NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 2: Sales Forecast

Beatriz, Monteiro, number: 20240591

Catarina, Gonçalves, number: 20230083

Margarida, Raposo, number: 20241020

Teresa, Menezes, number: 20240333

Group P

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

March, 2025

INDEX

EXECUTIVE SUMMARY

"If we want things to stay as they are, things will have to change." — Giuseppe Tomasi di Lampedusa, The Leopard

In the business world, particularly in sales forecasting, this principle is evident. Research from the Aberdeen Group highlights that **97% of companies with industry-leading forecasting processes achieve their sales quotas**, compared to only 55% of those without such processes. This statistic illustrates that to sustain or improve sales performance, companies must evolve by implementing robust forecasting methods. Accurate sales forecasts enable smarter decisions, stronger strategies, and sustainable growth, aligning with the idea that **change is essential to preserve success**.

Having said that, in this report we will assess the **Siemens sales data** in addition to important **external market data** in to develop an informed sales prediction. The primary focus will be on comprehending external **macroeconomic factors** that can impact sales trends as well as internal business success.

Our goal is to conduct **sales forecasting** monthly for 10 months, from May 2022 until February 2023, selecting product groups from one business unit of the Smart Infrastructure division, with a specific focus on Germany, the largest country for this business unit.

By examining internal sales data, production levels, and shipments in addition to including external variables like market prices, economic indicators, and industry trends, our goal is to **create an accurate model that can predict future sales**. This analysis will **help in determining the key variables** that affect sales success in addition to providing useful information for strategic decision-making.

Considering an average monthly salary of €2,824 for a data consultant, the three-week development period represents an estimated cost of €2,118 per person. With a team of four members, the total project investment is approximately **€8,472**—a highly cost-effective approach compared to external forecasting solutions.

BUSINESS NEEDS AND REQUIRED OUTCOME

A clear understanding of the business needs and requirements is essential for data-driven solutions to align with real world demands. As such, this chapter lays the foundation for suitable results, by assessing the business' current situation, defining primary objectives, and outlining how data mining can address key business challenges throughout the completion of the project.

BUSINESS OBJECTIVES

Siemens is a global technology company that provides multi-industry solutions, with a strong focus on automation, digitalization, and sustainability.

In October 2018, Siemens launched its Smart Infrastructure division, as part of the implementation of an optimized corporate structure. This new division emerged from the merger of former divisions, specifically the power distribution segment of the Energy Management division, the Building Technologies' business, and the control products business of the Digital Factory division.

Smart Infrastructure helps customers through their digital transformation journey by leveraging cutting-edge technology that enables clients to adapt to evolving environments and to seize new opportunities. It mostly provides B2B services, focusing on energy efficiency and resource optimization, which in turn improves clients' asset performance, availability, reliability, and operational efficiency. All in all, this division future-proofs infrastructures, fostering collaborative ecosystems.

From 2018 to 2022, several economic events that impacted Smart Infrastructure's activity took place:

2018: The global economy was affected by the US monetary policies and the fear of a global trade war. The political tension created by the US sanctions on Iran, and the uncertainties caused by the Brexit and Italy's budget negatively affected the investment sentiment in Europe. *[1]*

2019: The US and China trade dispute and tensions in the Middle East took a tool on global investments, while the decline in global trade negatively impacted regions dependent on trade. The uncertainties surrounding Brexit continued to weigh on Europe. *[2]*

2020: The first effects of COVID-19 took place, leading to a global recession and supply chain disruptions. The energy performance and data center markets grew with the pandemic conditions, as the focus shifted to energy efficiency, digital services and remote work. *[3]*

2021: A second wave of COVID-19, led by new variants, continued to disrupt the supply chains and created logistic bottlenecks. Home bounded consumers saw a growth in savings, which led to higher demand of goods, and consequent inflationary pressures. In particular, a semiconductor crisis, due to the increasing demand of the good, affected the Smart Infrastructure division. *[4]*

2022: The war in Ukraine further increased energy prices, heavily impacting Europe and the industrial sector. Additionally, China's strict COVID-19 policies affected the global supply chain. Smart Infrastructure proved resilient, growing in the electrical products and electrification, despite the economic conditions. *[5]*

This project primary business objective is to accurately forecast the monthly sales of Siemens' Smart Infrastructure division from May 2022 to February 2023, providing the best foundation for strategic planning. Additionally, we hope to provide Siemens with a functional and long-lasting sales forecasting model, including all the important components they need to account for when facing changes in the economic environment.

BUSINESS SUCCESS CRITERIA

Throughout the development of the project, the quality of the model was continuously assessed, using various evaluation metrics. However, the success of this project can only be determined with time,

once the actual sales data becomes available and we are able to compare both results. A reliable sales forecast should provide the company with valuable insights on how to efficiently allocate resources, budget, identify key problems and opportunities, and plan financially. As such, the project can be considered successful if the insights generated lead to better decision-making and improvements across these business areas.

SITUATION ASSESSMENT

This project, executed by a team of four business analytics students in collaboration with Siemens' business unit in Germany, aims to forecast its sales for 10 months, from May 2022 to February 2023, and to provide extra insights about the sales behaviour of different Smart Infrastructure's products. The data made available consisted of two csv files, one with previous years' sales data, and the other with macroeconomic indicators. The first contained three variables, the date at which the sale was recorded, the product group to which the product sold belongs, and the sales amount in euros. The second file contained forty-seven variables, consisting of production and shipment indexes for different regions/countries, world prices of materials, exchange rates, and producer prices. **(software)** The analysis of this project will be completed using Jupyter Notebook, with the estimated duration of three weeks.

Confidentiality is ensured by distorting the sales values and masking the product groups. The results must be comprehensible, high-quality, and actionable while adhering to security and legal standards. The lack of feature information led the team to make assumptions about the definition of the macroeconomic indicators used. Constraints include limited access to product characteristics, that would be valuable while choosing variables that influence their sales, as well as omitted sales characteristics (e.g. customer details, count of products sold, etc.) that would provide a better context to the analysis.

The project faces risks such as data inconsistencies, time limitations, or a misalignment between assumptions and the reality. Contingency plans involve iterative data validation and adaptive modelling approaches.

Terminology such as **CRISP-DM**, **Stationary**, **Autocorrelation Function**, **Partial Autocorrelation Function**, **Lags**, **RMSE**, and **Overfit Score** will be used. Refer to the glossary in the appendix for better understanding of these concepts.

While costs associated with forecasting sales include hiring specialized labour to develop the model, the time needed to develop and assess its accuracy, and the employee training to best deploy and maintain the model, potential benefits of implementing its insights include better production planning according to the anticipated demand, revenue growth, and expanded market share, ultimately strengthening Siemens' Smart Infrastructure competitive positioning.

DETERMINE DATA MINING GOALS

The primary objective, from a data mining point of view, is to develop a model that best predicts the sales for the next 10 months, by utilizing the historical sales and market data. The success of this task will be evaluated based on model performance metrics, which will measure the error or difference between our predictions and the actual sales values. Beyond accurate predictions, we aim to provide a long-lasting model, that Siemens can implement for months to come, as well as insights on which macroeconomic indicators most influence each product's sales, helping future forecasters to narrow the scope of macroeconomic factors they need to monitor.

