NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 2: Monthly Sales Forecast | Siemens

Bernardo, Pinto Leite, number: 20230978

Emília, Santos, number: 20230446

Nicolás, Zerené, number: 20230779

Ricardo, Kayseller, number: 20230450

Stephan, Kuznetsov, number: 20231002

Group F

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

April, 2024

INDEX

1. EXECUTIVE SUMMARY

Siemens Advanta Consulting has launched an ambitious initiative to innovate sales forecasting methods within its organization. The project seeks to offer monthly sales forecasts by analyzing data spanning from October 2018 to April 2022. The objective is to improve the accuracy of predictions by optimizing resource allocations, minimizing biases, and enhancing the data quality.

A detailed process of data analysis was employed, starting with the gathering and organizing of data, followed by machine learning algorithms with time series variation, including ARIMA, CrostonTSB Model, Prophet and various machine learning algorithms for prediction, such as XGBoost, LightGBM, LSTM and Random Forest. This comprehensive strategy guarantees a thorough analysis, effectively addressing the varied details of Siemens' product lines and market environments.

The findings reveal that with tailored modeling and the integration of market data, forecasting accuracy, as measured by RMSE, shows considerable improvement. The analysis also highlights the critical influence of seasonality and market indices on sales performance, underscoring the importance of incorporating these factors into forecasting models. The deployment of the hybrid forecasting models that combines traditional statistical methods with advanced machine learning algorithms capture the complex patterns in sales data. This initiative shows a notable enhancement in forecasting precision and capability.

Through the adoption of advanced analytics and machine learning, our project fulfills the requirements and further assists Siemens in attaining enhanced precision in forecasting. This enables more effective strategic decision-making and resource allocation, thereby boosting competitiveness in a quickly changing market.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. BUSINESS OBJECTIVES

Leading in Digital Transformation and Sustainability: Siemens aims to be the front-runner in digital industries, smart infrastructure, and mobility by harnessing technological innovations and sustainability. This objective is underpinned by Siemens' efforts to address the pressing challenges of urbanization, energy efficiency, and environmental sustainability through advanced solutions leveraging IoT, AI, and data analytics.

Integration of Advanced Technologies into Infrastructure: The core objective in the smart infrastructure market is to integrate cutting-edge technologies into physical infrastructures, thereby enhancing their efficiency, sustainability, and connectivity. This strategy is pivotal for tackling the challenges associated with rapid urbanization, the demand for energy efficiency, and the pressing need for environmental sustainability.

Expanding Market Position through Technological Innovations: Siemens is focused on expanding its market position within Germany and globally by pioneering technological innovations. This involves a commitment to research and development, with an emphasis on digitalization, electrification, automation, and renewable energy solutions.

Sustainability and Efficiency in Mobility Solutions: In the mobility sector, Siemens is committed to developing smart mobility solutions for rail and public transport that are not only efficient and reliable but also sustainable. This aligns with global trends towards reducing carbon emissions and improving urban mobility.

Enhancing Customer and Stakeholder Value: Siemens is dedicated to enhancing value for its customers and stakeholders by delivering integrated solutions that meet the demands of a changing global market. This includes improving operational efficiency, reducing time to market for new products, and embracing digital transformation across all sectors of its business.

2.2. PRELIMINARY INFORMATION

Market Growth Analysis: The global smart infrastructure market has experienced significant growth, underscored by a solid upward trend in urbanization, progressive technological advancements, and mounting concerns for environmental sustainability. The bar chart analysis reflects a Compound Annual Growth Rate (CAGR) of 20.5% from 2024 to 2030, signifying a critical demand for sustainable and efficient infrastructure solutions that can support the expanding urban landscapes and evolving global needs (Figure 1).

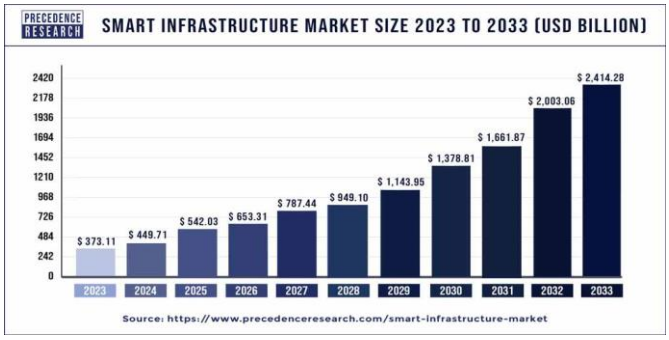


Figure 1 – Smart infrastructure market size in USD

Global Market Drivers and Restraints: The primary drivers bolstering this market growth include the rapid pace of urbanization and population growth, necessitating efficient resource management and the development of sustainable cities. Additionally, a series of government initiatives and breakthroughs in technology have been pivotal in advancing the growth of smart infrastructure solutions. Despite these drivers, the market faces challenges, particularly the rising threats to cybersecurity, emphasizing an indispensable need for secure and resilient smart infrastructure systems to mitigate risks and ensure dependable operations.

