NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 3: Siemens Advanta – Sales Forecast

Carolina, Caldeira, number: 20230440
Goncalo, Caldeirinha, number: 20230469
Madalena, Figueirinhas, number: 20230436
Martyna, Kmiecik, number: 20230452
Alina, Metzger, number: 20230998

Group C

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

April, 2024

INDEX

1 EXECUTIVE SUMMARY 2
2 BUSINESS NEEDS AND REQUIRED OUTCOME 3
2.1 Background 3
2.2 Business Objectives 3
2.3 Business Success Criteria 3
2.4 Situation Assessment 3
2.5 Determine Data Mining Goals 4
3 METHODOLOGY 4
3.1 Data Understanding 4
3.2 Data Preparation 6
3.3 Modeling 8
3.4 Results Evaluation and Interpretation 11
4 EVALUATION 14
5 DEPLOYMENT AND MAINTENANCE PLANS 15
6 CONCLUSIONS 16
6.1 Considerations for Model Improvement 17
7 REFERENCES 18
8 APPENDIX 19

LIST OF TABLES

Table 1: Market Indexes Related with each Product Group 8
Table 2: Model Performance per Product Group 11
Table 3: Glossary 19

LIST OF FIGURES

Figure 1: Total Sales by Product Group & Figure 2: Aggregated Sales over Time 5
Figure 3: Total Sales per Month 6
Figure 4: Sales Forecast for Product Groups 1, 3, 9 and 11 12
Figure 5: Sales Forecast for Product Groups 5 and 8 13
Figure 6: Sales Forecast for Product Group 13 and 16 13
Figure 7: Sales Forecast for Product Group 6 and 4 13
Figure 8: Sales Forecast for Product Group 36 14
Figure 9: Project Plan 16

1 EXECUTIVE SUMMARY

This report presents the development and application of a monthly sales forecasting model for selected product groups within the Smart Infrastructure Division in Germany. The initiative, carried out by a collaborative team from Siemens and Nova IMS, leverages advanced data science techniques to enhance inventory management, operational efficiency, and customer satisfaction through precise sales volume predictions from May 2022 to February 2023.

Employing the CRISP-DM methodology, the project involved meticulous phases of data understanding, preparation, and modeling. The data spanned sales records from October 2018 to April 2022, enriched with relevant macroeconomic indices to provide a robust basis for analysis. Advanced statistical and machine learning techniques—including Prophet, XGBoost, SARIMAX, and Linear Regression—were applied to develop models capable of accurately forecasting future sales volumes.

The project's success was assessed through several metrics, primarily the accuracy of the sales forecasts, as indicated by RMSE and MAPE scores. These models have enabled Siemens to refine their inventory and supply chain management strategies, becoming more responsive to market demands and fluctuations.

The results demonstrated varying levels of forecast accuracy across the product groups, with specific models excelling in particular contexts. This variance underscores the need for a tailored approach in predictive modeling to capture the unique sales patterns and market dynamics for each product group.

To enhance the strategic impact of our forecasting models, we recommend regular updates and recalibration of the models to maintain their accuracy as market conditions evolve. Furthermore, leveraging up-to-date data can improve the responsiveness of the forecasting models. Developing specific inventory strategies for each product group based on their unique sales patterns and forecast reliability can also improve business performance by enhancing customer satisfaction and optimizing resource allocation.

These strategies aim not only to improve forecast accuracy but also to boost overall business performance by enhancing customer satisfaction and optimizing resource allocation. By continuing to integrate advanced analytics into operational strategies, Siemens can sustain its competitive advantage in a dynamic market landscape.

2 BUSINESS NEEDS AND REQUIRED OUTCOME

2.1 BACKGROUND

Siemens AG is a diversified corporation operating globally across various industrial sectors, including digital industries, smart infrastructure, mobility, and medical engineering. Within the group, the Smart Infrastructure Division holds a vital position, focusing on providing innovative solutions for sustainable and efficient buildings, smart grid technology, and intelligent power distribution. In this context, the present business case aims to delve into the sales dynamics of selected product groups (PGs)[1] of one business unit within the Smart Infrastructure Division, particularly focusing on the German market, which stands as the largest market for the respective business unit.

2.2 BUSINESS OBJECTIVES

The core objective is to develop a monthly sales forecasting model for designated PGs of a business unit within the Smart Infrastructure Division in Germany for the months of May 2022 to February 2023. This includes leveraging data-driven methodologies, particularly machine learning and statistical techniques, to accurately predict monthly sales volumes for the defined period. By achieving this, Siemens aims to optimize inventory management, enhance operational efficiency, and ensure better alignment between supply and demand, ultimately contributing to improved business performance and customer satisfaction.

2.3 BUSINESS SUCCESS CRITERIA

The success of the project will be evaluated based on predefined criteria, primarily focusing on the accuracy and reliability of the sales forecasts generated. Specifically, the business criteria include:

- Accuracy of sales forecasts: The developed models should provide accurate predictions of monthly sales volumes, minimizing errors and discrepancies.
- Reduction in resource intensity: The approach should lead to a significant reduction in the resources (time and personnel) required for manual forecasting processes, thereby improving operational efficiency.
- Mitigation of bias: Leveraging data-driven methodologies, should mitigate biases inherent in manual forecasting, ensuring more objective and unbiased predictions.
- Integration of scattered information: By integrating data from multiple sources and aligning them with ever-changing market dynamics, a comprehensive understanding of sales trends should be facilitated.
- Reduction in opportunity costs: The implementation of AI-driven sales forecasting is expected to mitigate the opportunity costs associated with poor forecasting, particularly in terms of working capital management and customer satisfaction.

2.4 SITUATION ASSESSMENT

Regarding resources, a comprehensive dataset comprising daily sales records per PG from October 2018 to April 2022, coupled with important macroeconomic indices relevant to Siemens' operations in Germany, will serve as the primary data sources. A multidisciplinary team consisting of domain experts

[1] Product group is being referred to as “PG” from here on.

from the respective business unit of the Smart Infrastructure Division, sales analysts, and IT professionals from the Siemens side and five data scientists from the Nova IMS side collaborate closely on the project. The data is managed using Jupyter Notebook, a web-based interactive computing platform, capable of handling complex data mining tasks, along with data visualization tools.