Our **project** plan will begin by developing a comprehensive economic background for the years 2018 to 2022, to get a better understanding of external factors affecting the sales data. The exploratory phase of this project is crucial to its success, due to the lack of information, thus a lot of effort was put into exploring the sales and market data, visualizing its behaviour and understanding its intricacies. We started the modelling stage by assessing sales data stability, meaning if it has consistent patterns

over time, which is a crucial step to ensure a reliable forecasting. Using the appropriated techniques, we determined which factors most influenced sales, from macroeconomic indicators to lag features, which allowed us to focus only on the most influential variables, increasing the accuracy of the modelling stage. After the selection of impactful features, we tested and evaluated various modelling approaches to determine the best fit to the data.

METHODOLOGY

Following the **CRISP-DM methodology**, the process began with **Business Understanding** to establish the key goals and business context for forecasting sales. **Data Understanding & Preparation** involved exploring and transforming the data to gain insights into product sales behaviour in relation to market (macro) features. To support this, we performed feature selection and analysed time series lags to better understand temporal patterns and dependencies. In the **Modelling** phase, we implemented a range of forecasting models, including **TimeGPT, XGBoost, TCN (Temporal Convolutional Networks), ARIMA**, and **Prophet**. These models were then evaluated based on their performance using **Root Mean Square Error (RMSE)** as the key metric. Finally, insights were translated into business recommendations. For implementation details, refer to the notebooks *BCwDS_Case2_groupP_part1* and *BCwDS_Case2_groupP_part2*.

DATA UNDERSTANDING

We were provided with 2 datasets: a **sales dataset** and a **market dataset**. We will begin our explanation by focusing on the sales dataset, as it forms the foundation of our analysis.

The **sales dataset** contains daily sales information, structured across 3 key variables: '*DATE*', '*Sales_EUR'*, and '*Mapped_GCK*', which identifies a total of 14 distinct products. Our analytical approach involves grouping the date by month and treating each product individually, allowing for tailored feature selection and model development specific to the characteristics and behaviour of each product.

We identified 9,802 rows in the sales dataset where '*Sales_EUR'* equals 0, which we interpreted as instances of "no sales." As these entries do not contribute meaningful information to the modelling process, we decided to exclude them from the analysis.

Upon analysis, we observed the presence of **negative values** in '*Sales_EUR'*, with no clear or recurring pattern to explain their occurrence. Several hypotheses may account for these entries. First, although **data recording errors** (such as mistakenly entering a '-' sign) are a possibility, this seems unlikely given the automated nature of the system, as evidenced by the consistent recording of all dates, including those with no sales. A more plausible explanation is that these negative values reflect **product returns**, which are common in the infrastructure sector due to the complexity and scale of the products involved. Returns may result from defective components discovered during installation or from delivery of items that do not meet the required specifications for a given project. Such returns typically lead to the issuance of refunds or credit notes, generating a negative cash flow to offset prior sales. Another possible explanation is **inventory reconciliation**, where previously recorded sales are corrected or adjusted due to stock discrepancies or administrative updates. Given the context and supporting evidence, we will proceed with the assumption that these negative values primarily reflect product returns.

The analysis of negative sales (***Figure 1*** *and* ***2***) reveals that Products 1 and 5 consistently represent the highest financial losses, not necessarily due to the highest number of returns, but likely due to their higher value or complex return logistics. Product 11, with 7 returns and 520K negative sales, shows the highest average loss per return (74,286 euros per return), reinforcing this trend. In contrast, Product 3, despite having the highest return volume, results in significantly lower losses per unit, suggesting it is a lower-cost item. This indicates that not all returns have equal financial impact.

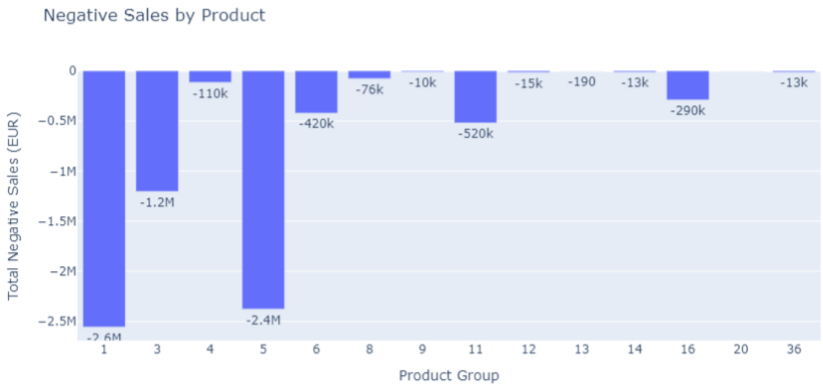

Figure 1: *Total negative sales per product*

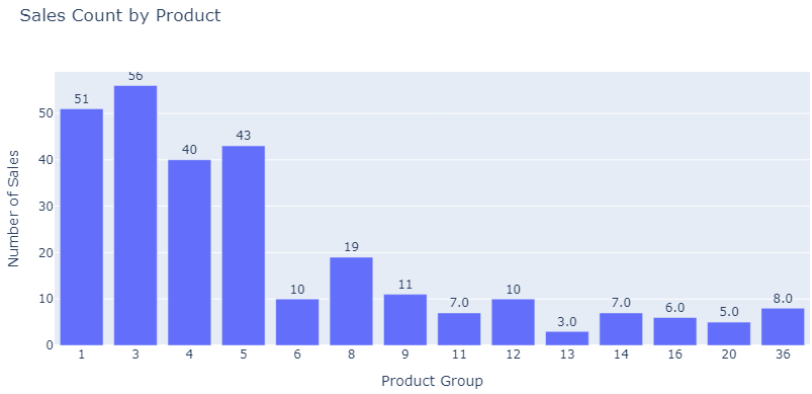

Figure 2*: Number of Sales per product*

The year-by-year breakdown shows that 2019 concentrated several issues, particularly for Products 1, 5, 11 and 16, while 2020 and 2021 continued to reflect challenges with Products 1, 3, and 5. These patterns suggest that some issues may have been localized to specific periods, but the recurring high losses associated with Products 1 and 5 require deeper investigation and possibly strategic action to mitigate ongoing financial impact.

Figure 3: *Yearly Evolution of Negative Sales by Product*

In our analysis of the **positive sales**, the highest values were from Products 1, 3 and 5, meaning they generate the most revenue. However, in negative sales per return, Product 11 stands out. This discrepancy suggests that some products may be far more expensive to handle during returns than they are to sell. This could be due to handling complexity, dismantling costs, or high depreciation upon return.

A deeper analysis of our product variables revealed that all products, except for Product 8, are **stationary**. This means their behaviour remains consistent over time, with **stable** average values and fluctuations. Regarding Product 8, which is **non-stationary**, reveals a clear **upward trend** in sales from 2019 to 2022, indicating consistent long-term growth. A distinct **seasonal pattern** is also present, with regular fluctuations suggesting that sales vary predictably across specific periods—possibly due to recurring demand cycles.

Based on the **Autocorrelation Function (ACF)** and **Partial Autocorrelation Function (PACF)** analysis (check **8.1. Glossary**), we identified clear sales patterns across several products. Products like **P3, P9, and P11** show strong spikes at lag 12, indicating yearly seasonality likely driven by **annual** budgets or calendar-based purchasing. **P4, P5, and P14** exhibit semi-annual cycles, with spikes at lag 6, suggesting biannual planning or reporting influences. Meanwhile, **P1 and P6** show a gradual decline in correlation over time, meaning their sales are influenced by recent months but without sharp seasonal effects. **P20** also shows a gradual tail-off but with a spike at lag 13, suggesting a potential yearly seasonal component in its sales behaviour. **P12** is shaped by short-term trends (1-3 months), where recent sales strongly impact current performance, while **P13** follows a longer, less common sales cycle (17 months), possibly linked to contracts or large-scale projects. **P16** reveals sharp cutoffs after lag 3, meaning the relationship between current sales and past sales is mostly limited to the last three months. Finally, **P36** shows a strong 10-month seasonal pattern, due to a spike at lag 10. This analysis was key to understanding product behaviour and selecting appropriate forecasting approaches.

