MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 2: Siemens Monthly Sales Forecast

André Moreira, number: 20222132

Beatriz Teixeira, number: 20211313

Catarina Nunes, number: 20230083

Pedro Catarro, number: 20230463

Group V

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

April, 2024

INDEX

1. EXECUTIVE SUMMARY

We were asked to solve a monthly sales forecasting challenge for Siemens through a project proposal, starting in April 2022.

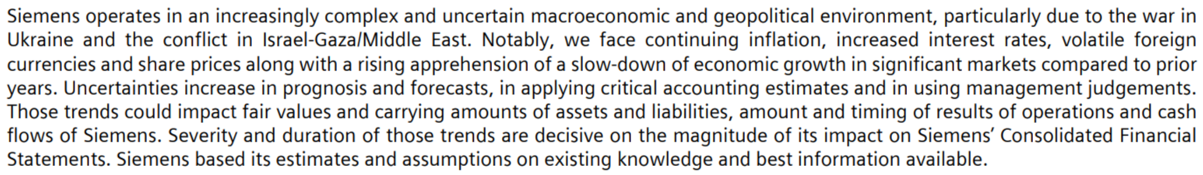
Founded in 1847, Siemens is a sustainable technologic company that act in sectors like industry, energy, healthcare, mobility, and infrastructure. This project focuses on Siemens' Smart Infrastructure Division in Germany, leveraging digital technology to optimize efficiency and sustainability.

Our objectives include automating sales forecasting, minimizing biases, centralizing data, and enhancing forecasting accuracy. Key questions will be addressed regarding sales trends, macroeconomic impacts, seasonal patterns, and leveraging historical sales data.

Success involves delivering a robust sales forecasting model, minimizing costs of imprecise predictions.

We were provided with two datasets from Siemens: 'Sales data.csv', containing sales data from October 2018 to April 2022 for 14 products, and 'Market data.xlsx', featuring crucial macro-economic indices for Siemens' key countries. Additionally, we've conducted supplementary research to identify additional macroeconomic and microeconomic factors impacting our analysis.

Our data mining goals involve creating two forecasts one to predict monthly sales over the specified 10-month period provided by Siemens, and second, to generate a final forecast for monthly sales throughout 2023.



Siemens operates in an increasingly complex and uncertain macroeconomic and geopolitical environment, particularly due to the war in Ukraine and the conflict in Israel-Gaza/Middle East. Notably, we face continuing inflation, increased interest rates, volatile foreign currencies and share prices along with a rising apprehension of a slow-down of economic growth in significant markets compared to prior years. Uncertainties increase in prognosis and forecasts, in applying critical accounting estimates and in using management judgements. Those trends could impact fair values and carrying amounts of assets and liabilities, amount and timing of results of operations and cash flows of Siemens. Severity and duration of those trends are decisive on the magnitude of its impact on Siemens' Consolidated Financial Statements. Siemens based its estimates and assumptions on existing knowledge and best information available.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

Through a project proposal by Siemens, our group was assigned to deliver an answer regarding a monthly sales forecasting problem. The starting point of this project is to situate, in time, the problem, being it during the month of April, 2022 and to adopt the CRISP-DM methodology as it is the most used methodology for data science and analytics projects.

Siemens, a global technology company, faces challenges in managing resources and inventory due to the dynamic nature of the business environment. An effective management strategy is crucial for predicting performance, finances, sales, inventory, and resources.

Monthly sales forecasting is a critical tool for evaluating overall company performance and developing strategic plans. Its analysis aids in improving financial planning, inventory management, performance evaluation, and identifying growth opportunities but it can be costly in resources to maintain if the company doesn't have an efficient approach.

BUSINESS UNDERSTANDING

Founded in 1847 and known for its innovative and sustainable technologies and solutions, Siemens, a German multinational corporation that operates in many sectors, like industry, energy, healthcare, mobility and infrastructure. With a strong focus on research and development, the company invests heavily in emerging technologies such as AI, IoT and digitalization to drive efficiency, productivity, and environmental sustainability across its many business sectors.

This project is focused on the Business Unit of Siemens' Smart Infrastructure Division in Germany. Smart infrastructure refers to the integration of digital technology into physical infrastructure systems, such as buildings, transportation networks, energy grids, and water management systems, to optimize their efficiency, sustainability, and performance. Smart infrastructure solutions typically involve the use of sensors, data analytics, automation, and connectivity to gather information, monitor operations, and make real-time adjustments to improve functionality and reduce resource consumption. Siemens is a major player in the smart infrastructure sector, providing a wide range of solutions and services to help cities and industries become more connected, efficient, and sustainable.