This analysis confirms the global smart infrastructure market's direction, characterized by dynamic growth, and influenced by pivotal trends and critical challenges. Companies operating within this space, including Siemens, can draw on this information to adapt strategies, anticipate market needs, and align their offerings with the forecasted growth and challenges ahead.

2.3. MARKET END-USERS

Non-Residential End-Users: Market Revenue Share (2023): Non-residential users held the largest share of the market revenue, which suggests a significant investment and adoption of smart infrastructure solutions in commercial and industrial settings;

Growth Factors: This dominance is due to the increasing need to digitize business operations, improve the consumer experience, and enhance digital security for sensitive data;

Key Solutions: The integration of automation systems is pivotal in this segment, promoting streamlined operations and effective resource utilization.

Residential End-Users: Projected Growth: The residential segment is expected to experience the highest growth rate in the forthcoming period;

Drivers of Growth: Demand is spurred by the adoption of smart home solutions like HVAC management, smart lighting, smart meters, and security systems such as smart door locks;

Additional Benefits: Increased awareness of smart grids, the expansion of the consumer electronics sector, rising personal incomes in developing countries, and improvements in power line communication also contribute to this segment's growth (Figure 2).

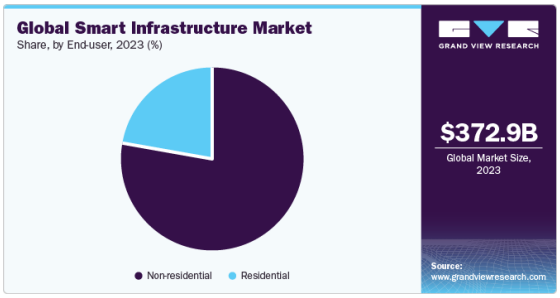


Figure 2 – Global smart infrastructure market

2.4. Competitive Analysis and SWOT Framework

Competitive Analysis: Siemens operates in a highly competitive landscape, with major competitors including ABB, Honeywell, and Schneider Electric in the smart infrastructure sector. The company differentiates itself through a comprehensive portfolio of products, systems, solutions, and services, emphasizing innovation, sustainability, and digitalization.

SWOT Analysis:

Strengths: Strong brand reputation, comprehensive technology portfolio, leadership in digitalization and sustainability initiatives, and global presence.

Weaknesses: Complexity of global operations, challenges in cybersecurity for smart infrastructure systems, and integration of new technologies.

Opportunities: Growing demand for smart infrastructure solutions, advancements in AI and IoT technologies, and increasing global emphasis on sustainability and energy efficiency.

Threats: Intense competition, rapid technological changes, cybersecurity threats, and geopolitical and economic uncertainties affecting global operations.

2.5. Strategic Recommendations for Siemens in the Smart Infrastructure Market

Considering the data and trends observed in the smart infrastructure market, the following strategic recommendations are proposed for Siemens AG:

1. **Enhancement of Non-Residential Solutions:** Siemens should continue to innovate and broaden its product and service offerings for the non-residential sector, capitalizing on the high demand for digitalized operations and secure infrastructure systems. By focusing on these areas, Siemens can strengthen its market position and provide compelling value propositions to commercial and industrial customers.

2. **Expansion in the Residential Segment:** Given the projected growth in the residential sector, Siemens is advised to intensify its development and marketing efforts in smart home technologies. Tailoring solutions to meet the increasing consumer needs for energy management and advanced security will likely capture the growing market segment and respond to the shifting consumer behaviors and expectations.
3. **Strategic Alignment of R&D and Marketing:** Siemens could harness the market dynamics to realign its research and development, as well as its marketing strategies, to target the most lucrative growth areas in the smart infrastructure market. Focusing investments in these directions will enable Siemens to innovate in line with market trends and consumer demands, ensuring long-term growth and competitiveness.

3. METHODOLOGY

3.1. DATA UNDERSTANDING

3.1.1. Data collection

In order to conduct this forecasting, two files were given to us: a CSV file related to Sales Data and a Market Data excel file. These were provided from Siemens Advanta Consulting team and consist of daily sales data from Siemens product groups and key market indices.

3.1.2. Data description

Regarding **Sales Data**, which corresponds to our training set, this file has a total of 9801 observations and is composed by three variables: “DATE” (daily data, to be converted to monthly so that it, eventually, corresponds to the test set), “mapped_GCK” and “Sales_EUR”, i.e., a datetime variable, a categorical and a numerical one, respectively. This CSV file contains data from October 2018 to April 2022.