The **market dataset** had 47 variables that represent important macro-economic indices for Siemens in its most important countries. These variables range from production and shipment indexes to producer prices. They were obtained from various countries, including China, France, Germany, Italy, Japan, Switzerland, the United Kingdom, and the United States, providing insights into both European and global markets. The dataset also contained the prices of base metals, of energy, of metals and minerals, the natural gas index, the average price of crude oil and the copper price.

To better understand market dynamics, we first examined the **shipments and production levels** of various countries (***Figure 4***). Our analysis revealed that the majority share of shipments and production is held by European countries, while China individually dominates the market for both.

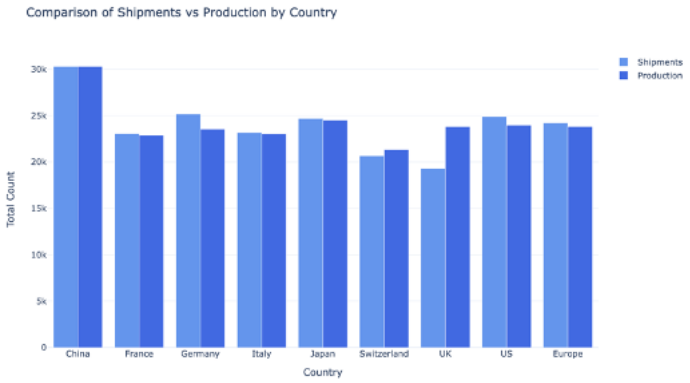

Figure 4: *Comparison of Shipments vs Production by Country*

However, there were significant negative peaks in 2010 and 2020. These declines may be attributed to the **sovereign debt crisis** in 2010 and the **COVID-19 pandemic** in 2020. We also noticed that the majority of nations have larger shipments than production levels. China, Switzerland, and the United Kingdom are an exception, though. We think that **Brexit's effects** may have contributed to the decline in the UK in 2020.

Beyond production and shipments, another crucial factor influencing the market is **the fluctuation of metal prices** over the years, with notable lows in 2009 and 2020, which may be linked to the previously mentioned factors. Now taking a closer look, the Natural Gas Price Index is consistently higher than other indices, especially in 2022. Both the average price of energy and crude oil reached a significant

peak in June 2008, most likely due to speculation in the oil market and the global financial crisis. The start of the conflict between Russia and Ukraine may also have contributed to the notable rise in all indices around 2022.

By analysing **the price of electrical equipment**, we noticed that the United States saw the largest increase, peaking in April 2022. In contrast, in China, prices have generally been decreasing from January 2006 to April 2022, though they have remained relatively stable over the years. The United Kingdom's electrical equipment prices have no further records after the end of 2020. Meanwhile, in Italy, France, and Germany, prices have experienced some increases but have generally remained stable overall.

The **production index for machinery and equipment** in France declined in 2010 and never recovered. Globally, the production index has generally increased, though there was a noticeable dip at the end of 2009. In the UK, Japan, Italy, and Germany, the production index has remained relatively steady, whereas the US and Switzerland exhibit an irregular pattern over the years.

In the US, France, and Italy, the **electrical equipment production index** indicates a slight but noticeable drop. In contrast, Switzerland has experienced a boost, along with the worldwide production index. There is no apparent pattern for the UK, Japan, or Germany because their output indices do not differ much.

To gain a deeper view on how production trends impact pricing, we compared the **producer prices for electrical equipment with the production index**, a number of trends become evident. Producer prices typically either rise or stay the same, with the majority exhibiting growth, especially in the US, where the increase is most noticeable. Despite a declining production index, producer prices are comparatively constant in France and Italy. In contrast, both the production index and producer prices have increased in the UK and Germany. A clear pattern can be observed: when the **production index rises**, **prices tend to increase** as well. However, the reverse is not always true, as seen in the cases of the US, Italy, and France.

DATA PREPARATION

"While a lot of low-quality information is available in various data sources and on the Web, many organizations or companies are interested in how to transform the data into cleaned forms which can be used for high-profit purposes." (Zhang, Zhang, & Yang, 2003). Effective **data preparation techniques**, such as handling missing values, removing inconsistencies, and normalizing data, play a crucial role in ensuring the reliability and accuracy of analytical insights.

In the **sales dataset**, we start by optimizing the data types and creating new features to analyse the overall sales trends and changes for each product. We identified missing values for products 14 and 20, which are assumed to represent months with no sales activity. These missing entries were imputed with zeros to accurately reflect the absence of sales during those periods.

After that, we brought in external data, a CSV file with Germany's **Consumer Price Index (CPI)**, and merged it with our sales dataset. From there, we adjusted each product's sales to account for inflation and made the analysis more accurate over time. We repeated the stationarity analysis on the CPI-adjusted series and found no significant changes in the results.

In the **market dataset**, no errors or outliers were identified. However, some variables contained missing values, particularly between 2004 and 2006. Since our sales dataset only includes data from October 1, 2018, onward, keeping market data dating back to 2004 would not be relevant for our analysis and we started only to analyse our data after October of 2018.

To fill in the **missing values**, we calculate the monthly growth rate of each variable by computing the percentage change for each month. Next, we calculate the average of these monthly growth rates. Using this average growth rate, we can fill in any missing values by multiplying the value from the prior

month by (1 + average growth rate). This method allows us to estimate the missing value based on the observed trend in the data.

There were two variables from the United Kingdom, 'MAB_ELE_SHP826' and 'PRI27826_org', which had 41.86% missing values. Since this percentage was high (above 30%), we decided to drop these columns from our analysis.

Following the previously mentioned modifications, we used the date as the index to combine the two datasets because it was the common variable. Our objective with this connection was to **increase the accuracy of sales forecasting** for the products in the sales dataset by using market variables for feature selection. By combining both datasets, we ensured a **more comprehensive analysis**, capturing external factors that could influence sales trends.

In **feature selection**, we used 7 distinct methods: Pearson Correlation, Spearman Correlation, Random Forest, XGBoost, SHAP values, R^2 and RMSE to assess which features are most predictive of our target variable.

To capture temporal dependencies within our time series data, we engineered **lag-based features** that reflect past values of the target variables. Specifically, we constructed lag features (*e.g. P8_lag_6*) and corresponding moving averages (*e.g. P8_ma_6*) to help the model learn patterns across time. The best lag combinations were chosen based on statistical significance (p-values < 0.05) and model fit.

The correlations present on the data suggest that **products 1, 6 and 12** are strongly **US driven electrical equipment** demonstrating sensitivity to prices such as those of European goods, oil and other raw materials.

Regarding **products 5, 9 and 11** there is a clear dependency on **UK-based manufactured electrical equipment**, and also a weaker dependence on the world production and shipment index, namely that of China, Japan and the US.

Switzerland also plays a crucial role in the demand for **products 13, 16 and 36** given their high connection to its production index (in particular, that of electrical equipment). Moreover, with less strength we can associate their demand with the elasticity of the American and European markets.

The remaining products, i.e., **products 3, 4, 8, 14 and 20** are contingent on **electrical equipment supply**, specially from the **EU** (in particular, France and Italy) and **Japan**. Nevertheless, energy costs and raw materials also take part in the variation of sales of these products. Intuitively, these results are reasonable given that their production requires, from the start, a strong electrical and raw input driven reliance.