To gather more knowledge about the company and the historical factors that affect the present state, we delve into the Siemens Annual Reports and gather information across the years.

2.1.1. Fiscal Year 2019

Revenue Growth: Siemens targets annual revenue growth of 4% to 5%, measured on a comparable basis. This excludes currency translation and portfolio effects, which are calculated as absolute differences divided by revenue for the comparison period. The same calculations apply to orders.

Overall Economic Conditions: Global GDP growth slowed to 2.6% in fiscal 2019 due to escalating trade tensions, geopolitical uncertainties, and declining industrial production. This affected trade-reliant regions like Europe, the Americas, and Asia. Advanced countries saw GDP growth drop to 1.6%, while emerging countries' growth fell from 4.6% to 4.1%.

Smart Infrastructure: Formed in fiscal 2019, it combines several divisions to offer solutions for energy supply and building technology. R&D focuses on decentralization, decarbonization, and digitalization, with investments in digital offerings and renewable energy technologies.

Results of Operations – Orders and Revenue by Region: Emerging markets, particularly China, drove growth, while Europe, C.I.S., Africa, Middle East had mixed order development. The Americas saw significant order growth, driven by currency effects, with double-digit growth in SGRE and Mobility. Asia, Australia region experienced significant order growth, led by SGRE in Taiwan. In terms of revenue, growth was moderate in Europe, C.I.S., Africa, Middle East, led by SGRE, with declines in Gas and Power. Americas revenue increased significantly due to currency effects, with declines in SGRE. Asia, Australia revenue rose moderately, led by Siemens Healthineers and Digital Industries, but SGRE revenue decreased.

Subsidiary by region:

- Germany – 136 companies
- Europe, C.I.S., Africa, Middle East (without Germany) – 525 companies
- Americas – 171 companies
- Asia, Australia – 234 companies

2.1.2. Fiscal Year 2020

Revenue Growth: In Siemens financial strategy, targets annual revenue growth between 4% to 5% on a comparable basis. Comparable growth excludes currency translation and portfolio effects, which are calculated as absolute differences divided by revenue for the comparison period. This approach helps us manage and control revenue growth effectively.

Overall Economic Conditions: In fiscal 2020, the global economy faced a severe recession due to COVID-19, with an expected 4.5% contraction in global GDP. Siemens' business performance was significantly impacted, with varying effects across industries, customer markets, and regions. Some sectors accelerated to meet increased online activity, while others suffered. Reduced sales and service activities led to cost savings, but stable production was maintained. China experienced an early return to growth, while Europe and the Americas continued to struggle with COVID-19 impacts, affecting Siemens' volume and income.

Smart Infrastructure: Despite COVID-19 impacts, revenue remained stable. Market trends indicate opportunities in building tech and electrification. Geographically, Asia shows recovery, while Europe faces challenges. Overall, Smart Infrastructure maintains a resilient business mix amid changing market dynamics.

Results of Operations – Orders and Revenue by Region: COVID-19 impact – Orders declined notably, especially in Mobility. Europe saw a significant decrease, with Germany showing growth in Mobility but declines in other sectors. Americas also experienced a drop, mainly due to Mobility. Asia, Australia saw slight overall growth, driven by Digital Industries, particularly in China. Revenues slight decline overall, driven by Digital Industries and Smart Infrastructure. Mobility saw slight growth. Europe, C.I.S., Africa, Middle East with a moderate decrease, with slight decline in Germany. Americas were stable, with Mobility growth offsetting declines elsewhere. Asia, Australia with a moderate decline, impacted by India, but China saw moderate growth.

Subsidiary by region:

- Germany – 112 companies
- Europe, C.I.S., Africa, Middle East (without Germany) – 284 companies
- Americas – 105 companies
- Asia, Australia – 130 companies

2.1.3. Fiscal Year 2021

Revenue Growth: Siemens aimed for 4% to 5% annual revenue growth up to fiscal 2021, measured by comparable growth. This accounts for currency and portfolio effects. Now, under a modified framework, the target is 5% to 7% comparable revenue growth over a three to five-year cycle.

Overall Economic Conditions: Global GDP is set to rebound by 5.5% in 2021 after a 3.4% contraction in 2020 due to COVID-19. Stimulus programs and vaccine developments supported economic recovery, despite setbacks from new virus variants and supply disruptions. China's economy thrived, while other major economies also saw strong rebounds: EU (5.0%), U.S. (5.4%), Japan (2.3%), India (7.7%). Advanced countries are expected to grow by 4.9%, while emerging markets by 6.4%.