Potential risks include data quality issues, model overfitting (poor performance on new data), and changes in market dynamics. To mitigate these risks, contingency plans include robust data preprocessing techniques, regularization methods, and ongoing model monitoring and recalibration.

As for terminology, it is essential that stakeholders understand key project terms, such as the evaluation metrics 'Root Mean Square Error (RMSE)' and 'Mean Absolute Percentage Error (MAPE)' as well as the machine learning algorithms 'Prophet', 'XGBoost', 'SARIMAX', and 'Linear Regression'. A glossary of these terms will be provided in the appendix for clarity and alignment.

While the project entails initial investment in resources and technology, the expected benefits include improved forecasting accuracy, reduced operational costs, and enhanced customer satisfaction, leading to better working capital management and increased revenue potential for Siemens in the German market.

2.5 Determine Data Mining Goals

The technical objective is to develop predictive models capable of forecasting monthly sales volumes for selected PGs within the Smart Infrastructure Division in Germany. This involves three main phases: data understanding and preparation; model selection and assessment; and result evaluation and interpretation. The ultimate goal is to deploy a scalable and adaptable solution that can support informed decision-making and drive operational excellence within Siemens' sales processes.

The performance of the models will be evaluated by the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), which are metrics used to assess the accuracy and precision of sales forecasts. The success criteria will be achieved if the developed models demonstrate a RMSE and MAPE values within acceptable thresholds, indicating accurate and reliable sales forecasts. Specifically, the aim is to minimize the values, ensuring that the predicted sales volumes closely align with actual sales data. Additionally, the models should be capable of adapting to changing market conditions and dynamics, thereby enhancing their predictive capabilities over time.

3 METHODOLOGY

3.1 Data Understanding

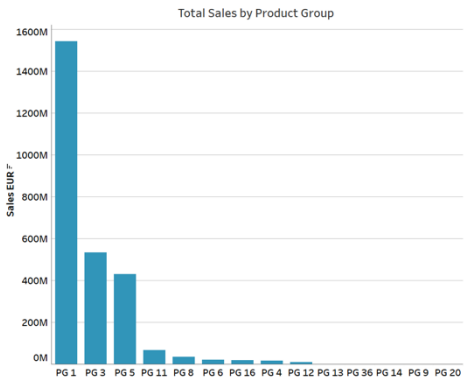
In the critical domain of sales forecasting for Siemen's Smart Infrastructure Division in Germany, comprehending the complexities of the datasets is crucial to navigating the complex landscape of business analyses. This understanding not only helps in crafting predictive models but also in fine-tuning strategic business decisions. The dataset provided offers a glimpse of the dynamic interaction between product sales and economic indicators.

The sales data is concise yet detailed in crucial details, it records the transactions between October 2018 and April 2022. This data set is pivotal for analyzing sales trends, assessing product performance, and forecasting future sales volumes within the diverse products. It tracks:

- **Date of transaction:** Offering a temporal snapshot of sales.
- **Mapped GCK (Generalized Commodity Key):** This identifies the PG, providing a link to the specific items being sold.
- **Sales in EUR:** Quantifying the revenue generated from each sale, thus serving as the primary output variable for forecasting models.

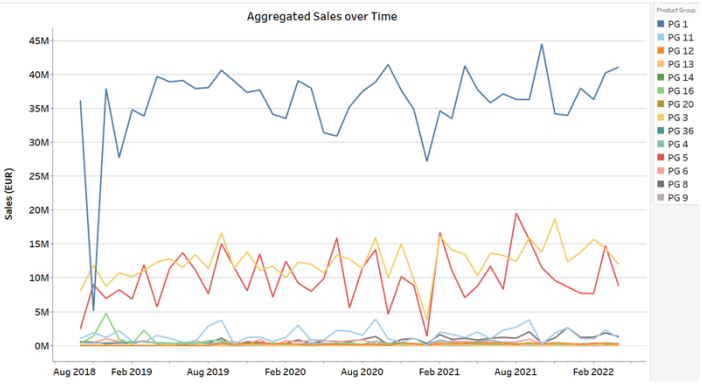
To gain a comprehensive understanding of the products, the objective was to examine both the sales of each product over the months, as well as the total sales of these products, thereby providing a more in-depth preview of their performance. It is important to mention that PG 1, 3, and 5 had much higher total sales compared to the other groups, as illustrated in figure 1.

Figure 1: Total Sales by Product Group



Source: Own elaboration based on sales data.

Figure 2: Aggregated Sales over Time

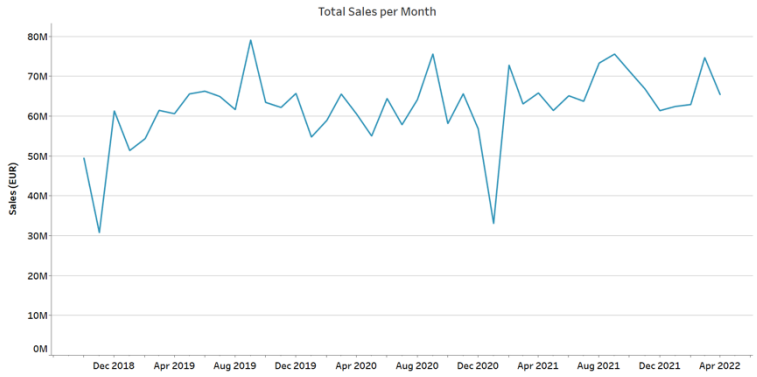


Source: Own elaboration based on sales data.

It should also be noted that a more in-depth analysis of product sales data over time (see figure 2) reveals not just trends, but also peculiar anomalies and irregularities in the data. For instance, it is notable that certain PGs show unexpected peaks or dips in sales. These deviations from the norm may correspond to external market forces, operational changes or other factors.

Advancing beyond the scope of individual product analyses, the synthesis of monthly sales data encapsulates the broader sales trends across all PGs. It became clear that there exists month-to-month variance, revealing fluctuations in market demand and seasonal patterns. For instance, a significant peak was observed in the early months of 2019, subsequently followed by a decline and a period of stability. However, the onset of 2020 brought about a noteworthy downturn, which can be attributed in big parts to the global impact of the Covid-19 pandemic. The global health crisis disrupted supply chains, altered consumer spending habits, and led to lockdowns and social distancing measures, all of which significantly affected sales across various sectors. As the sales progress into 2021, an observed pattern of fluctuation becomes particularly important to consider in the context of macroeconomic conditions provided in the market data. From this one may also conclude that the products mentioned above have the same pattern as the total monthly sales.