The **Market Data** excel file enables an analysis and an overview of the key indices relative to macro-economic data in the most important countries for Siemens, such as Germany (biggest country of business unit), China, France, Italy, Japan, Switzerland, United Kingdom, United States and, in a more general approach, Europe and World. This CSV file comprehends monthly data from February 2004 to April 2022. Concerning the variables present on the file, the ones related to the indices can be checked bellow (Table 1), as well as the countries or region related to them. In total, initially without any data preparation, there were 48 variables and a total of 221 observations.

Index	Region/Country
"Production Index Machinery & Electricals"	China, France, Germany, Italy, Japan, Switzerland, United Kingdom, United States and Europe
"Shipments Index Machinery & Electricals"	China, France, Germany, Italy, Japan, Switzerland, United Kingdom, United States and Europe
"Price of Base Metals"	World
"Price of Energy"	World
"Price of Metals & Minerals"	World
"Price of Natural gas index"	World
"Price of Crude oil, average"	World
"Price of Copper"	World
"Producer prices on electrical equipment"	United States, United Kingdom, Italy, France, Germany, China (Producers)
"Production index on machinery and equipment"	United States, Switzerland, United Kingdom, Italy, Japan, France, Germany and World
"Production index on electrical equipment"	United States, Switzerland, United Kingdom, Italy, Japan, France, Germany and World

Table 1 – Indices per Region/Country available on the data

3.1.3. Data Exploration

After some initial data preprocessing, for instance, setting the date as the index, removing the row relative to the indices' codes (for example: 'MAB_ELE_PRO156'), since it was giving no additional and relevant information, the following exploratory analysis was conducted.

Firstly, considering correlation between variables. China's Production/Shipments Index Machinery and Electricals are highly correlated, in a positive way, with Producer Prices of Electrical Equipment from all the countries considered, except China ("Producer Prices.5"). There is a highly negative correlation between China's Production/Shipments Index Machinery and Electricals and its own prices on Electrical Equipment, i.e., the higher the indices, the lower the prices of the equipment.

A high positive correlation index was as well observed between "World: Price of Natural gas index" and "World: Price of Base Metals"; "World: Price of Crude oil, average" and "World: Price of Energy"; "World: Price of Copper" and "World: Price of Metals & Minerals". These relations all seem sensible, since, for instance, copper is a metal and crude oil is needed to produce some types of energy, thus if one increases, the other has the same tendency (Figure 3).

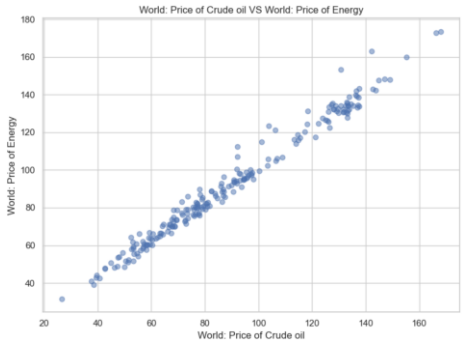


Figure 3 - World: Price of Crude oil VS World: Price of Energy

A high and positive correlation between Production Index of Electrical Equipment and Production Index of Machinery and Equipment, from a world perspective, was analysed. This same correlation was seen in Italy, Japan, and France, although not us much as high as in the World. This said, there is indeed a relation between machinery and equipment and electrical equipment (Figure 4).

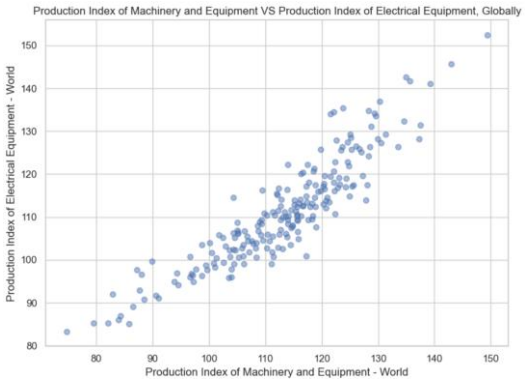


Figure 4 - Production Index of Machinery and Equipment VS Production Index of Electrical Equipment, Globally

Now having the biggest country of business unit, Germany, in mind and a period between January 2018 and April 2022 (approximately the period contained in our training data set), it is possible to say that there was indeed a pandemic effect, which can be observed in the graph below (Figure 5), since the prices of energy, metal, minerals and natural gas all had a significant increase on 2020, ending

up influencing slightly a bit the prices of electrical equipment production of Germany (more notable since 2021).

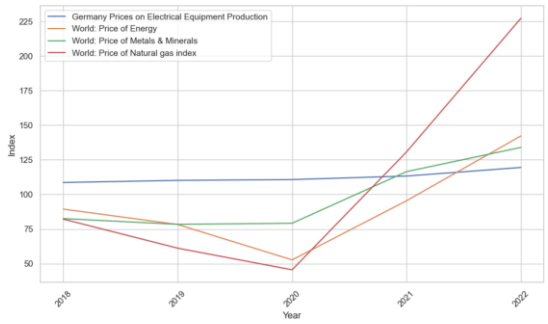


Figure 5 – Evolution of World Indices and Germany Prices on Electrical Equipment Production

Regarding the data about Simens' Sales, the group of products which sold the most between October 2018 and April 2022 was GCK 1, followed by 3 and then group 5 (Annex 1).