Smart Infrastructure: Fiscal 2021 saw growth in orders and revenue across regions, driven by industrial demand recovery and strong performances in Asia. Adjusted EBITA and profitability rose, supported by cost savings and recovery efforts. Market growth is expected to continue, led by pharmaceuticals, data centers, and utilities, with Asia driving the fastest expansion.

Results of Operations – Orders and Revenue by Region: Orders surged across sectors globally, driven by Mobility's record deal in the U.S., while Siemens Healthineers and Digital Industries saw significant growth, particularly in China. Revenue up significantly YoY, led by Siemens Healthineers and Digital Industries. Strong growth in Europe, C.I.S., Africa, Middle East. Americas and Asia, Australia also show increases.

Subsidiary by region:

- Germany – 121 companies
- Europe, C.I.S., Africa, Middle East (without Germany) – 306 companies
- Americas – 124 companies
- Asia, Australia – 154 companies

2.1.4. Start of Fiscal Year 2022

Siemens saw its profit fall by 49% to 1.21 billion euros in the second quarter of the fiscal year, partly due to the costs of exiting the Russian market. Revenues globally increased by 16% to 17.04 billion euros, with orders rising by 32% to 20.98 billion euros. The company need to source resources from other countries.

BUSINESS OBJECTIVES

The project aims to address the following motivations: Efficiency automating sales forecasting to save time and reduce manual effort; Bias Reduction implementing data-driven algorithms to minimize human biases in forecasting; Data Centralization utilizing AI to gather and analyze data from multiple sources, enhancing accuracy compared to manual methods; Cost Savings improving forecasting accuracy to allocate resources more effectively and reduce potential losses from poor forecasting.

To answer the challenges we have to define the following questions: What are the primary factors influencing sales trends in the targeted product groups? How do macroeconomic indices impact sales performance within the German market? Are there any seasonal patterns or trends affecting sales fluctuations? What are the key drivers behind variations in sales volumes over the specified period? How can historical sales data be leveraged to predict future sales trends accurately?

The previous business questions will be worked under the following assumptions: The input/tool should be interchangeable and usable in other businesses leading to reusability and scale; The solution should integrate into the existing business tools and system landscape integrating the business process; The solution must generate a net positive value for the organization; The cost of the use case must surpass the expected value add of it; The model must be feasible; Have a realistic effort required for full development and maintenance of the solution; and must be effective for business professionals. To meet the previous points, we have the next requirements: Targeting selected product groups within one Business Unit of the Smart Infrastructure Division; Focusing on the largest country of the business unit, Germany; Utilizing sales data from October 2018 to April 2022; Incorporating important macroeconomic indices.

The benefits of this project are expected to be the achievement of a robust sales forecast model and to address the challenge of generating a monthly sales forecast based on real data from a Siemens business unit in Germany.

BUSINESS SUCCESS CRITERIA

The business success criteria for this project are to deliver a solution that will turn the monthly sales forecasting into a robust model, avoiding the cost of opportunity that comes from unprecise predictions. The data mining success criteria is to create a model that will eventually predict with more accuracy Siemens sales in months to come, based on their previous sales and macro and microeconomic factors, and understand the correlations between them. Several approaches will be tested to find which combo of features and models better suits the analysis of each product and finally reach an overall model with the lowest RMSE possible. This will be further discussed in the modeling phase.

SITUATION ASSESSMENT

In terms of available resources crucial for our analysis, our team has received two datasets from Siemens: The first one regarding the **Training Set 'Sales data.csv'** comprising the sales data from

October 2018 to April 2022 for 14 products and the second, **'Market data.xlsx',** that has important macro-economic indices for Siemens in its most important countries.