Figure 3: Total Sales per Month



Source: Own elaboration based on sales data.

The market data is comprehensive, capturing key economic indices from influential countries namely China, France, Germany, Italy, Japan, Switzerland, the United Kingdom, and the United States. These indices can be divided into macro- and microeconomic indicators.

Macroeconomic indicators:

- **Production and shipments indices:** These indices give insights into the manufacturing output and distribution volume especially when covering categories like machinery and electricals, reflecting the economy of the nations and industry-wide scale.
- **Commodity prices:** These include base metals, energy, metals & minerals, and specific commodities like natural gas and crude oil. They mirror the market trends and economic conditions that can and will affect industries at the macro level.

Microeconomic indicators:

- **Producer prices:** Specific to electrical equipment, they provide insights into the costs of production for individual firms or industries, which can influence pricing strategies and competitive dynamics within the market.

Notably, there are instances of missing values and anomalies within the data, requiring careful attention and careful handling to ensure the integrity of the results. The methods deployed to address these gaps will be elaborated in subsequent sections.

3.2 Data Preparation

After understanding the data available and the complexity associated with dealing with time series, it is important to implement a detailed and thought-out preprocessing strategy, that addresses all issues. Since there are two data sets, the first to be treated was the Sales data. Firstly, it was confirmed that there was no more than one record per month, per PG, meaning that there were no duplicated records. Additionally, there were no missing values found in this data.

As this is a time series forecasting problem, it is important to check for stationarity.[2] This concept is associated with static statistical characteristics of the data through time, meaning that variances, means, and autocorrelation are maintained throughout the timeframe. Also, some models assume

[2] Cf. Brownlee, J. (2020).

stationarity in time-series and this characteristic facilitates prediction. After performing the test[3], it was concluded that only the sales time series for PG 8 was a non-stationary one. Some methods of handling non-stationarity were tested however, to simplify interpretability and the complexity of the model, this fact was ignored, and the results of the test were near the stationarity threshold value.

Afterwards, outliers were treated in a very delicate manner as there was a considerable amount in the data and there was not a substantial number of records. This step was included in the process due to the fact there were unique and extreme sales values for some PG that were assumed as not recurrent. Additionally, most outliers could represent trends in the data or for example the end of fiscal years for clients, which might lead to a need for Siemens products, translated into a high number of sales for that period. Also, some external factors could impact these sales leading to extreme values in some periods of sales. For this reason, it was decided to change these values which most likely did not represent patterns, to the second highest ones. This way the extreme values are still represented but with a smaller impact. This was done for PG 1, 3, 9 and 36. Through this process, there was a concern with keeping stationarity in the data as well. For PG 16, since the extreme values were only in the beginning, it decided to replace them with the average of this data to not influence forecasted values as much.

Afterward, feature engineering was done. Features with information about the month and quarter were added to highlight possible relationships between a specific time and sales behavior. Additionally, for each PG sales the following features were added: a feature with the percentage of growth for every month in relation to the previous ones, added to highlight these short-term changes in case they would be relevant; a *rolling-mean* feature, giving the average between the past 3 months of data, to smooth short-term fluctuations and highlight longer-term patterns; lastly, lagged features, which correspond to the sales but with specific lags – these were chosen looking at the sales autocorrelation, in a way to capture the immediate past's influence on the present.[4]

For the Market data, the first step was treating missing values. Distributions and the line plots were analyzed, making sure that after the changes the feature's statistical characteristics would not change. Imputation was made based on similarity with other records, using the common method KNN[5], and then distributions and plots were rechecked to make sure there were no drastic changes in the relevant period, the sales period - October 2018 until April 2022. After this stage only data from this period was considered as it was the only one for which there were sales, and, for that reason, it would not be useful to use past Market data information that would not be accounted for already in the features created. These features included, similarly to the Sales data, the percentage change, and the rolling mean for each.

Outliers were not treated because the impact of these variables would not be as important as the past sales information and it was decided to keep the variability in this data set, preserving external factors.

After preprocessing, the next stage was Feature Selection. For stage, 3 methods were chosen, according to what would be most suitable for this type of data. The first one was Recursive Feature

[3] Cf. Verma, Y. (2021).

[4] Cf. Agrawal, R. (2021).

[5] Cf. Brownlee, J. (2020).

Elimination,[6] which builds a model for forecasting sales with all features, sequentially removing the least important each repetition, until the defined number of features is reached; Mutual Information[7], which measures how much information the sales shares with other variables (the higher, the more important that variable is for the model); lastly, LassoCV[8] which is indicated for when there is a considerable amount of features but only a few are relevant for the prediction. After using all 3 methods and looking at the top 10 features for each PG, the features in common were the ones chosen.

After this selection, it was possible to identify that for all PG, short-term changes in sales were relevant as well as external factors, the Market information, which impacted the sales or at least behaved in similar ways. The relationships are illustrated in the following table:

Table 1: Market Indexes Related with each Product Group

PG	Related to
1	Switzerland Machinery and Electricals Production
3	Japan Machinery and Electricals Production
4	USA's Machinery and Electricals Production
5	Italy, Japan, USA, France Machinery and Electricals Shipping Great Britain Electrical Equip. Producer Price
6	Germany Electrical Equipment Production
8	World Natural Gas Price USA Electrical Equip. Producer Price
9	Germany, Japan Machinery and Electricals Production Great Britain and World's Electrical Equip. Production

PG	Related to
11	France Machinery and Electricals Shipping, Italy Electrical Equip. Production
12	Great Britain Electrical Equip. Production, USA EUR in LCU
13	China Electrical Equip. Producer Price
14	China Electrical Equip. Producer Price Japan Machinery and Electricals Production
16	$USA to EUR conversion rate
20	Great Britain Electrical Equip. Production World Electrical Equip. Production
36	Europe Machinery and Electricals Production Great Britain Electrical Equip. Producer Price

Source: Own elaboration based on market data.