MODELLING

Our pipeline employs a walk-forward validation approach, in which the validation window is gradually advanced as we make predictions one step ahead of time. This guarantees that our model is consistently assessed using future data that has not yet been seen, closely resembling predicting scenarios found in the real world.

GridSearch is created using the PredefinedSplit method for hyperparameter optimisation purposes. This prevents incorrect cross-validation from disrupting the chronological order of the training and validation datasets.

Feature selection is crucial because each product has only 43 monthly data points, making the dataset small. Overfitting could result from having too many features since the model might record noise rather than actual patterns. We investigate the following feature combinations in an effort to reduce this: no macro/lags features (baseline model), single-feature models and multi-feature models.

We evaluate five forecasting models based on their scalability, interpretability, and efficacy in managing trends and seasonality (***Figure 5***).

Feature	TimeGPT (Zero-Shot)	XGBoost	TCN	ARIMA	Prophet (Facebook)
Scaling Needed?	❌ No	❌ No	✅ Yes	❌ No (if stationary)	❌ No (logistic/linear trends handle scaling internally)
Handles Trends	✅ Excellent	✅ (With features)	✅ Excellent	✅ (With differencing)	✅ Piecewise linear/logistic
Handles Seasonality	✅ Automatic	❌ (Requires manual lags)	✅ (With large receptive field)	✅ (SARIMA)	✅ Fourier terms (fixed but interpretable)
Multivariate Support	✅ Yes	✅ Yes	✅ Yes	❌ Univariate only	Limited
Training Speed	⚡ Instant (pre-trained)	🏎 Fast	🐢 Slow (GPU helps)	🚴 Moderate	🏎 Fast
Interpretability	❌ Black box	✅ Feature importances	❌ Black box	✅ Model coefficients	✅ Best-in-class (clear trend/seasonality plots)
Best For	Zero-shot forecasting	Tabular data with temporal features	Long-range dependencies	Simple univariate series	Business-friendly forecasts with explainability
Weakness	Limited control	Poor with long sequences	Computationally heavy	Linear assumptions only	Struggles with short-term dependencies (no AR terms)

Figure 5: *Time Series Model Comparison Table*

Our **evaluation** follows a sequential approach: TimeGPT as a baseline, XGBoost with lag features for feature engineering, TCN for deep learning insights, and Prophet & ARIMA for interpretability. This ensures a balanced comparison of accuracy, complexity, and explainability.

We evaluate using a step-by-step methodology: Prophet & ARIMA for interpretability, TCN for deep learning insights, XGBoost with lag features for feature engineering, and TimeGPT as a baseline. This guarantees an accurate comparison of explainability, complexity, and accuracy.

The final choice of modelling algorithm to use in each product was based on the RMSE and Overfit score, with the respective selection of features.

EVALUATION

Table 1: *Best Model's Per Product Evaluation*

Product	Model	RMSE	Overfit_Score
P1	ARIMA	2077977.613	-0.9453
P3	XGBoost	1142677.479	-0.1396
P4	ARIMA	66180.87	-0.5615
P5	XGBoost	1331522.065	-0.5324
P6	ARIMA	178899.4184	-0.7267
P8	XGBoost	108604.7529	6.7598
P9	XGBoost	4008.9303	0.0336
P11	XGBoost	626291.7346	-0.1633
P12	ARIMA	100594.6894	-0.0542
P13	XGBoost	10281.0456	-0.0769
P14	Prophet	10862.7748	-0.2337
P16	XGBoost	56763.369	4.0353
P20	XGBoost	685.6289	-0.1966
P36	XGBoost	17449.4	-0.3049
Sales_Total	XGBoost	3708435.169	-0.2707

RESULTS EVALUATION

On the final forecasting for the months of May 2022 to February 2023, we achieved the following results:

The sales of **product group 1** are expected to grow nearly 14% in 2022, when compared with 2021 results. The 10 predicted months sales average to a total amount of **43 019 464€**, which represents a 16% increase against the actual sales from the same period, starting in 2021. In the first 2 months of 2022, the year-to-date growth is expected to increase in 12% compared to 2022.

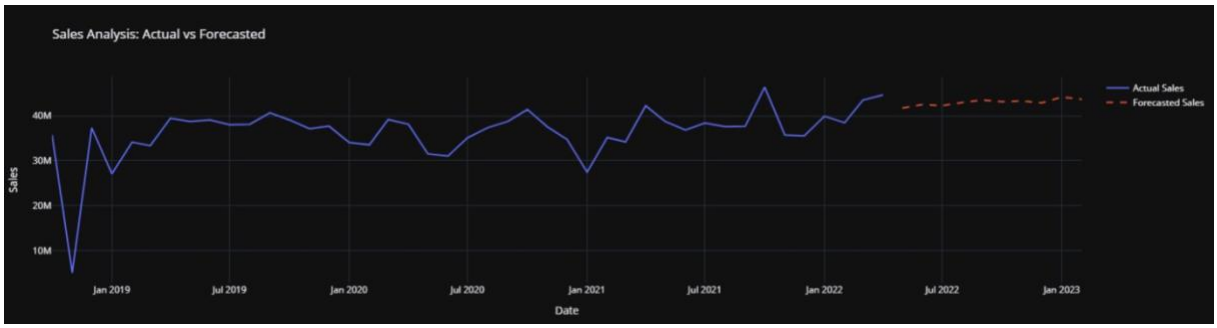

Figure 6: *Sale Analysis: Actual vs Forecast – P1*

Product group 3 sales are expected to grow by about 8% in 2022 compared to 2021. The 10 predicted months' sales average a total of **14 858 782 €**, representing a 10.26% increase over actual sales for the same period in 2021. In the first two months of 2023, year-to-date growth is projected to rise by 3.62% compared to 2022.

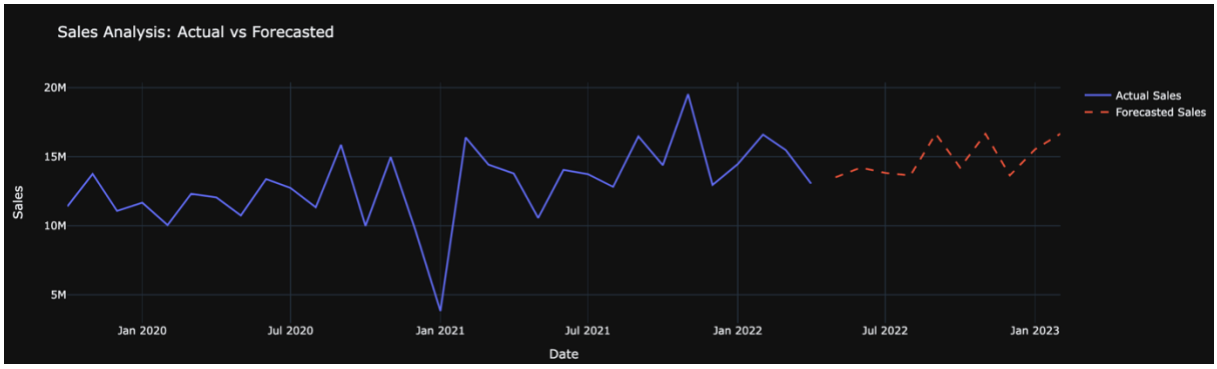

Figure 7: *Sale Analysis: Actual vs Forecast – P3*

The sales of **product group 4** are expected to decrease almost 22%, in 2022, when compared with 2021 results, averaging a total amount of **339 579 €**. The 10 predicted months' sales reflect a decline of 21.96% against the same period in 2021. However, in the first two months of 2023, year-to-date growth is expected to rebound by 23.17% compared to 2022.