In addition to these datasets, our team has undertaken supplementary research to identify further macroeconomic and microeconomic indices that could significantly impact our analysis. The macroeconomic indicators are the monthly interest rates for China, Eurozone, France, Germany, India, Russia, United Kingdom, and United States, the consumer price index (CPI) for China, Eurozone, France, Germany, India, OCDE (EU), Russia, Saudi Arabia, United Kingdom, and United States, exchange rates: CNY/EUR, GBP/EUR, HKD/EUR, INR/EUR, RUB/EUR, SEK/EUR, and USD/EUR, nominal effective exchange rates for China, Germany, European Union, India, Russia, Saudi Arabia, Sweden, United Kingdom, and United States, gross domestic product (GDP) for China, European Union, France, Germany, India, Russia, Saudi Arabia, the United Kingdom, and the United States, unemployment rate for France, Germany, the United Kingdom, and the United States, consumer confidence index for China, France, Germany, India, OCDE (EU), Russia, United Kingdom, and United States, business confidence indicator for China, France, Germany, India, OCDE (EU), Russia, United Kingdom, and the United States, business tendency (manufacturing) for the European Union, Germany, United Kingdom, United States, China, and India, business tendency (construction) for the European Union, and geopolitical events and conflict data (Ukraine and Russia) such as the conflict in Ukraine that can have significant impacts on global markets, including commodity prices and investor sentiment. These events can indirectly affect Siemens' operations, especially if there are disruptions in supply chains or changes in global economic conditions that affect demand for its products and services. It's noteworthy to mention that while we have selected pertinent countries for our analysis, the datasets comprise additional countries available for deeper exploration and analysis. As a microeconomic indicator we selected Siemens Stock Performance that monitors the stock performance of specific companies, such as Siemens, offering insights into industry-specific trends, market sentiment towards a particular company, and broader economic conditions.

DETERMINE DATA MINING GOALS

The data mining can be distributed in 2 goals being the first one the creation of a model that predict the monthly sales through a forecast on the 10 months required from Siemens and the second goal is to create a final forecast of the monthly sales for the year of 2023.

3. METHODOLOGY

Describe only the major steps involved in the process. Do not replicate what is already described in the source. If necessary, refer the reader to the source code.

DATA UNDERSTANDING

3.1.1. Data Loading

We've simplified the format of the sales file using Power Query before loading it to the notebook to simplify the formatting process. Since our project focuses on monthly sales forecasts, we've aggregated sales by month-year. The rest of the datasets are treated in the notebook.

3.1.2. Univariate EDA – Missing Values

The raw data presented some null values that we will have to look into in the future. It is an alert but will probably be solved when ranging the data sets on the time period we are analyzing.

3.1.3. Univariate EDA – Data Visualization

The following text provides insightful analyses into the sales patterns, anomalies, and trends for each product. These analyses are crucial for guiding decision-making and further investigation during the data preparation phase. Unusual sales patterns suggest potential outliers, like how the P1 relation with total Sales_EUR shares a similar distribution, indicating P1's significant impact on the overall dataset.

Anomalies are noticeable, particularly in November 2018, where P3 and P5 surpass P1 or in January 2021 for total sales where P1 dominates total sales, making it challenging to access seasonality for other products. Each product displays unique sales trends. For instance, P1 exhibits a notable negative anomaly in Nov 2018, while P3 and P5 show negative anomalies in Jan 2021. P6 demonstrates strong overall performance but significant negative anomalies in specific months. P8 indicates a growth trend, with October as a comparatively weaker sales month inducing seasonality. P11 presents seasonal peaks in September, and P12 suggests a possible recent launch because the sales started growing after August 2019. P14 displays positive anomalies in certain months, while P16 initially performs well but declines after April 2019. P20 shows inconsistency with minimal representation, and P36 features positive anomalies in specific months. Sales_EUR shows anomalies, with September standing out as the most positive month.

Analyzing the month-to-month variation provides deeper insights into the significant fluctuations observed earlier. Sometimes, these changes hide certain patterns, like when product sales went up by 600% in December 2018 for P1. It doesn't necessarily mean there were a lot more sales that month; it might just be the result of the previous month's lower sales. The same thing happened with other products, like P3 and P5, in February 2021.

Additionally, it's important to note that some variations contain gaps due to sales registering as zero. To fix this in our charts, we replaced those zeros with the number 1 to ensure continuity. When we calculated how much sales changed, we used the absolute sales numbers. This way, we could handle negative sales correctly; otherwise, our calculations might wrongly show that sales went down when, in reality, they grew..

3.1.4. Bivariate EDA: Data Visualization

Based on the data, it can be observed that products P1, P3, and P5 make up a significant portion, ranging between 89% and 95%, of the total sales across the years. Notably, P1 alone contributes to over half of this total. Sales_EUR is a reliable indicator of the overall sales in the dataset, as it shows a strong positive correlation of 0.72 with P1. This suggests that as sales of P1 increase the total sales tend to increase as well. Similarly, P3 and P5 also exhibit good correlations with Sales_EUR, which aligns with their significant contributions to total sales.