After obtaining the features for each PG it was necessary to forecast their values for de period of May 2022 until February 2023. This is important since it was evidenced that there are variables impacting specific PG sales between 2018 and 2022, and so, it can be concluded that for the next period, they would also be relevant. Due to a lack of external data, it was decided to forecast these values as they would be needed for the predictions made afterward. This forecasting was made using the computational model Prophet (which will be explained further in this report) only taking into account the past information of each variable for their own predictions, making this a more straightforward and simple approach.

3.3 MODELING

To predict the future sales of each PG, different forecasting models were applied to each PG and afterwards evaluated based on the training and validation sets. The performance of the models was compared through the evaluation metric Root-Mean-Square Error (RMSE) and, in addition, the Mean Absolute Percentage Error (MAPE)[9].

A diverse set of models were developed to forecast future sales of each PG. Each modelling technique was selected for its unique strengths and applicability to the available dataset, which included not only

[6] Cf. Brownlee, J. (2020).
[7] Cf. Guhanesvar. (2021).
[8] Cf. Agrawal, S. (2023).
[9] Cf. Gong, L., & Shi, J. (2010).

historical sales from October 2018 to April 2022, allowing the models to identify patterns and trends in the data, but also additional sales-related features, as well as market indicators relevant to Germany. The selection of multiple models and the subsequent validation on their basis allowed for a more objective selection resulting in better forecast accuracy.

The models selected to obtain predictions included: Prophet, XGBoost, SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors) and Linear Regression. SARIMAX and Prophet are models common to time-series forecasting, unlike the other two, which, nevertheless, can be adapted for that and yield beneficial results as well.

Prophet follows a procedure for forecasting time series data based on an additive model in which nonlinear trends are adjusted for seasonality. It is also an open-source software tool delivered by Facebook platform. Prophet was selected as one of the training models for the Siemens problem due to its robustness in handling time series data with strong seasonal patterns and its flexibility in including other characteristics possibly related to the predicted variables. In addition, it was recognized for the simplicity of its application[10].

The second time series forecasting model used is SARIMAX, which is an extension of ARIMA, but also taking into account seasonality of variables, if there is any, and exogenous variables. This model was chosen, not only because of these characteristics, but also due to its broad popularity. While predicting, such a model takes into consideration past values (autoregression, moving average), seasonality patterns and external values and predicts future values based on that. The model handles the complexity of sales data particularly well, which can be influenced by external factors such as economic indicators or marketing activities, making it ideal for our dataset, which includes market indices[11].

Another one of the models adapted for the problem - XGBoost - is an optimized distributed gradient tree model designed to be highly efficient and flexible, handling classification and regression well. While it is not the traditional time series forecasting tool of choice, it was chosen for its powerful machine learning capabilities in data processing. By transforming time series data in a way that helps computers learn from examples, XGBoost can capture and predict the complex nonlinear relationships between predictors that vary in the developed models for different PGs, so this is a particularly valued asset of the model. In addition, the model is extremely fast, which is another considerable advantage[12].

Finally, linear regression has been leveraged for its simplicity and ease of interpretation, providing a baseline model against the performance of more complex models. It is particularly useful for understanding the direct impact of individual characteristics on sales, although it may not capture very complex patterns as effectively as other models.

After training and fitting the models to the data, the ones with the lowest RMSE for the validation set were selected, as they are expected to provide the most accurate predictions for each PG for future data. Based on the obtained results, there is no one best model to treat all of the PGs at once, and different ones for specific products is the most precise approach. After all, the model employed for as

10 Cf. Menculini, L., et al. (2021).

11 Cf. Alharbi, F. R., & Csala, D. (2022).

12 Cf. Dairu, X., & Shilong, Z. (2021).

many as half of the PGs was found to be linear regression. XGBoost performed best for four, Prophet for two and SARIMAX for only one PG.

- **PG 1:** The optimal model for PG 1 was determined to be linear regression. The model included predictors such as the average sales value from the three previous months (rolling mean), monthly growth rate of the sales and Swiss Machinery and Electricals Production Index, which were identified during the feature selection process.
- **PG 3:** For PG 3, linear regression also proved to be the most accurate model. This one, however, in addition to the past sales data also considered the variable the changes in Japanese Machinery and Electricals Production Indexes.
- **PG 4:** XGBoost outperformed linear regression for PG 4's sales forecasting due to its ability to capture complex nonlinear relationships. This model was enriched with features like rolling mean and growth rate related to its past sales and changes in prices of electrical equipment in the USA.
- **PG 5:** For PG 5, linear regression has been identified as the most accurate forecasting model based on RMSE score. Predictors included a rich set of explanatory variables related to past sales and various market dynamics.
- **PG 6:** PG 6 is another example for which XGBoost has performed best. The model incorporated rolling mean of sales, and features connected with German Electrical Equipment Production from external predictors.
- **PG 8:** The sales of the PG 8 were most accurately predicted with the Prophet model based on training and validation set, that is why it was decided to apply this one for predicting the future sales as well. The factors included in the model were the ones connected to its past sales and some regarding the world prices of natural gas price USA prices of electrical equipment.
- **PG 9:** The best performing linear regression model for PG 9 was forecasting based on features connected with German and Japanese Machinery and Electricals Production Index as well s Great Britain and World's Electrical Equip. Production Index.
- **PG 11:** The sales of PG 11 seemed to be most successfully obtained by the SARIMAX technique. This is the only product for which this model performed best. The factors on which the model was based were as follows: rolling mean of its sales and of the French Machinery and Electricals Shipment Index and change in Italian Electrical Equipment Index.
- **PG 12:** PG 12's sales were predicted based on many characteristics, some of which in link with the sales themselves, whereas others were informative about market indices of Electrical Equipment Production in Great Britain and USA currency. The best predictions were obtained with XGBoost model in this case.
- **PG 13:** PG 13 is another example for which Linear Regression has performed best. The model included only two predictors, such as: growth rate of sales of PG 13 and producer prices of electrical equipment in China.
- **PG 14:** Sales of the PG 14 appeared to be most accurately forecasted by the Prophet model based on RMSE score, and the features chosen for this process were related to its past sales and China Electrical Equip. Producer Price and Japan Machinery and Electricals Production Indexes.
- **PG 16:** Model trained for predicting sales for PG 16 incorporated variables concerning their rolling mean and growth rate as well as US currency. The outperforming forecasting technique for this article was linear regression.