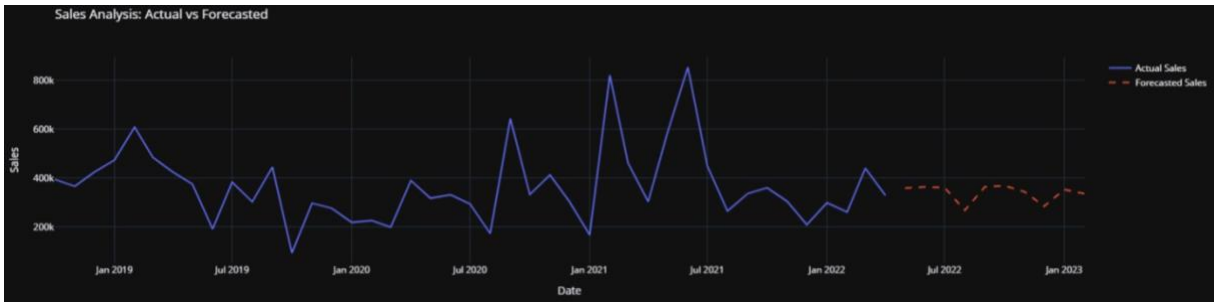

Figure 8: *Sale Analysis: Actual vs Forecast – P4*

Comparing 2022 sales to 2021 results, it is anticipated that product **group 5** sales will drop by over 3%, averaging **10 834 543 €.** When compared to the same period in 2021, the sales for the ten anticipated months show a further drop of 6.16%. Nonetheless, year-to-date growth is anticipated to increase by 26.65% in the first two months of 2023 in comparison to 2022.

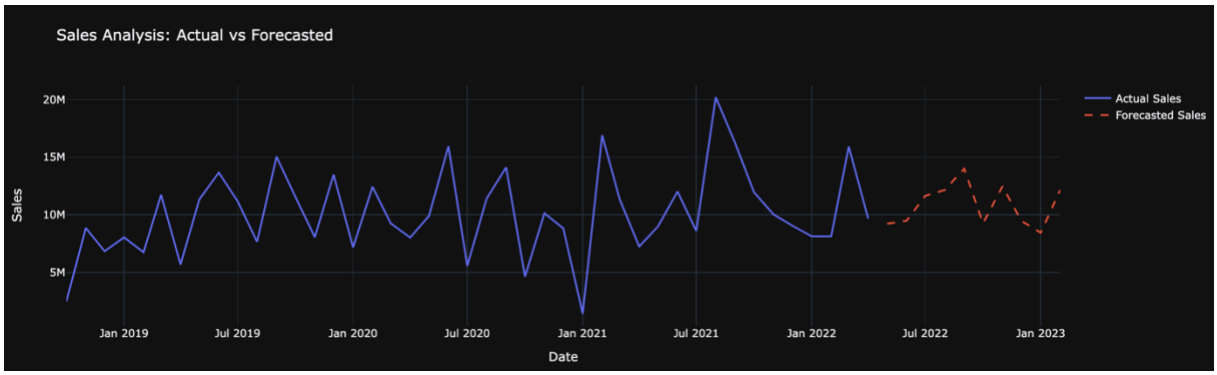

Figure 9: *Sale Analysis: Actual vs Forecast – P5*

Sales for product **group 6** are projected to drop by over 40% in 2022 compared to 2021, averaging **259 748 €.** In the following 10 months, sales are expected to decline by 42.20%, with an additional year-to-date decrease of 21.54% in early 2023 compared to 2022.

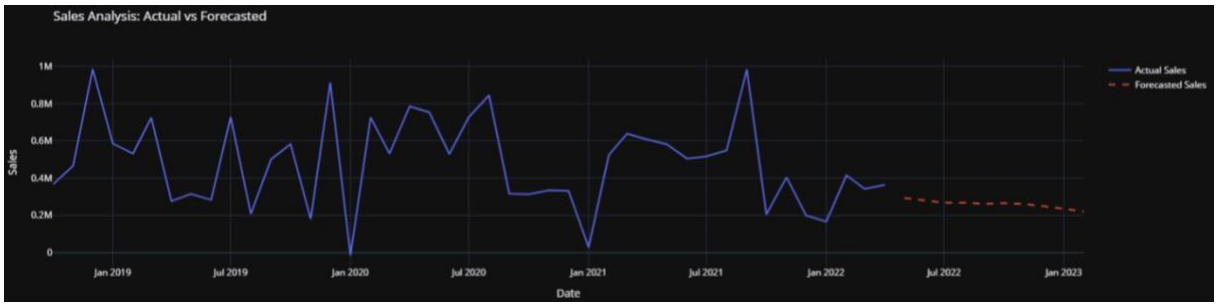

Figure 10: *Sale Analysis: Actual vs Forecast – P6*

When compared to 2021 figures, product **group 8** sales are anticipated to increase by around 24% in 2022, reaching an average of **1 532 561 €.** Over the next 10 months, sales are expected to grow by 20.24%, with an additional year-to-date increase of 19.31% in early 2023 compared to 2022.

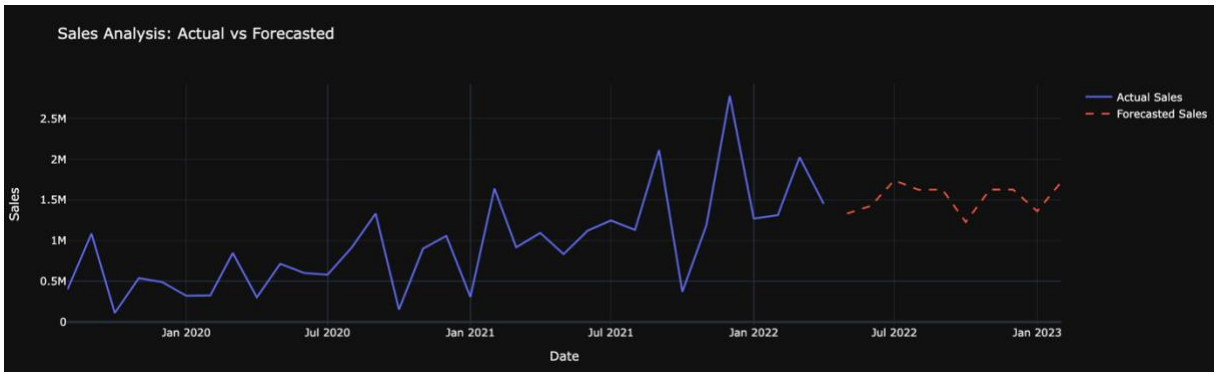

Figure 11: *Sale Analysis: Actual vs Forecast – P8*

Product **group 9** sales are predicted to drop by almost 22% in 2022 compared to 2021 results, with an average total of **10 777 €.** However, over the next 10 months, sales are forecasted to grow by 35.53%, with an impressive year-to-date increase of 595.21% in early 2023 compared to 2022.