In addition, there are moderate positive correlations between certain pairs of products, including P8 and P12, P8 and P11, P3 and P5, P3 and P12, and P3 and P8. These correlations imply that these products may be moderately affected by similar market changes or external factors, such as shifts in customer preferences or market dynamics, which are all present in macroeconomic analysis.

DATA PREPARATION

During our data preparation for analysis, we removed columns that had missing values from some essential datasets, such as "market_final," "cpi_df," "exchange_rates," "cci_df," and "bci_df." This was done to ensure that our data is accurate and reliable. However, we did not include "GDP_df" in our analysis because we could not find any data for the year 2022.

For the "conflict_data" dataset, we had missing values, but we filled them with zeros since a lack of conflict incidents is not the same as incomplete information in this context.

When analyzing sales data, we set negative values to zero since sales cannot be negative. To identify outliers, we established lower and upper bounds, which helped us remove anomalies.

While merging datasets, we kept all variables and will determine the optimal variables for each product in the next step (modeling) using an algorithm. For instance, variables like 'business confidence' and 'business tendency for manufacturing' may have similar distributions, which could lead to redundancy. However, we will resolve this through the selection process.

MODELING

Regarding the modeling part, we will go product by product analyzing the combination of feature selection and the model that outcomes the lowest RMSE while visualizing which one of the predictions is less prone to overfit. This will lead us to the best combination of features per model per product.

3.1.5. Product 1

Analyzing one of the most prominent products we found that the model was high in overfitting. Overfitting occurs when a model learns noise present in the training data rather than the underlying pattern, potentially leading to poor generalization on unseen data. This could pose significant challenges, as it may result in inflated performance metrics during training but poor performance when applied to new data. To address this, simplifying the approach and eliminating redundant data show up is essential. Taking this into account we find Test 2 the best one for P1 although it has a more moderate nature compared to the target variable, it has a good performance (keeps up with the movement) during the forecast phase. Among the models evaluated, the AdaBoostRegressor() emerged as the top-performing predictive multivariate model. AdaBoostRegressor is a boosting ensemble method that combines multiple weak learners, typically decision trees, to create a strong learner. It sequentially trains each weak learner, adjusting the weights of training instances based on the performance of the preceding learners. This process allows AdaBoostRegressor to focus on the observations that are difficult to classify, ultimately improving the model's predictive accuracy.The feature selection of the model included the following variables: “ITA_M&E_Ship lag10”; “GBR_M&E_Prod lag1”; “GBR_Prod_Index_Elect lag1”.

3.1.6. Product 3

Regarding product 3, Tests 4 and 5 are the best fitting models and similar. The combo variable/model is the same, but we will use Test 4 because it is more moderate and trends with the shape/mode swing during the forecast test. Additionally, we chose this one because it does not overfit on the training data. Among the evaluated models, the RANSACRegressor() emerged as the top-performing predictive multivariate model. RANSACRegressor is a robust regression algorithm that is particularly useful when the dataset contains outliers or noise. Given that P3 exhibits some problematic spikes, even though we treated the outliers during preparation, RANSACRegressor proves to be a suitable choice. The feature selection of the model included "WLD_Prod_Index_Elect”, “CHN_M&E_Prod lag11” and “GBR_Prod_Index_Elect lag8”.

3.1.7. Product 4

For product 4 the Tests 2 and 3 are similar to each other and performed the best for P4. We are going to use test 2. The combination of variables and the model remains consistent across both tests. These tests followed the curve's general shape, although they are not perfect, they do not show signs of overfitting. Among the evaluated models, only the SVR() emerged as the best model. SVR, which stands for Support Vector Regression, is a technique used for regression tasks. It works by fitting a hyperplane in a high-dimensional space to minimize the error between the actual and predicted values. The feature selection of the model included “ITA_Prod_Index_Elect”, “CNY/EUR lag6”, “HKD/EUR lag1”

3.1.8. Product 5

Regarding product 5, Tests 1 and 2 are essentially the same trend and performed the best for P5. The combination of variables and the model remains consistent across both tests and so we are choosing Test 1. Among the evaluated models, only the `SVR()` emerged as the best model.The feature selection of the model included “CHN_M&E_Ship lag3” and “WLD_Prod_Index_Elect lag11”.