As can be seen on Figure 6, some of the highest points correspond to September sales, which is caused by the "September wonder" effect. This effect is due to the last day of September being the end of the fiscal year, when bonuses are defined, KPI's, and other important aspects, therefore, in this month companies and their employees do their best to sell the most.

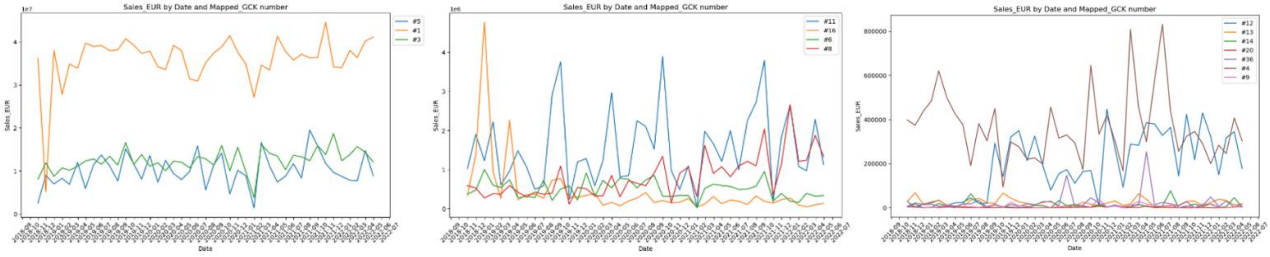


Figure 6 – Sales by GCK divided the order of magnitude, i.e., 1e7, 1e9, and the rest (October 2018 to April 2022)

3.1.4. Data Quality

About the data on "Market data", a very small number of missing values were detected, most of them related to producer prices, and no duplicates were found.

Talking about Sales data, 7134 observations out of 9801 had as values '0' and some negative numbers were shown (2.82% of the total values).

3.2. DATA PREPARATION

The initial step was to transform the 'DATE' column from a string format to a datetime format, specifying the day, month, and year arrangement. Sales data, stored as strings is converted to a floating-point format. A threshold for Sales Data is defined to treat very small sales values, which are practically insignificant, as zeros.

3.2.1. Understanding the impact of "Holidays" and "Weekends"

To recognize the impact of public holidays on sales patterns, a list of German public holidays is created and converted into a datetime format. Similarly, weekends are identified as periods which potentially have different sales patterns compared to regular weekdays. Furthermore, the sales data is aggregated on a monthly basis, excluding the identified holidays and weekends, to provide a clearer

view of sales trends over time. The final dataset, organized by month and year, is then merged with market data to create a full picture.

3.2.2. Feature Engineering

After merging the datasets relative to the market and the one from sales, a new variable was created. This variable had as objective to explore the effects of the pandemic (Covid Lockdown – from 22 March 2020 until around 17 March 2022) and the blockage of the Suez Canal (from 23 March 2021 until 29 March 2021). These were abnormal dates, since they affected prices, events and, for instance, demand globally.

3.2.3. Missing Values

We found some missing values in our merged dataset in columns representing sales figures and production indices for various countries (e.g., Switzerland, the United Kingdom, the United States) and specific market indicators such as Producer Prices. Subsequently, these missing values were imputed using the average of their respective columns.

3.2.4. Data Scaling

The MinMaxScaler is used to normalize feature scales to a range between 0 and 1, ensuring that all data points are uniformly evaluated by the model. This process adjusts the training data first by learning its scale and then applies the same scaling to the validation data to prevent data leakage. The scaled data is then converted back into a pandas DataFrame, maintaining the original structure with transformed values for efficient model training.

3.2.5. Feature Selection

In the data preparation process, significant features for each product were identified and selected based on their importance. This feature selection process was conducted through an automatic code which applies the methods to every GCK.

First, **Decision Tree Regressor** was applied on scaled training data, from which feature importances were extracted to determine the relevance of each feature to the product's sales predictions. Features deemed insignificant (with approximately zero importance) were excluded, allowing for a streamlined data set focused on the most influential factors. Secondly, **Recursive Feature Elimination** (RFE) was utilized to identify the most impactful features for predictive modeling in a regression context, using linear regression as the basis for feature evaluation. The process involved systematically creating models with a varying number of features, ranging from 1 to 30, to determine the optimal set that achieves the lowest Root Mean Squared Error (RMSE) on validation data, indicating the best predictive performance. Lastly, **Lasso Regression**, regularized linear regression, was developed in order to improve the selection, discarding useless or redundant features.

To sum up, for each GCK was selected the best group of features, according to their importance to each product.