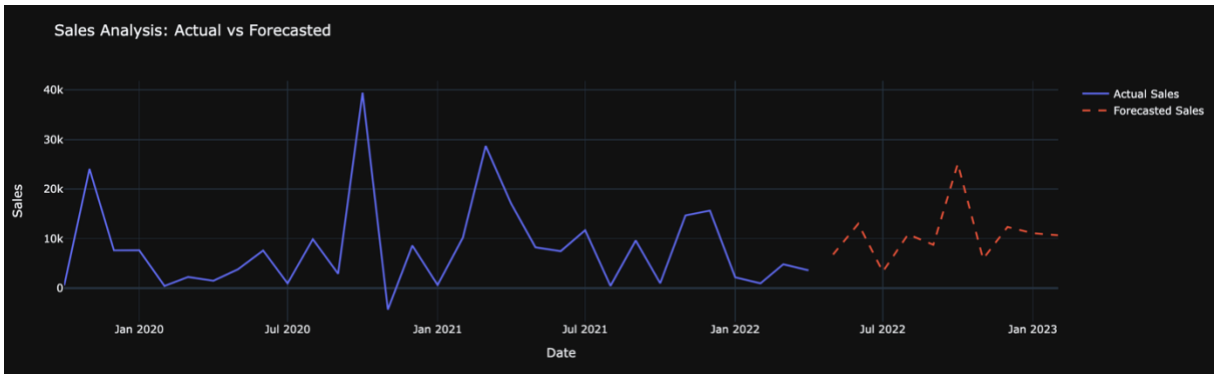

Figure 12: *Sale Analysis: Actual vs Forecast – P9*

The sales of **product group 11** are expected to decrease about 8% in 2022, when compared with 2021 results, averaging a total amount of **1 722 006 €**. The next 10 months are forecasted to see a further decline of 9.57%, but year-to-date growth in early 2023 is projected to rise by 33.95% compared to 2022.

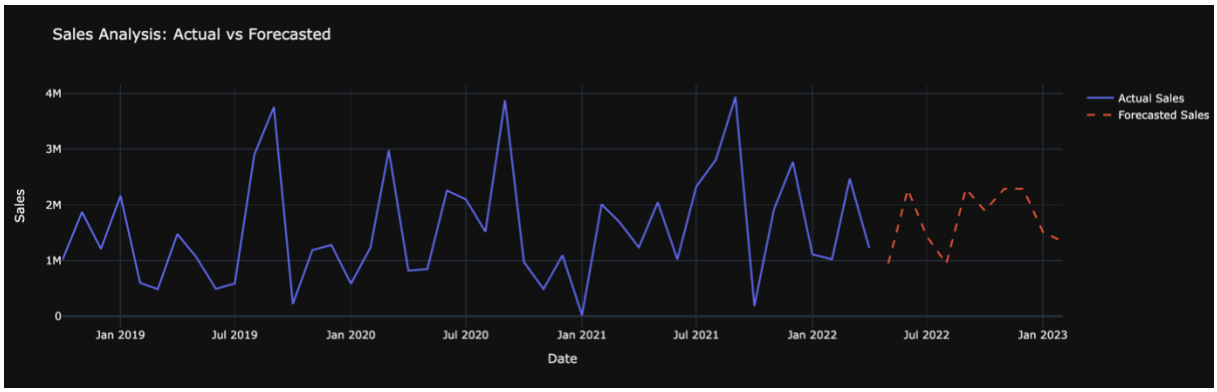

Figure 13: *Sale Analysis: Actual vs Forecast – P11*

When compared to 2021 statistics, **product group 12** sales are predicted to drop by 11% in 2022, averaging a total of **281 314 €**. Over the next 10 months, sales are expected to decline by 8.93%, but year-to-date growth in early 2023 is projected to increase by 6.31% compared to 2022.

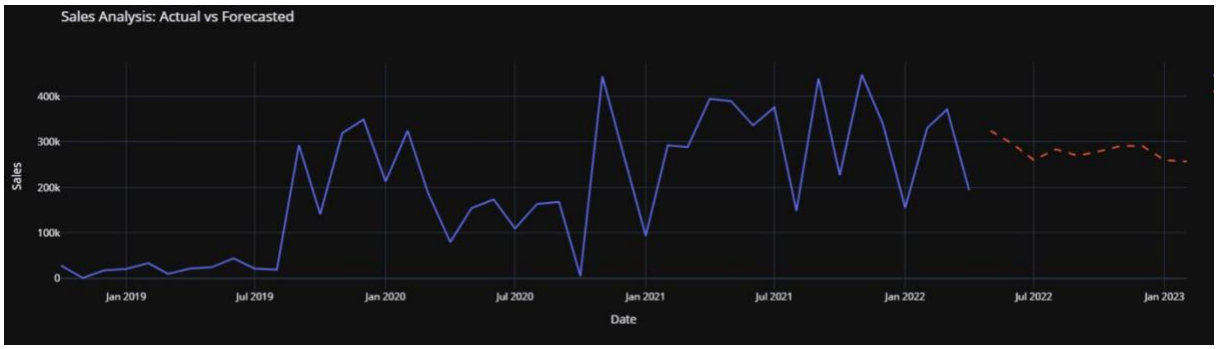

Figure 14: *Sale Analysis: Actual vs Forecast – P12*

The sales of **product group 13** are expected to decrease by nearly 1% in 2022 compared to 2021 results, averaging a total of **18 352 €**. Over the next 10 months, sales are forecasted to grow by 19.43%, but year-to-date growth in early 2023 shows a decline of 52.65% compared to 2022.

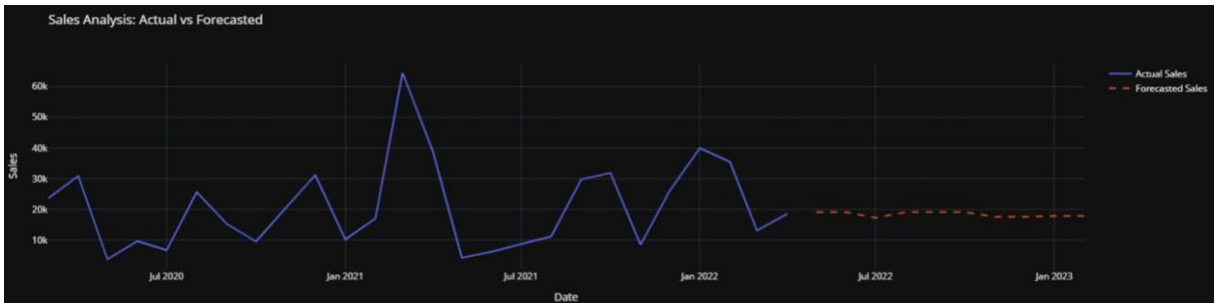

Figure 15: *Sale Analysis: Actual vs Forecast – P13*

It is anticipated that sales of **product group 14** would rise by almost 42% in 2022 over 2021, averaging **19 182 €.** Sales are expected to increase by 40.32% over the following ten months, with an outstanding year-to-date gain of 236.87% in early 2023 compared to 2022.