3.1.9. Product 6

- In this analysis we will have to be very careful with overfitting. To address this, simplifying the approach and eliminating redundant data show up as essential. All the tests carried out are prone to overfit being the `Test 3` the one that gives more guarantees of achieving more moderate results performing better in forecast. The `BaggingRegressor()` is the most robust predictive model due to its ability to reduce overfitting and enhanced generalization. BaggingRegressor is an ensemble learning method that aggregates predictions from multiple estimators to improve stability and accuracy. It uses bootstrap sampling to train each base estimator on different subsets of the training data.The feature selection of the model included ”FRA_M&E_Prod lag5”, “CHN_IR lag6” and “CHN_IR lag7”.

3.1.10. Product 8

In this case we are confronted with a target variable that shows some big variations which will difficult in the choice of models. On one hand we have `Test 1` presenting a conservative solution that follows the variations and inclination trend while being moderate and on the other hand we have `Test 2` presenting a more aggressive solution. For that matter we will choose `Test 1`, the best performance during forecast. The `MLPRegressor()` emerged as the top-performing predictive multivariate model. MLPRegressor is a neural network for regression tasks. It utilizes multiple layers of nodes to learn patterns in data. It optimizes weights via backpropagation, refining predictions through hidden layers and activation functions and it's also effective to capture nonlinear relationships in large datasets. The feature selection of the model included “CHN_M&E_Prod lag9”, “CHN_M&E_Ship lag9”, “WLD_P_Minerals lag2” and “OCDE(EU)_CPI lag4”.

3.1.11. Product 9

For product 9, we must test regarding overfitting as seen before. By doing so, we reach Test 2 as the best one for P9, although it has a more moderate nature compared to the target variable, it has a good performance (keeps up with the movement) during the forecast phase. Among the models evaluated, the `AdaBoostRegressor()` emerged as the top-performing predictive multivariate model.The feature selection of the model included "ITA_M&E_Ship lag10", "GBR_M&E_Prod lag1" and "GBR_Prod_Index_Elect lag1".

3.1.12. Product 11

In this analysis we will have to be very careful with overfitting. All the tests carried out are prone to overfit being the `Test 5` the one that gives more guarantees of achieving more moderate results performing better in forecast. The `BaggingRegressor()` is the most robust predictive model due to its ability to reduce overfitting and enhanced generalization. The feature selection of the model included "JPN_M&E_Prod lag6" and "JPN_Prod_Index_Elect lag6"

3.1.13. Product 12

In this analysis we have again the overfitting problem. All the tests made are prone to overfit being the `Test 3` the one that gives more guarantees of achieving more moderate results performing better in forecast. The GradientBoostingRegressor() is based on the idea of an ensemble method derived from a decision tree. The feature selection of the model included "CHN_CPI", "CHN_CPI lag1" and "OCDE(EU)_CPI lag10"

3.1.14. Product 13

Regarding product 13, Test 2 is the test that performed the best for P13. Among the evaluated models, only the ` LinearSVR()` emerged as the best model. The feature selection of the model included "WLD_P_Natural_Gas lag12", "CHN_IR lag1", "FRA_IR lag9" and "FRA_Unemployment_Rate lag7".

3.1.15. Product 14

Regarding product 14, test 3 was the one with the best performance. Among the evaluated models, only the `XGBRegressor()` emerged as the best model. XGBRegressor is used for regression problems where the intent is to predict continuous numerical values emerged as the best model. The feature selection of the model included "ITA_M&E_Prod lag12", "ITA_Prod_Index_M&E lag12" and "ITA_Prod_Index_Elect lag12"

3.1.16. Product 16

Regarding product 16, test 1 was the one with the best performance. Among the evaluated models, only the ` LinearSVR()` emerged as the best model. The feature selection of the model included "EURO(19)_IR", "EURO(19)_IR lag1", "EURO(19)_IR lag2", "IND_CPI lag4", "IND_CPI lag5"

3.1.17. Product 20

Regarding product 20, test 2 was the one with the best performance. Among the evaluated models, the RANSACRegressor() emerged as the top-performing predictive multivariate model. The feature selection of the model included “USA_CPI lag2” and “FRA_CConfidence lag10”

3.1.18. Product 36

Regarding product 36, test 5 was the one with the best performance while avoiding overfitting issues. Among the evaluated models, the RANSACRegressor() emerged as the top-performing predictive multivariate model. The feature selection of the model included "CHN_M&E_Prod lag6”, “CHN_M&E_Ship lag6”, “GBR_Prod_Index_Elect lag9”, “SIESTOCK_%VAR lag12”.

EVALUATION

Evaluation of predictive models reveals their effectiveness in meeting business objectives. Among various features, the model exhibits strong predictive capabilities while addressing concerns such as overfitting and outlier detection.