- **PG 20:** Another example where linear regression worked best was for PG 20. The model contained predictors related to both sales and market indexes of British and World's Electrical Equipment Index.
- **PG 36:** PG 36 is the last of the four products for which the sales were chosen to be forecasted by XGBoost model. It incorporated features regarding the groups' 36 past sales, Europe Machinery and Electricals Production and Great Britain Electrical Equipment Price Indexes.

3.4 Results Evaluation and Interpretation

The table presented represents the performance of four forecasting models across fourteen distinct product categories, using two evaluation metrics: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The lowest RMSE values, highlighted in green for each PG, denotes the model that best forecasts the sales of that PG.

Table 2: Model Performance per Product Group

Product Group	Eval. Metric	Prophet	XGBoost	SARIMAX	Linear Reg.
PG 1	RMSE	1756845.82	2258949.49	3712102.1	1726766.81
	MAPE	3.42 %	4.69%	8.57 %	3.86 %
PG 3	RMSE	1312025.9	2001356.59	2548446.27	880732.96
	MAPE	7.15 %	19.95%	14.23 %	5.53 %
PG 4	RMSE	86107.16	67682.54	357156.75	80243.92
	MAPE	24.75 %	17.38%	120.7 %	20.73 %
PG 5	RMSE	3553411.89	4105563.53	4511660.37	2757638.52
	MAPE	28.59 %	19.33%	28.68 %	19.49 %
PG 6	RMSE	248636.14	193028.20	280786.59	195225.34
	MAPE	64.11 %	56.78%	61.63 %	52.86 %
PG 8	RMSE	397826.27	527526.53	545438.77	486231.62
	MAPE	26.5 %	36.73%	35.25 %	40.47 %
PG 9	RMSE	7051.52	6605.21	10107.33	4473.23
	MAPE	487.12 %	425.43%	440.59 %	284.51 %
PG 11	RMSE	792597.41	882826.47	684899.79	965508.54
	MAPE	92.85 %	70.32%	96.66 %	149.63 %
PG 12	RMSE	78016.31	50310.34	285457.27	97272.71
	MAPE	31.9 %	15.54%	112.57 %	41.94 %
PG 13	RMSE	14495.51	11622.46	19882.31	11503.46
	MAPE	61.16 %	67.95%	73.15 %	61.95 %
PG 14	RMSE	12586.24	13710.92	17656.99	13641.5
	MAPE	93545.5 %	34143.79%	280849.57%	134158.8 %
PG 16	RMSE	188619.54	48145.72	242779.74	43319.29
	MAPE	156.12 %	21.83%	177.94%	25.47 %
PG 20	RMSE	2006.94	1093.75	2114.92	1043.58
	MAPE	35095.57 %	6637.44%	29329.57%	20322.33 %
PG 36	RMSE	18274.62	7775.59	18083.38	10525.52
	MAPE	242.95 %	84.94%	521.03%	196.61 %

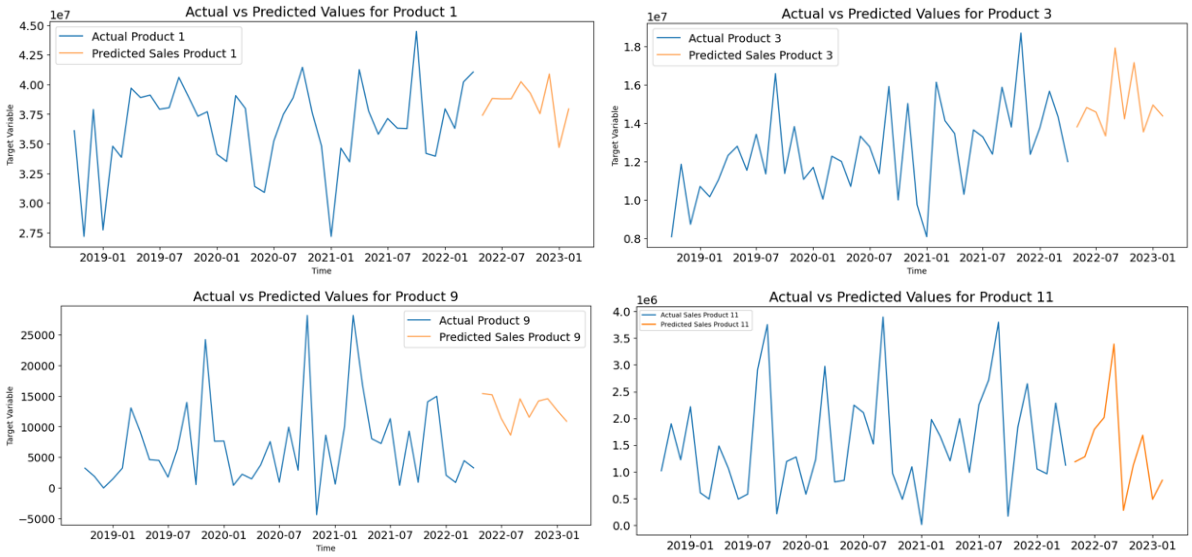
Source: Own elaboration based on Jupyter Notebook.

The MAPE values, presented as percentages, are relevant to compare PGs. This relative measure compensates for the difference in sales volumes across PGs, offering an accurate assessment. Four categories can be analyzed based on this metric:

- **High Accuracy** (<10% MAPE): for **PG 1** and **PG 3**, the Linear Regression model shows great forecasting precision which suggests a consistent and relatively predictable linear sales pattern within these groups.
- **Reasonable Accuracy** (10%-30% MAPE): for **PG 4**, **PG 5**, **PG 8**, **PG 12** and **PG 16**, which indicates reliable forecasts although there is potential for improvement.
- **Low Accuracy** (>30% MAPE): **PG 6**, **PG 9**, **PG 11**, **PG 13** and **PG 36** have a higher level of forecast errors. Despite the models being the best among the options, the values suggest challenges in accurately forecasting for these specific PGs, due to fluctuating demand patterns or other market dynamics.
- **Exceptional variability** (extraordinarily high MAPE): **PG 14** and **PG 20**, marked by extreme MAPE values, likely due to outliers or volatile product demand. This indicates a critical need for model reassessment or data reevaluation.