3.3. MODELING

To obtain the forecasting of sales, different kinds of algorithms were considered. Different algorithms required different input data. With this, there were 2 categories of predictive models based on the required inputs: models using only sales values and models using market values as inputs.

Using Sales:

- **ARIMA:** The first model used was Auto-Regressive Integrated Moving Average (ARIMA). This model uses past data (past sales) to predict the future, accounting for trends (seasonality) (Hayes, 2022). The ARIMA algorithm takes the historical sales and performs smoothing and makes the predictions based on the order parameter (p, d, q). For each product, it was performed an ACF and PACF test (Autocorrelation function and Partial Autocorrelation function, respectively) (Ch 6. Model Specification Time Series Analysis Time Series Analysis Ch 6. Model Specification, n.d.). With the results of the tests and graphical analysis (see Annex 2 as an example), the model was created and the forecast was done.
- **Croston TSB:** The second model using only sales was a variation of the Croston (model used for forecasting in logistics). The Croston model was created to predict demand with small variation over time. This modelthat accounts for the spikes in demand (in this case sales) (Croston, 1972). It has been shown that the Croston model is outperformed by its variant TSB method (that accounts for when demand is 0) (Babai et al., 2019). For the purposes of this project, it was initially thought that the input values would be daily sales (with which this model works very well), however, it was noted that the input should be the monthly values as the expected output is in terms of monthly sales. The RMSE of this model shows that for 5 products it performed above all other models.
- **Prophet:** This model has only 2 inputs: time in days and sales of those days (Quick Start, 2022). A daily sales dataframe was used and the model was applied (as the documentation requires). It was noted that as a predictive model, this was not the best (it accounted very poorly for the big sales, or spikes), but it gave useful information about trends.

Using Market Values:

The following models required doing data partition, scaling and feature selection on the input data, consequently they were developed after doing these steps, which were explained above.

- **XGBOOST:** It focuses on optimizing a predefined objective function, utilizing techniques like gradient descent and regularization to improve model performance while controlling for overfitting. XGBoost's key features include parallelized computation, tree pruning, and built-in handling of missing values. This model was shown to be **the best one** in predicting the majority of the GCK's sales, since the RMSEs were the lowest in almost every GCK.
- **Light GBM:** This model is a light gradient boosting machine (light gbm). This means that it is a gradient boosting framework that uses tree-based algorithms (Welcome to LightGBM's Documentation! — LightGBM 3.3.5 Documentation, n.d.). In order to perform the forecasting of sales, the class LGBMRegressor was used. This allowed for a quick calculation and the model was fairly quick to run. It produced the second best results overall (annex 3) of the models of this category.
-
- **Random Forest**: This approach enhances predictive accuracy and reduces the risk of overfitting compared to a single decision tree. Combining Random Forest with TSCV allows for robust model training and evaluation in time series analysis, ensuring the reliability and generalization of the predictive model.
- **LSTM (Long Short-Term Memory):** By leveraging the memory capabilities of LSTMs to capture temporal patterns and the sequential nature of TSCV to validate model performance realistically, this approach enables the development of accurate and reliable forecasting

models for time series analysis. Slight changes had to be implemented to the input data, since this model requires a different shape of data relative to the one we were using.

3.4. EVALUATION

Upon analyzing the forecast results for each product, it's evident that they exhibit distinct behaviours over the predicted 10-month period. This observation suggests that the company may need to prioritize individual products and devise tailored strategies to address their unique challenges.

Working with lags also proved to be significant factors, as several market shifts only manifest in sales figures months later. This underscores the criticality of recognizing and accounting for these delays when forecasting and strategizing for the future.

That's why, for this model evaluation, it was needed to look for each product what it is the best. So, the general process was to look for the main indicator RMSE. Why this? Because is essential for forecasting because it quantifies accuracy, allows for comparison between models, identifies areas for improvement, informs decision-making, and effectively communicates forecasting performance.

After completing the process for each product, we could compare the different models which have the best RMSE (annex 3). This comparison showed XGBoost to have the higher number of products with the smallest RMSE.

4. RESULTS EVALUATION

In this case, XGBoost was considered as the best model overall (in almost every GCK), since it shows the best performance on RMSE (the lowest). However, this metric is still very close to the LightGBM one, as can be seen on Figure 7. Since in this problem the model had to be trained several times, a closer look to the evaluation metric was taken, coming to the conclusion that, most of the times, XGBoost behaviour was slightly better than LightGBM presenting, in fact, much less variance than LightGBM (shown through the boxplots referent to the model's RMSE, as the one on Annex 4 – in this last run the variance statement did not verify).