For example, models such as AdaBoostRegressor, RANSACRegressor, and SVR exhibit robust performance, provide more accurate forecasts and reduce the risk associated with volatile market conditions. and improve efficiency

4. RESULTS EVALUATION

The model creation was a success regarding the criteria established before, we have a robust model able to forecast Siemens monthly sales with precise information. As we used the lowest RMSE in each combination, now we only wait for Siemens global RMSE test and approval.

5. PLANNING, DEPLOYMENT AND MAINTENANCE

This project deployment and maintenance will be done through a 5 step schema characterized by: Planning and assessment, in a prior phase, during classes with the professor and Siemens invited professionals tutoring where we could analyze which were the challenges and requirements needed to approach; Development, being delivered at this point, with a raw solution to the problem having done a thorough analysis of the data provided, assigned the best performing model to each of the projects and arrived to a final solution/overall model with the professor oversight; Testing and quality assurance will be held on the next 6 months through a 3 phase plan: Organizing a meeting, two weeks from now, with the involved departments to gather brainstorming ideas, training employees and clarification of doubts; During the next three months, conduct a new forecast on data collected in this time period while doing a market research on the investment needed to expand the smart infrastructure network on the most relevant markets; At the end of the year, execute a final forecast test with the involved parts for the next year, in this case 2023, and present it. Software deployment on the end of the first trimester to the end users; Monitoring and maintenance will be provided on a yearly basis where we will train employees to get to know the software, take the most useful insights on each product during the course of the year and present some cases that can have impact on the models such as war events or natural disasters while refining the model to ensure that it takes into account the last cases, more precisely the Russia-Ukraine war.

6. CONCLUSIONS

IMPLICATIONS FOR BUSINESS

With the analysis made before in combination with the deployment plan there will be answers to the business questions before established. This will lead to an increase in revenue to Siemens because now the company can create strategies more suited to each product.

This project will also implicate bigger costs in appropriated staff to deliver the solution while adding value.

CONSIDERATIONS FOR MODEL IMPROVEMENT

Add new data to the model such as smart infrastructure on the most important countries particularly in the United States and data regarding Ukraine's or any other country with war implications supply chain. As seasonality is a point of friction throughout all products, there is the need to improve the model in this theme.

LIMITATIONS

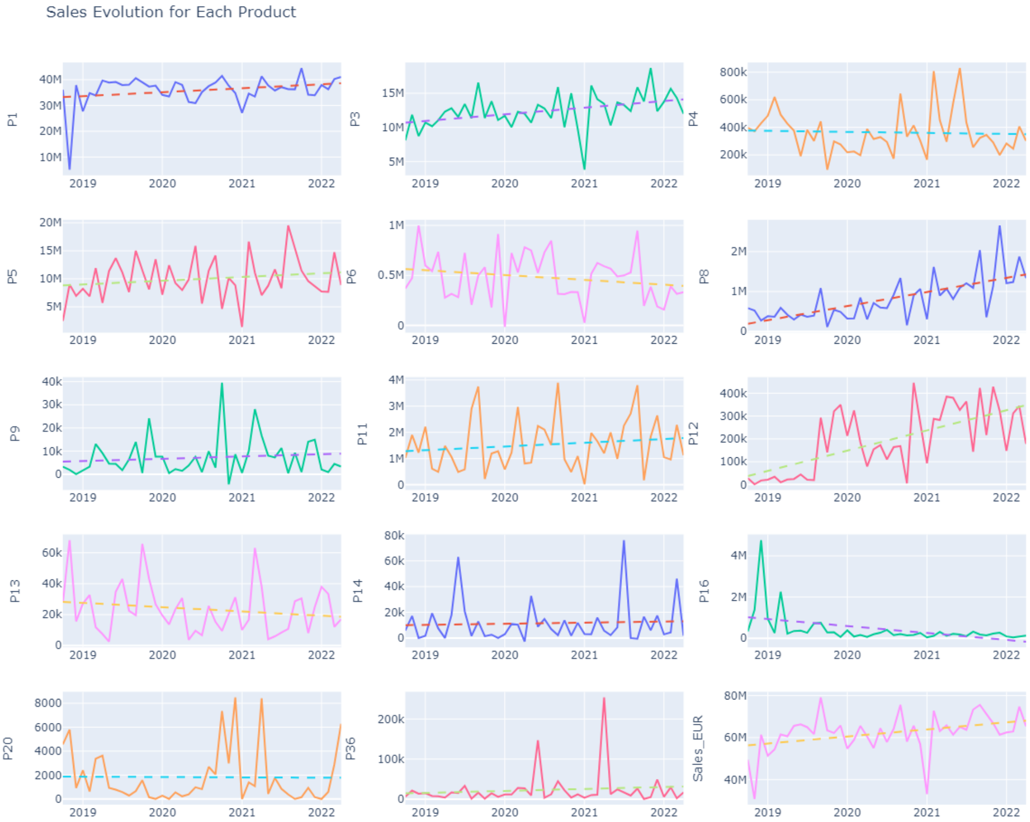
As limitations we identified several topics that deserve further discutions such as low historical data input that would rise the robustness of the model, a better product contextualization in order to identify the most important features of each product and how they correlate with the rest of the data, add a market sales view instead of the product sales view improving the way the model identifies the best markets the company is at regardless of the product and a suppliers analysis to add features quantifying the risk of each supplier country.