The forecast evaluation of the PGs has provided valuable insights into sales trends, seasonality, and overall performance, which informs how to manage inventory more effectively.

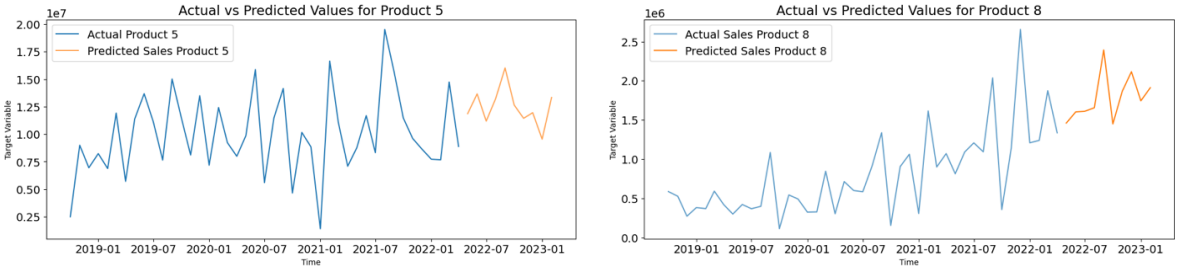
Figure 4: Sales Forecast for Product Groups 1, 3, 9 and 11



Source: Own elaboration based on Jupyter Notebook.

Firstly, **PG 1**, **PG 3**, **PG 9** and **PG 11** present significant seasonal trends. **PG 1** and **PG 3** sales show a decline in January, suggesting a post-holiday fall. On the other hand, **PG 9** experiences an increase in sales around October/ November and **PG 11** around August. For these groups, adjusting inventory levels to match these predictable patterns will be key to meeting customer demand and avoiding over or under stocking.

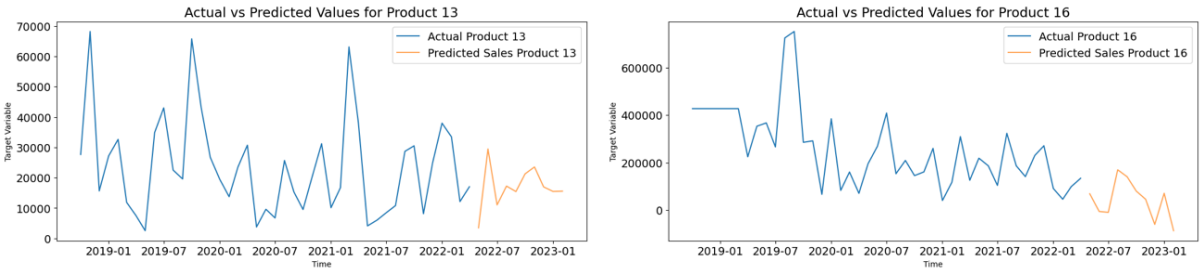
Figure 5: Sales Forecast for Product Groups 5 and 8



Source: Own elaboration based on Jupyter Notebook.

Secondly, the **PGs 3**, **5** and **8**, are experiencing growing demand over time. It is crucial to ensure sufficient stock for these groups to capitalize on this upward trend and not miss out on potential sales.

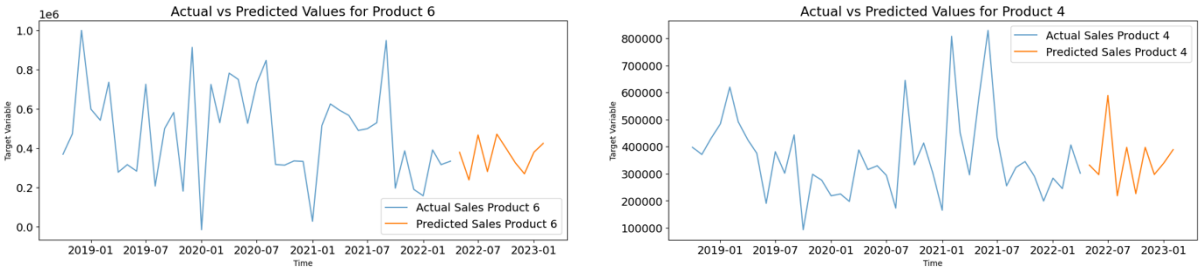
Figure 6: Sales Forecast for Product Group 13 and 16



Source: Own elaboration based on Jupyter Notebook.

Thirdly, there are some product groups showing downward trends, as in the case of **PG 13** and **PG 16**, which indicates that they may be entering a decline phase. Especially, for **PG 16**, there is an increased need to manage inventory carefully to avoid overstocking that could tie up resources, as the product group may be reaching the end of its lifecycle. **PG 6** is expected to have low sales thus being important to oversee inventory levels in order to avoid overstocking too.

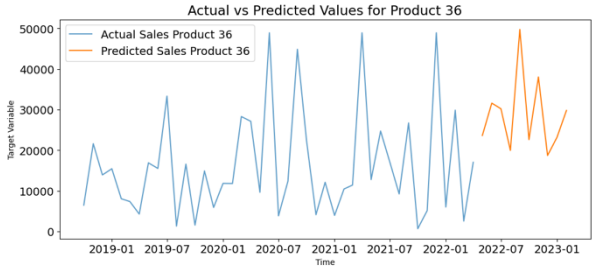
Figure 7: Sales Forecast for Product Group 6 and 4



Source: Own elaboration based on Jupyter Notebook.

In addition, for **PG 1**, **PG 4**, **PG 9** and **PG 11**, despite their sales volatility, they show consistency in the long run. This requires flexible inventory management that can accommodate short-term fluctuations, while still meeting the overall stable demand.