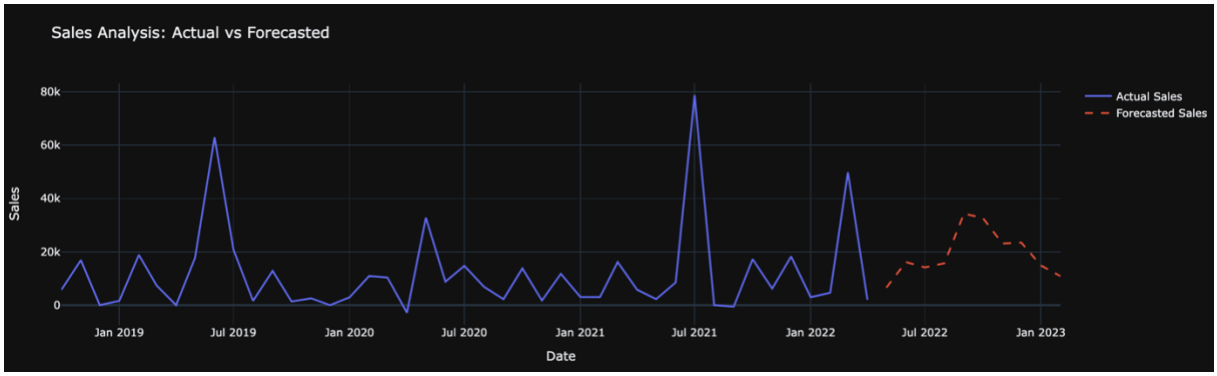

Figure 16: *Sale Analysis: Actual vs Forecast – P14*

The sales of **product group 16** are expected to decrease about 42% in 2022, when compared with 2021 results, averaging a total amount of **120 845 €**. Looking ahead, sales are forecasted to decline by 35.84% over the next 10 months, though year-to-date growth in early 2023 shows a notable increase of 82.29% compared to the same period in 2022.

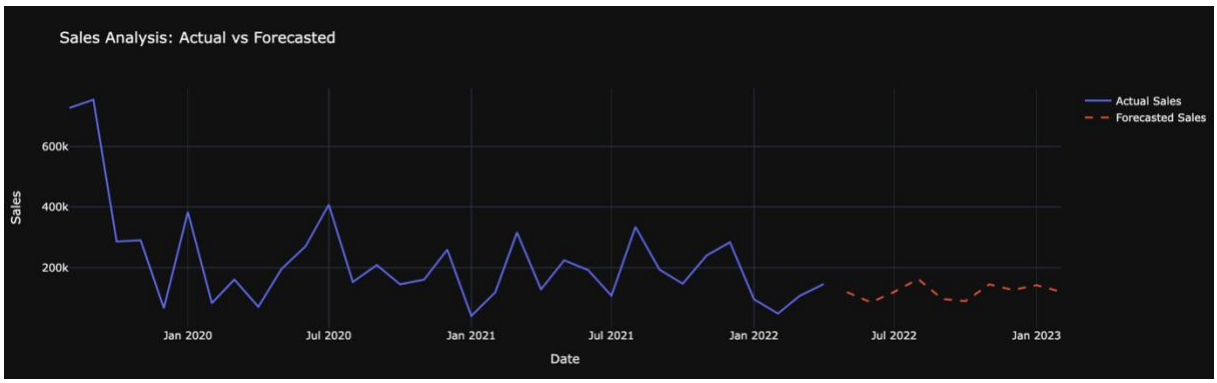

Figure 17: *Sale Analysis: Actual vs Forecast – P16*

Product group 20 sales are projected to grow by around 81% in 2022 compared to 2021, averaging **1 919 €**. The forecast for the next 10 months indicates a strong growth of 209.06%, with year-to-date growth in early 2023 up by 59.31% compared to the same period in 2022.

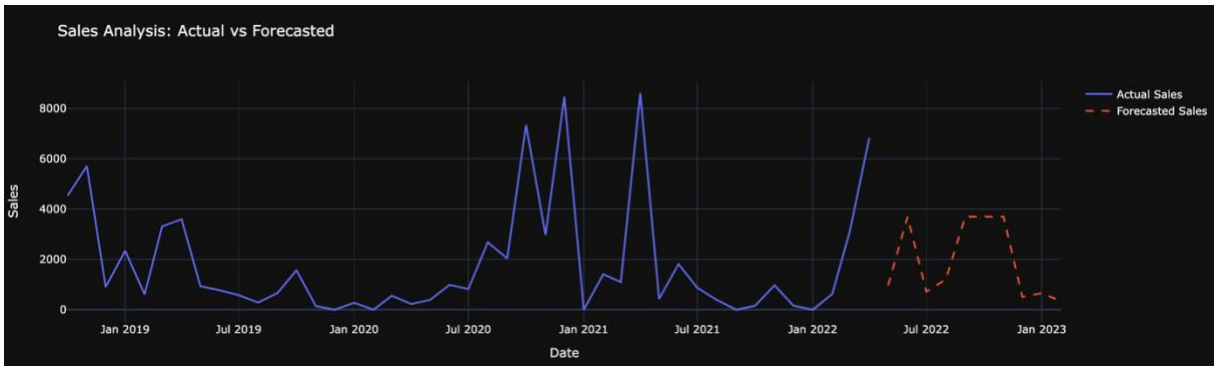

Figure 18: *Sale Analysis: Actual vs Forecast – P20*

The sales of **product group 36** are expected to decrease about 53% in 2022, when compared with 2021 results, averaging a total amount of **17 576 €**. Looking ahead, the next 10 months are expected to see a growth of 6.19%, with year-to-date growth in 2023 down by 17.82% compared to the same period in 2022.

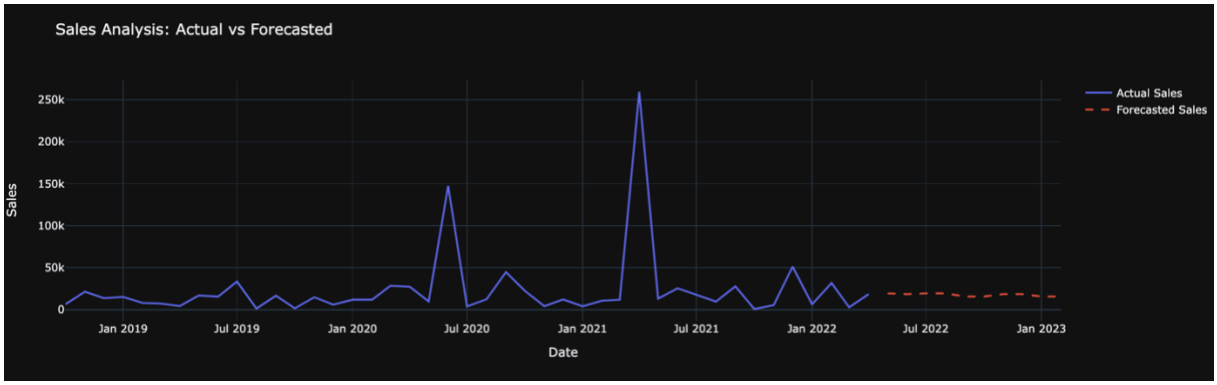

Figure 19: *Sale Analysis: Actual vs Forecast – P36*

DEPLOYMENT AND MAINTENANCE PLANS

Successful deployment and long-term value delivery in data-driven projects rely heavily on structured planning and proactive maintenance. While the technical implementation is crucial, the real challenge lies in ensuring that solutions remain reliable, scalable, and aligned with evolving business needs. The *Bull Survey (Spikes Cavell, 1998)* found that **57% of IT project failures** stem from communication failures, followed by **39% due to inadequate planning** and **35% due to poor quality control**. More recently, *Gartner (2019)* reported that **only 20% of analytic insights** effectively translate into business outcomes.

To mitigate these risks, this section outlines a clear and practical approach for deploying our solution effectively, along with the mechanisms in place to monitor, update, and maintain its performance over time.

PROPOSED NEXT PHASES

- **Next Milestone: Validation & Feature Enhancement**

Once approval is secured, the next phase will focus on refining the forecasting approach and enhancing its practical value for Siemens' commercial planning through the **development of an integrated platform**. This platform will embed the forecasting model and connect directly with the company's internal systems to calculate monthly sales in real time, monitor goal achievement, and assess stock levels by linking to the logistics system. This will enable more dynamic and informed decision-making across commercial and operational areas, including inventory management, demand planning, and resource allocation.

