.

Once deployed, the model will require ongoing monitoring and maintenance to ensure that its performance does not decline over time. Siemens' in-house tools will be employed to track model performance, with RMSE being monitored monthly to track accuracy against actual sales data. At the same time, data drift and concept drift will be monitored through statistical tests to detect any significant shifts in input data patterns or model behavior. Input data, projections, and errors will be tracked consistently to provide auditing and visibility. In the event of the system detecting irregularities or a drop in performance, automatic alerts will be initiated via Siemens' internal communication infrastructure.

As for model updates, retraining will occur quarterly using new sales and market data. If performance drops sharply (e.g. a significant increase in error metrics) retraining can be triggered earlier. Every new version of the model should be version-controlled, ensuring traceability and the ability to roll back if needed. Before any updated model is deployed, it will be validated on a hold-out dataset and reviewed by relevant stakeholders to confirm it meets performance standards.

6. CONCLUSIONS

This project demonstrates the benefits of using a sales forecasting model in assisting Siemens achieve its business goals. We compared different algorithms with historical sales and macroeconomic factors and found that XGBoost with hyperparameter optimization gave the best model based on RMSE. While the forecasting pipeline shows promising initial results, it is not yet production-ready because of inconsistent preprocessing between training and test sets. These must be addressed for reliability and performance with real-world deployment. Nevertheless, the project provides a solid foundation and a definite direction. With further advancement, the model can be integrated within Siemens' setup to aid in planning, improve working capital management, and ease data-driven decision-making.

7. REFERENCES

Deepa, K. and Raghuram, G., (2021) "Sales forecasting using machine learning models", *Annals of the Romanian Society for Cell Biology*, 25(5), pp. 3928-3936.

Gustriansyah, R., Ermatita, E. and Rini, D.P., (2022) "An approach for sales forecasting", *Expert Systems with Applications*, 207.

Lawrence, M., O'Connor, M. and Edmundson, B., (2000) "A field study of sales forecasting accuracy and processes", *European Journal of Operational Research*, 122(1), pp. 151–160.

Pavlyshenko, B. M., (2019), "Machine-Learning Models for Sales Time Series Forecasting", *Data*, 4(1), 15.

Siemens AG, (2024), *Press Material*, 14 November. Available at: https://press.siemens.com/global/en/article/press-materials-siemens-ag [Accessed: 20 March 2025].

Wacker, J.G. and Lummus, R.R., (2002), "Sales forecasting for strategic resource planning", *International Journal of Operations & Production Management*, 22 (9), pp. 1014-1031.

Winklhofer, H., Diamantopoulos, A., and Witt, S., (1996), “Forecasting practice: A review of the empirical literature and an agenda for future research”, *International Journal of Forecasting, 12*(2), pp. 193-221.