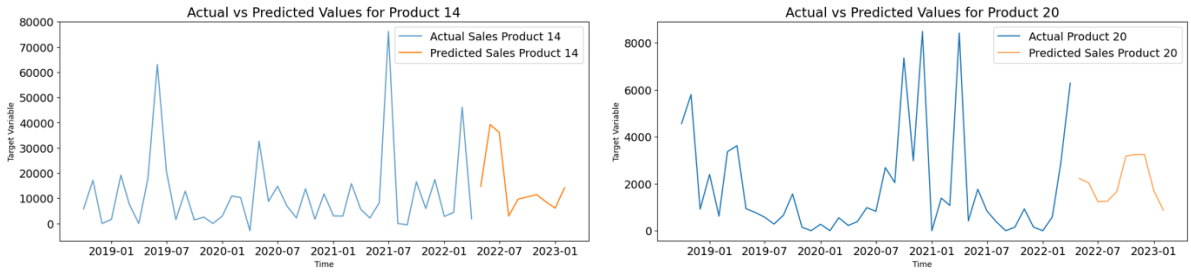
Figure 8: Sales Forecast for Product Group 36



Source: Own elaboration based on Jupyter Notebook.

Furthermore, there is a special product group, **PG 36**, that stands out for its very inconsistent sales patterns, however it has the potential for high sales. Inventory management for this group needs to be particularly agile to avoid overstock during low-demand periods and shortages during peak times.

Figure 9: Sales Forecast for Product Groups 14 and 20



Source: Own elaboration based on Jupyter Notebook.

Regarding **PG 14** and **PG 20**, as their forecasts are not reliable due to isolated sales peaks, identified as outliers, it would be impractical to base decisions on these evaluations currently.

The application of these insights can improve efficiency, reduce costs, and enhance customer satisfaction. Continuous improvement of forecasting and inventory strategies will be crucial for adapting to market changes and sustaining growth.

4 EVALUATION

During the evaluation of the sales forecasting model, it was found that the majority of the developed models performed satisfactorily during both training and testing phases. While some predictions may not have achieved the desired level of accuracy, they effectively captured and represented significant fluctuations and trends within the sales data. This ability to highlight critical episodes within the sales process provides valuable insights for decision-making and resource allocation within the company.

Moreover, even when the predictions were not as accurate as desired, they still offered a general idea of the most significant moments and changes in the business. This is a valuable insight, as it is crucial to have an understanding of these sensitive episodes within an industry.

Overall, the business objectives have been met. The models and forecasts generated proved to be insightful, offering valuable perspectives on the behavior of each PG. Despite potential limitations, such as occasional inaccuracies, the models provided meaningful guidance for understanding sales

dynamics and organizing internal processes effectively. This underscores the importance of leveraging data-driven approaches to inform strategic decision-making and enhance operational efficiency within the organization.

It is important to note that the final quantitative evaluation of the results will be conducted by Siemens, utilizing the RMSE score on the test set. Siemens holds the real values necessary for this evaluation, ensuring an accurate assessment of the model's performance against actual sales data. While the preliminary analysis provides valuable insights into the model's efficacy, Siemens' evaluation will offer a comprehensive validation of the forecasting accuracy and reliability. This collaborative effort underscores the importance of bridging data science methodologies with real-world business applications to drive informed decision-making and achieve tangible business outcomes.

5 DEPLOYMENT AND MAINTENANCE PLANS

To deploy the monthly sales forecasting model into production, a comprehensive strategy will be implemented to ensure seamless integration into existing business processes within Siemens' Smart Infrastructure Division in Germany. The deployment process will involve the following steps:

- **Integration planning**: A cross-functional team comprising data scientists, IT specialists, and business stakeholders will collaborate to devise a deployment plan specifically tailored for the monthly sales forecasting model. This plan will outline the necessary steps to integrate the model into existing sales and inventory management systems, as well as related business processes.
- **System compatibility assessment**: The model's compatibility with existing IT infrastructure, including databases and analytical tools, will be assessed to identify any potential integration challenges. Necessary changes or updates to systems and databases will be implemented to ensure smooth deployment and interoperability with the forecasting model.
- **Stakeholder training**: Training sessions will be conducted to familiarize relevant stakeholders, including sales teams and inventory managers, with the functionality of the forecasting model. Training will focus on interpreting the model's predictions and integrating them into decision-making processes to optimize inventory management and sales strategies.
- **Rollout and monitoring**: The forecasting model will be gradually rolled out across the organization, starting with pilot testing for selected PGs. Feedback from users will be collected and incorporated into the deployment process to address any usability issues or performance concerns.

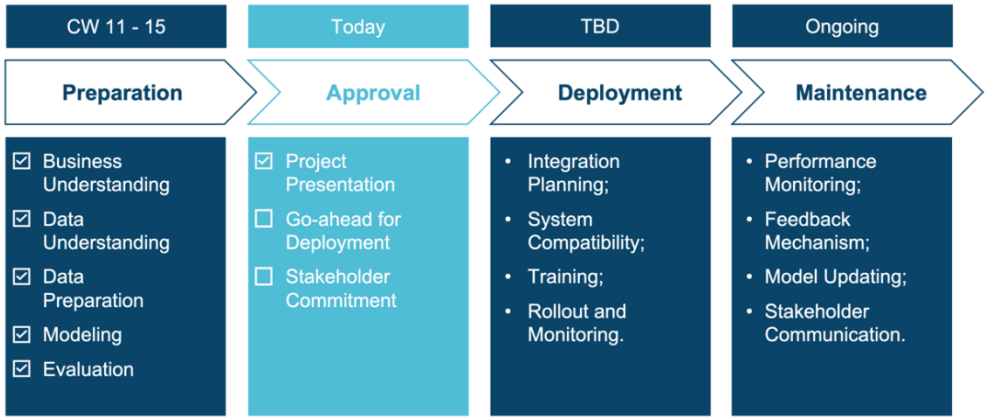
Following deployment, ongoing monitoring and maintenance of the monthly sales forecasting model will be essential to ensure its continued effectiveness and relevance. The following measures will be implemented for performance assessment and model maintenance:

- **Performance monitoring**: Regular monitoring of the model's performance will be conducted to assess its accuracy and reliability in forecasting monthly sales volumes. The key performance metrics RMSE and MAPE will be tracked to identify any deviations from expected performance levels.