7. REFERENCES (IF APPLICABLE)

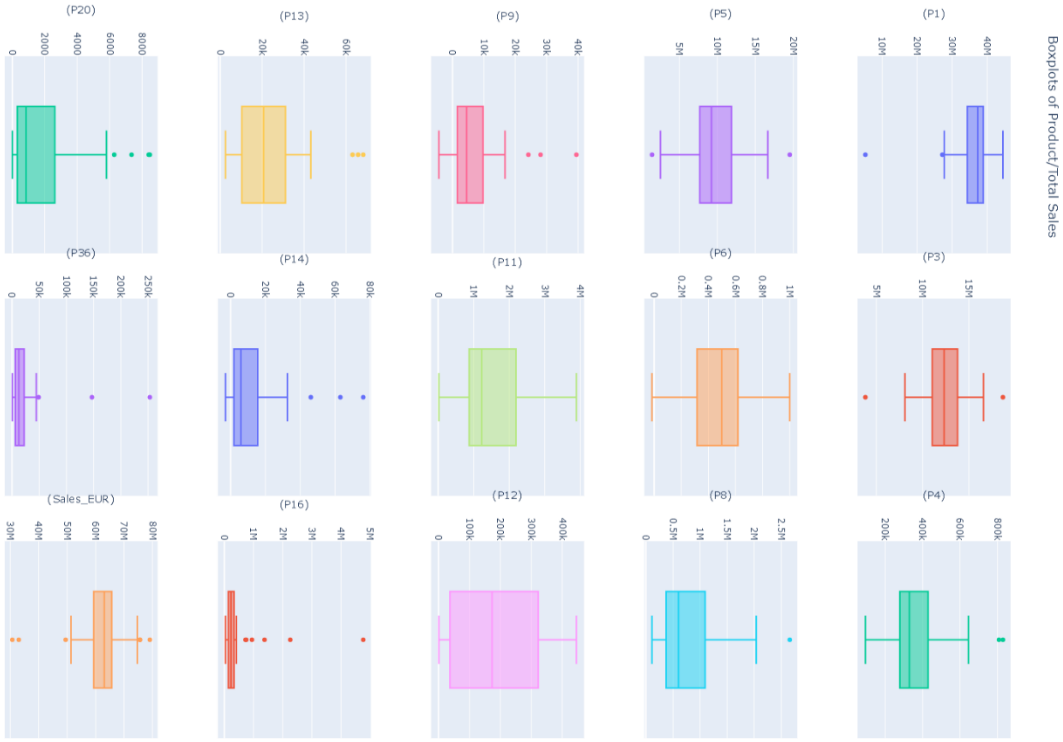
Team, C. O. (2023, June 28). *The software deployment Process: steps, importance, and best practices*. Codefresh. https://codefresh.io/learn/software-deployment/the-software-deployment-process-steps-importance-and-best-practices/

Automate the deployment of an Amazon Forecast time-series forecasting model | Amazon Web Services. (2023, May 4). Amazon Web Services. https://aws.amazon.com/pt/blogs/machine-learning/automate-the-deployment-of-an-amazon-forecast-time-series-forecasting-model/

8. APPENDIX



Sales Evolution for Each Product
P1
P3
P4
P5
P6
P8
P9
P11
P12
P13
P14
P16
P20
P36
Sales_EUR
2019
2020
2021
2022



Boxplots of Product/Total Sales
(P20)
(P13)
(P9)
(P5)
(P1)
(P36)
(P14)
(P11)
(P6)
(P3)
(Sales_EUR)
(P16)
(P12)
(P8)
(P4)