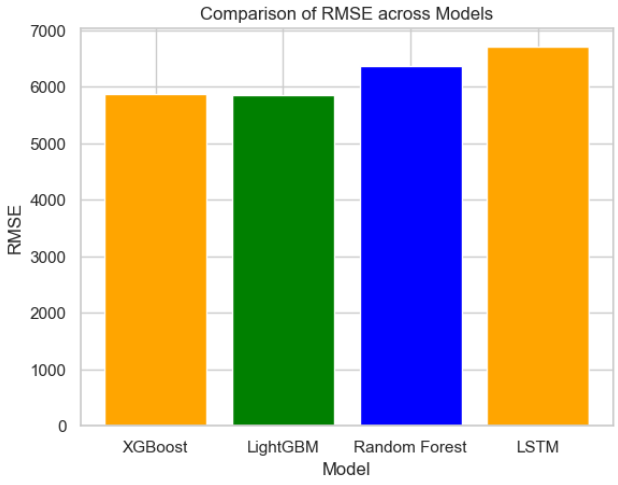


Figure 7 – Comparison of model's RMSE

Concerning the results of forecasting given by XGBoost and observing Figure 8, GCK #1 will stay consistent and still the most selling group of products, as happened on the actual data. At a lower level, #3 and #5 also show themselves to be very stable in terms of sales, being that sometimes #3 will have higher sales that #5, and vice-versa. The rest of the products have a way lower sales amount than the ones previously mentioned, some of the products being almost insignificant, since they seem to appear close to zero. This forecast shows that there is a stable trend overtime among most of the products, i.e., there is no prediction of major drops or peaks in sales., which can indicate effective sales strategies on the previous years, consequently, reflected on this forecast.

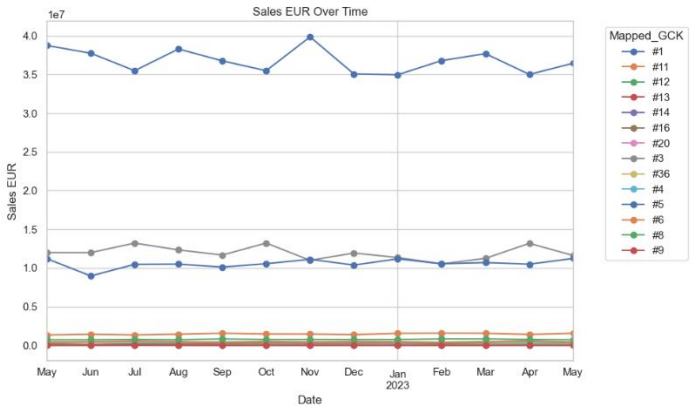


Figure 8 – Sales forecast overtime, in EUR (May 2022 to May 2023)

5. DEPLOYMENT AND MAINTENANCE PLANS

5.1 MODEL RELEASE MANAGEMENT CYCLE

Figure 9 – Release Management Cycle

1. **Request for Changes or New Features:**

Siemens identifies the need to transform their traditional manual sales forecasting methods for their Smart Infrastructure Division in Germany. The aim is to develop a more precise forecast model to enhance the accuracy and complexity of sales data, and to reduce the opportunity cost of poor forecasting.

2. **Release Planning and Design:**

In this phase, Siemens Advanta Consulting outlines the challenge's parameters: Focus on selected product groups and utilize sales data from October 2018 to April 2022 alongside key macroeconomic indices. The design of the forecasting solution must consider not only the data provided but also the complexities of real-world application, reflecting Siemens' operational needs. Models are expected to be both robust and adaptable, potentially requiring iterative refinements.

3. Software Build:

Several predictive models were implemented, each leveraging different data sets to forecast Siemens' sales figures: ARIMA, Croston TSB, Prophet, XGBoost, Light GBM, Random Forest and LSTM.

Each model's development involved deep coding, training, and tuning to align with the demands of Siemens' sales forecasting objectives.

4. Review

Significant efforts were invested in the refinement of each forecasting model to meet the criteria set by Siemens. Each model was diligently enhanced, with a comprehensive consideration of various sales drivers ensured by our team. The approach was carefully adjusted to ensure it was both reasonable and could be easily replicated, following the CRISP-DM framework suggested by Siemens.

5. Test

Models are tested using a separate test set provided by Siemens, covering sales data from May 2022 to February 2023. The primary metric for testing is RMSE, evaluating the accuracy of the sales forecasts generated by the AI models. The optimization process specifically targeted improving the precision and dependability of each forecasting model.

6. Deployment

The deployment phase will introduce our best-performing model into Siemens' operational environment for a trial period. This strategic launch is designed to test the model's forecasts within real-world business workflows, assessing its impact on decision-making processes. Should the testing show promising results, the model stands to offer substantial benefits, improving the efficiency and accuracy of Siemens' sales forecasting tools. This step is very important, as it might transition the model from a theoretical construct into a practical tool for the business.

7. Support

Following the model's deployment, any issues that may emerge from its operation in the real-world business context will be promptly addressed and resolved by our team. Adjustments and enhancements to the model will be made as necessary, ensuring its adaptation to the evolving dynamics of Siemens' business sectors and the changing landscape of economic conditions.