- **Feedback mechanism**: A feedback mechanism will be established to solicit input from end-users regarding the model's performance and usability. This feedback will be used to identify areas for improvement and inform future updates or enhancements to the forecasting model.
- **Model updating**: Periodic updates to the forecasting model will be conducted to incorporate new sales data, market trends, and business insights. This may involve retraining the model with updated data or refining forecasting algorithms to improve predictive accuracy.
- **Documentation and version control**: Comprehensive documentation of the forecasting model, including its architecture, data sources, and assumptions, will be maintained to facilitate transparency and reproducibility. Version control mechanisms will be implemented to track changes and updates to the model over time.
- **Stakeholder communication**: Regular communication with stakeholders, including sales managers and business unit leaders, will be maintained to keep them informed of any changes to the forecasting model and its impact on business operations. This will ensure alignment between the model's objectives and organizational goals.

To ensure a systematic deployment and maintenance of the monthly sales forecasting model, a well-defined timeline is essential. Below is a brief overview of the proposed schedule.

Figure 9: Project Plan



Source: Own elaboration.

By implementing a structured deployment strategy and establishing thorough monitoring and maintenance processes, the monthly sales forecasting model will be effectively integrated into Siemens' Smart Infrastructure business processes, enabling informed decision-making and driving operational excellence in sales.

6 CONCLUSIONS

This project aimed to develop robust sales forecasting models for Siemens' Smart Infrastructure Division targeting specific PGs crucial to their market strategy in Germany. Throughout the project, multiple data-driven techniques were employed, including Prophet, XGBoost, SARIMAX, and Linear Regression. The primary goal was to enable Siemens to optimize inventory management, enhance operational efficiency, and align supply with demand more effectively.

The models varied in their effectiveness, with certain models excelling in specific contexts due to the unique characteristics of the data related to different PGs. Linear Regression models provided clear and interpretable results, making it a valuable tool for straightforward forecasting needs where relationships are predominantly linear. In contrast, XGBoost and Prophet models excelled in scenarios requiring the identification of complex patterns and seasonal fluctuations, respectively.

The evaluation phase highlighted that while some models did not meet the predefined accuracy thresholds, they effectively captured and delineated critical sales trends and fluctuations, which are invaluable for strategic decision-making. The iterative model testing and validation process has proven essential in refining the approaches and ensuring that the models remain relevant under changing market conditions.

6.1 Considerations for Model Improvement

To further enhance the forecasting models, the following recommendations are proposed:

- **Integration of additional models**: Incorporating more diverse models may improve predictive performance by capturing different aspects of sales dynamics not addressed by the current models. Techniques such as ensemble methods could be explored to combine the strengths of multiple forecasting approaches.
- **Expanding external factors:** Including more granular macroeconomic indicators and potentially integrating real-time data streams could refine the accuracy of the forecasts. Factors such as geopolitical events, industry-specific trends, and consumer sentiment indices should be considered.
- **Advanced validation techniques:** Implementing cross-validation methods across different time periods could help in understanding the models' performance stability and robustness. This would also mitigate the risk of overfitting by ensuring that models are generalizable across various operational contexts.
- **Regular model updates and maintenance:** Establishing a routine for continuous model evaluation and updating is critical. This process should include regular recalibrations to align the models with the latest available data and market insights, thus maintaining their relevance and accuracy over time.

7 REFERENCES

Alharbi, F. R., & Csala, D. (2022). A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors Forecasting Model-Based Time Series Approach. *Inventions*, 7(4), 94. https://doi.org/10.3390/inventions7040094.

Agrawal, R. (2021). Time-series Forecasting -Complete Tutorial | Part-1. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/07/time-series-forecasting-complete-tutorial-part-1/.

Agrawal, S. (2023). Feature Selection Using Lasso Regression. Retrieved from *Medium*: https://medium.com/@agrawalsam1997/feature-selection-using-lasso-regression-10f49c973f08.

Brownlee, J. (2020). Data Preparation.

Brownlee, J. (2020). Introduction to Time Series Forecasting with Python.

Dairu, X., & Shilong, Z. (2021). Machine Learning Model for Sales Forecasting by Using XGBoost. In 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 480-483). Guangzhou, China. https://doi.org/10.1109/ICCECE51280.2021.9342304.

Gong, L., & Shi, J. (2010). On comparing three artificial neural networks for wind speed forecasting. *Applied Energy*, 87(7), 2313-2320. https://doi.org/10.1016/j.apenergy.2009.12.013.

Guhanesvar. (2021). Feature Selection Based on Mutual Information Gain for Classification and Regression. Retrieved from *Medium*: https://guhanesvar.medium.com/feature-selection-based-on-mutual-information-gain-for-classification-and-regression-d0f86ea5262a.

Menculini, L., Marini, A., Proietti, M., Garinei, A., Bozza, A., Moretti, C., & Marconi, M. (2021). Comparing Prophet and Deep Learning to ARIMA in Forecasting Wholesale Food Prices. *Forecasting*, 3(3), 644-662. https://doi.org/10.3390/forecast3030040.

8 APPENDIX

The glossary below ensures clarity for the stakeholders and enhances understanding of key technical terms used throughout this report. These concepts are the core of the analytical approach used.

Table 3: Glossary

Term	Explanation
Root Mean Squared Error (RMSE)	RMSE is a commonly used metric to measure how accurate a prediction model is. It looks at the difference between what the model predicts and what actually happens. The RMSE tells us on average how far off the predictions are from the actual outcomes. The lower the RMSE, the better the predictions.
Mean Absolute Percentage Error (MAPE)	MAPE is another metric to measure how accurate a prediction model is. It calculates the average percentage difference between the predicted and actual values. Like RMSE, lower values of MAPE indicate better predictions.
Prophet	Prophet is a forecasting tool developed by Facebook. It is particularly good at predicting future trends in data that has repeating patterns, like sales data with seasonal ups and downs. Prophet breaks down the data into parts like overall trend, yearly cycles, and special events like holidays, to make more accurate predictions.
XGBoost	XGBoost is a powerful machine learning algorithm used for making predictions based on data. It combines the predictions of many individual models to come up with the best possible prediction.
SARIMAX	SARIMAX is a statistical method used for predicting future values in a series of data points. It considers seasonal patterns, how today's value relates to yesterday's, and any external factors that might influence the data, to make predictions.
Linear Regression	Linear regressions is a statistical method used to model the relationship between two or more variables. It assumes a linear relationship between the variables and aims to fit a straight line that best describes the observed data. Linear regression is widely used for prediction and inference tasks in various fields.

Source: Own elaboration.