Furthermore, the platform will be designed to issue alerts in strategic scenarios—for instance, when a product shows a strong positive correlation with a macroeconomic variable, and that macro begins to shift significantly (either increasing or decreasing), users will be automatically notified. This ensures that commercial and supply chain teams are proactively informed of market movements that could impact demand, allowing them to adjust strategies accordingly and stay aligned with evolving business objectives.

Additionally, qualitative insights such as client feedback or satisfaction indicators may be explored to uncover patterns that could further inform demand forecasting and align commercial strategies with client expectations.

- **Continuous Improvement & Model Updates**

To maintain accuracy and relevance, the sales forecasting models are regularly updated to incorporate new data and reflect changing market dynamics and sales behaviour. Key variables, such as sales and macroeconomic indicators are systematically stored to support continuous improvement and long-term model enhancement. This ensures robust, efficient analysis and empowers teams with up-to-date insights to support strategic sales and operational decisions.

Figure 20: *Deployment Plan*

INTERNAL DATA ANALYSIS SOLUTION: A COST-EFFECTIVE ALTERNATIVE

Our solution provides an interactive, user-friendly, and cost-efficient alternative to traditional forecasting platforms, with a more personalized approach tailored to Siemens' specific needs.

Key Features of Our Solution

Table 2. *Internal data analysis solution*

Feature	Description
Advanced Forecasting Models	The project applies cutting-edge models like TimeGPT, XGBoost, Arima and NeuralProphet tailored for time series prediction with high accuracy and flexibility across product lines.

Proactive Alerting System	Automatically notifies users when relevant market indicators (e.g., macroeconomic variables) shift, enabling agile and informed responses to changing external conditions.
Strategic Decision Support	Forecast results directly support demand planning, inventory management, and resource allocation—strengthening cross-functional decision-making.
Continuous Learning & Model Updating	The model architecture supports frequent retraining with new data, ensuring forecasts remain accurate as business conditions evolve.

Cost Comparison

Considering the average monthly salary of a data consultant is €2,824, and the project took 3 weeks to develop, the equivalent cost per person is approximately €2,118. Since the team consisted of 4 members, the total estimated cost amounts to **€2,118 × 4 = €8,472**.

Table 3. *Cost Comparison*

Our Sales Forecasting	Workday Adaptive Planning	Aviso Predict
8,472€	Starts at around USD 15,000€	75€ per user per month

Why Choose Our Solution?

For companies seeking a powerful yet cost-effective forecasting tool, our solution offers the ideal balance between accuracy, usability, and integration. We will provide Siemens with an adaptable and ready-to-use platform that supports commercial planning.

- **Cost Savings**: No ongoing licensing or per-user fees, making it a sustainable internal alternative to external forecasting platforms.
- **Greater Flexibility**: The forecasting model can be continuously improved based on business feedback and evolving commercial needs.
- **Predictive Capabilities**: Includes built-in time series forecasting and macroeconomic correlation analysis without reliance on third-party tools.
- **Collaboration**: Commercial and operations teams can interact with real-time data, receive automated alerts, and adjust strategies together within the same environment.

CONCLUSIONS

This project successfully delivered a robust, insightful, and cost-effective sales forecasting solution tailored to Siemens' Smart Infrastructure division in Germany. By integrating historical sales data with relevant macroeconomic indicators, our model supports accurate monthly forecasts and empowers more informed strategic planning across commercial and operational teams.

Over the course of three weeks, and following the CRISP-DM methodology, our team explored the business needs and data insights, applied various feature selection techniques, and tested four forecasting models: ARIMA, Prophet, XGBoost, and TimeGPT. The final approach included product-level analysis, lag feature engineering, and macroeconomic trend alignment, ensuring that the model is not only accurate but also interpretable and adaptable to future changes.

Through this analysis, we identified product-specific patterns such as yearly or semi-annual seasonality, macro dependencies across global markets, and financial vulnerabilities linked to high-value returns. These insights, paired with our model's predictive power, provide Siemens with a

valuable tool for anticipating sales behaviour and optimizing decisions related to demand planning, inventory, and resource allocation.

REFERENCES

[1] Siemens. (2018). Annual report 2018. Siemens AG. https://www.siemens.com/global/en/company/investor-relations/events-publications-ad-hoc/annualreports.html

[2] Siemens. (2019). Annual report 2019. Siemens AG. https://www.siemens.com/global/en/company/investor-relations/events-publications-ad-hoc/annualreports.html

[3] Siemens. (2020). Annual report 2020. Siemens AG. https://www.siemens.com/global/en/company/investor-relations/events-publications-ad-hoc/annualreports.html

[4] Siemens. (2021). Annual report 2021. Siemens AG. https://www.siemens.com/global/en/company/investor-relations/events-publications-ad-hoc/annualreports.html

[5] Siemens. (2022). Annual report 2022. Siemens AG. https://www.siemens.com/global/en/company/investor-relations/events-publications-ad-hoc/annualreports.html

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. Applied Artificial Intelligence, 17(5-6), 375–381. https://doi.org/10.1080/713827180

Key enabling technologies. (n.d.). Siemens.com Global Website. https://www.siemens.com/global/en/company/about/strategy/siemens-megatrends/key-enabling-technologies.html

6 Shocking Statistics About Sales Forecasting: Argano. (2025). Argano.com. https://argano.com/insights/articles/6-shocking-statistics-about-sales-forecasting.html

Salesblink. (n.d.). *15 Best Sales Forecasting Tools To Use In 2024*. Retrieved April 1, 2025, from https://salesblink.io/blog/best-sales-forecasting-tools

Xactly. (n.d.). *Sales forecasting: What it is and why it's important*. Retrieved April 1, 2025, from https://www.xactlycorp.com/blog/forecasting/sales-forecasting-what-it-and-why-its-important

APPENDIX

GLOSSARY

- **CRISP-DM (Cross Industry Standard Process for Data Mining):** A structured methodology for data mining projects, consisting of six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment.
- **Stationary:** A variable is considered stationary when its behaviour remains consistent over time—meaning its average, variance (level of fluctuation), and overall pattern do not change. In practical terms, a stationary variable doesn't show long-term trends or increasing volatility, making it easier to forecast and analyse. Stationary data is predictable and stable, which is ideal for building reliable forecasting models.
- **Autocorrelation Function (ACF):** The ACF measures the correlation between a time series and its past values across different time lags. It helps identify the presence of patterns or randomness in the data and is useful for determining the *q* parameter in Moving Average (MA) models. High ACF values at specific lags indicate repeated patterns, while values close to zero suggest a lack of correlation.
- **Partial Autocorrelation Function (PACF):** The PACF measures the correlation between a time series and its past values, removing the effects of intermediate lags. It is mainly used to determine the *p* parameter in Autoregressive (AR) models. Unlike ACF, PACF focuses on the direct relationship with lagged values, making it helpful for identifying the true order of an AR process.
- **Lags:** Refers to the delay between an observed data point and its preceding values. Specifically, it is the time difference between two observations in a sequence, or the number of steps back in time a past observation is from the current time.
- **Root Mean Square Error (RMSE)**: RMSE measures the average difference between a statistical model's predicted values and the actual values. Mathematically, it is the standard deviation of the residuals. Residuals represent the distance between the regression line and the data points. RMSE quantifies how dispersed these residuals are, revealing how tightly the observed data clusters around the predicted values.
- **Overfit Score**: Measures how well the model generalizes to new data, beyond the dataset it was trained on.