8. Issue Reporting and Collecting

Finally, Siemens would monitor the performance of the deployed model, gathering feedback and performance data. Issues and suggestions for improvements would be documented to inform future development cycles, potentially leading to further iterations of the forecasting model.

6. CONCLUSIONS

The Siemens Advanta Consulting sales forecasting project represents a significant advancement in improving prediction accuracy and strategic decision-making capability. The comprehensive analysis and modeling undertaken, which included a variety of traditional statistical methods and advanced machine learning algorithms such as ARIMA, Croston TSB, Prophet, XGBoost, LightGBM, Random Forest, and LSTM, provided deep insights into the complex dynamics of Siemens' product lines and market environments.

The project findings underscore the critical influence of seasonality, market indices, and external events such as the COVID-19 pandemic and the Suez Canal blockage on sales performance. By incorporating these factors into the forecasting models, the project achieved significant improvements in accuracy and prediction capability, as evidenced by the reduction in RMSE.

Upon analyzing the forecast results, it's evident that each product group exhibits unique behaviors. This highlights the importance of tailored strategies for individual products, considering factors like market shifts and sales data lags. For model evaluation, RMSE was crucial, with XGBoost emerging as the best model overall, closely followed by LightGBM. Notably, XGBoost demonstrated lower RMSE and variance compared to LightGBM. Regarding forecasting results, GCK #1 consistently led in sales, while #3 and #5 showed stability with occasional fluctuations between them. Other product groups had lower sales volumes, indicating a stable trend without major peaks or drops, reflecting effective past sales strategies.

The strategic recommendations derived from the analysis are aligned with Siemens' business objectives, emphasizing the importance of enhancing non-residential solutions, expanding in the residential segment, and strategically aligning research and development with market trends. These recommendations provide actionable insights for Siemens to capitalize on emerging opportunities and strengthen its position as a leader in digital transformation, sustainability, and technological innovation.

Overall, the project's success in delivering enhanced prediction accuracy enables Siemens to make more informed strategic decisions, optimize resource allocation, and enhance competitiveness in a rapidly evolving market landscape. As Siemens continues to innovate and adapt to changing market dynamics, the insights and methodologies developed through this initiative will be valuable assets for driving future growth and success.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

Feature Engineering: Explore additional relevant features that may capture underlying patterns or dynamics in the sales data. This could involve incorporating external factors such as economic indicators, weather data, or industry-specific metrics that could influence sales trends.

Hyperparameter Tuning: Optimize the hyperparameters of the models to improve their predictive accuracy. This includes parameters such as learning rate, maximum depth of trees, number of

estimators, and regularization parameters. Techniques like grid search or random search can be employed to systematically search through the hyperparameter space.

Ensemble Methods: Investigate the use of ensemble methods to combine predictions from multiple models. Ensemble techniques like stacking or boosting can often yield better results by leveraging the strengths of different algorithms and mitigating their weaknesses.

Model Interpretability: Enhance the interpretability of the models to provide insights into the factors driving sales forecasts. Techniques such as SHAP (SHapley Additive exPlanations) values or partial dependence plots can help explain the importance of different features in the prediction process.

Handling Seasonality and Trends: Incorporate methods to better capture seasonality and long-term trends in the sales data. This may involve using techniques like seasonal decomposition, detrending, or differencing to preprocess the data before modeling.

Model Interpretation: Enhance the interpretability of the models to provide insights into the factors driving sales forecasts. Techniques such as SHAP (SHapley Additive exPlanations) values or partial dependence plots can help explain the importance of different features in the prediction process.

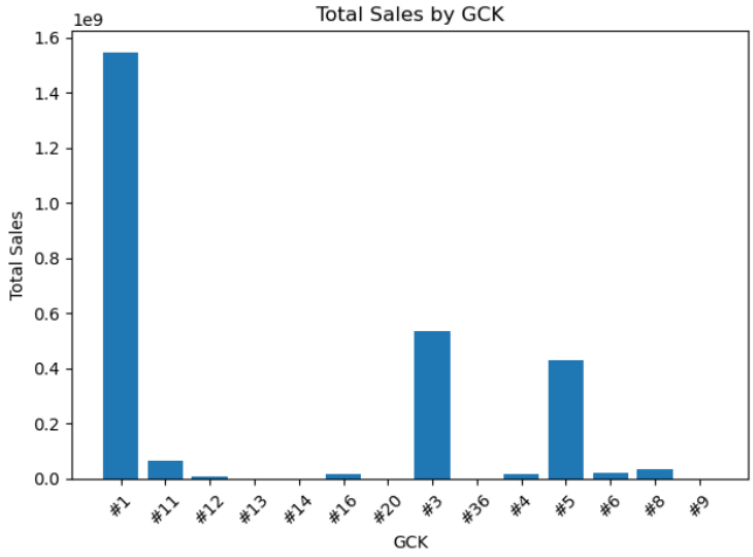
7. REFERENCES

- GfG. (2024, January 29). Time Series Analysis Visualization in Python. GeeksforGeeks. https://www.geeksforgeeks.org/time-series-data-visualization-in-python/
- Smart Infrastructure Market Size, Share & Trends Analysis Report by offering (Products, services), by type, by end-user (Residential, Non-residential), by region, and segment forecasts, 2024 - 2030. (2023, December 28). https://www.grandviewresearch.com/industry-analysis/smart-infrastructure-market-report
- Smart Infrastructure market size to hit USD 2,414.28 BN by 2033. (n.d.). https://www.precedenceresearch.com/smart-infrastructure-market
- Andrés, D., & Andrés, D. (2023, June 24). Clean your Time Series data III: Outliers removal - ML Pills. ML Pills - Machine Learning Pills. https://mlpills.dev/time-series/clean-your-time-series-data-iii/

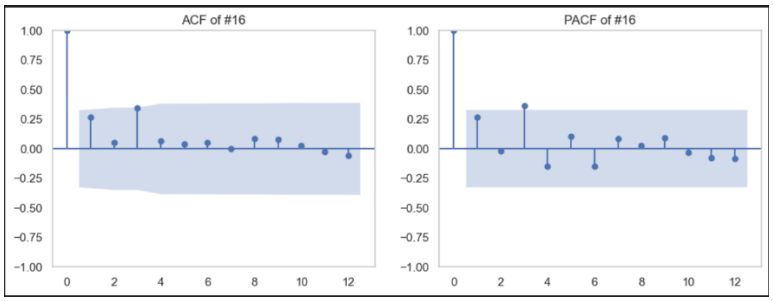
- Hayes, A. (2024, April 6). Autoregressive Integrated Moving Average (ARIMA) Prediction Model. Investopedia. https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp
- Croston, J. D. (1972). Forecasting and stock control for intermittent demands. Operational Research Quarterly, 23(3), 289. https://doi.org/10.2307/3007885
- Babaï, M. Z., Dallery, Y., Boubaker, S., & Kalaï, R. (2019). A new method to forecast intermittent demand in the presence of inventory obsolescence. International Journal of Production Economics, 209, 30–41. https://doi.org/10.1016/j.ijpe.2018.01.026
- Quick start. (2023, October 18). Prophet. https://facebook.github.io/prophet/docs/quick_start.html#python-api
- Ch 6. Model Specification Time Series Analysis Time Series Analysis Ch 6. Model Specification. (n.d.). Retrieved April 10, 2024, from https://people.missouristate.edu/songfengzheng/Teaching/MTH548/Time%20Series-ch06.pdf
- Siemens Report for fiscal 2023 (2023). Retrieved April 10, 2024, from https://assets.new.siemens.com/siemens/assets/api/uuid:be1828a9-2368-4c3b-a85f-f1bcb1f14a59/Siemens-Annual-Report-2023.pdf
- Welcome to LightGBM's documentation! — LightGBM 3.3.5 documentation. (n.d.). Lightgbm.readthedocs.io. https://lightgbm.readthedocs.io/en/stable/

8. APPENDIX

Annex 1 – Total Sales by GCK (October 2018 to April 2022)



Annex 2 – ACF and PACF test of product #16



Annex 3 – RMSE of each product for each model - the darker the blue, the smaller the RMSE value.

	ARIMA	CrostonTS	XGBoost	LightGBM	Random Forest	LSTM
#1	7364409	4938492	3419734,97	35440751,51	35943898,35	35582465
#11	777976,8	901822,9	989811,147	1544725,754	2170729,194	726970,3
#12	276479,8	266160	85194,40989	170101,8946	303286,911	92165,41
#13	11079,03	11258,81	14445,95709	18543,85699	30831,67594	13675,01
#14	15004,05	17838,86	14485,67838	7706,54521	17388,87561	12860,82
#16	1470684	759230,8	167866,0882	476991,947	338745,1672	139578,8
#20	3799,469	2720,558	2140,990732	6857,692936	7588,750506	2254,891
#3	4413599	3627575	3616849,122	12017644,62	12260426,61	10686712
#36	18362,53	15803,14	15844,85365	19694,39934	51578,17635	19382,03
#4	153960,6	140570,2	114280,0392	370772,3033	493274,0425	197197,8
#5	3000043	3089080	4025697,291	10017233,96	15587537,9	6212769
#6	316155,3	223599,4	248963,3574	512848,3545	623400,8561	401047,6
#8	1138245	1172356	707883,1299	690391,6715	1325129,362	874688,8
#9	5692,31	5320,187	5864,253498	5852,952251	6364,016661	6703,626

Annex 4 - Boxplots referent to the model's RMSE