MDSAA

Master's Degree Program in
Data Science and Advanced Analytics

Business Cases with Data Science

Case 2: Monthly Sales Forecast

André, Lourenço, number: 20240743

Emir, Kamiloglu, number: 20240945

Manuel, Andrade, number: 20240571

Rute, Teixeira, number: 20240667

Victor, Silva, number: 20240663

Group T

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

March, 2025

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

This report outlines a sales forecasting challenge conducted by Siemens Advanta Consulting. The goal is to develop an AI-driven model to improve monthly sales forecasts for selected product groups within Siemens' Smart Infrastructure Division in Germany. The challenge utilizes sales data and key macroeconomic indices to improve forecasting accuracy.

The primary objectives of this project are to enhance forecast accuracy for Siemens' German business unit, reduce the reliance on manual forecasting methods that are labour-intensive and prone to bias, and improve data integration across multiple sources to ensure more comprehensive and actionable insights.

The project addresses significant challenges in Siemens' sales forecasting, including fragmented data sources, resource-intensive processes, and potential judgment bias. Successful implementation is expected to improve forecasting precision, reduce opportunity costs, and enhance customer satisfaction. By leveraging AI-driven forecasting techniques, Siemens aims to streamline its sales prediction processes, enhancing both operational efficiency and strategic decision-making.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. Business Understanding and industry context

Siemens' Smart Infrastructure Division operates in a complex and dynamic market where sales forecasting plays a crucial role in effective planning and resource management. The company has identified critical business needs to improve sales forecasting in Germany.

The desired outcome is to implement an AI-driven solution that improves forecast accuracy, optimizes resource allocation, and reduces manual effort. The solution must seamlessly integrate sales data with key macroeconomic indicators to provide reliable and actionable insights. By achieving these outcomes, Siemens aims to enhance operational efficiency, minimize forecast errors, and improve decision-making processes.

2.2. Business Objectives

The primary objective of this project is to implement a robust forecasting model that enhances Siemens' ability to predict future demand accurately and efficiently.

By leveraging data-driven insights, this solution aims to improve decision-making processes, reduce manual efforts, and enhance operational efficiency.

Below are the key business objectives for this initiative:

1. **Accurate Demand Forecasting:**

   - Deliver precise insights into future demand, particularly over the next 10 months of sales, ensuring better alignment with business needs.

2. **Reduction in Manual Workload:**

   - Streamline processes to minimize resource-intensive tasks, reducing both staff allocation and time spent on forecasting activities. This will ease the burden on employees and improve productivity.

3. **Stakeholder Adoption & Satisfaction:**

   - Ensure the forecasting tool is widely accepted and effectively utilized by Siemens teams. The model should support stakeholders in making informed decisions regarding inventory levels, staffing, and financial planning.

4. **Unbiased Predictions:**

   - Rely on objective, data-driven insights to eliminate the influence of subjective projections from multiple stakeholders, ensuring more reliable forecasts.

5. **Operational Efficiency:**

   - Anticipate short-term operational needs to minimize opportunity costs and enhance resource allocation.

6. **Adaptability:**

   - Incorporate macroeconomic indicators to adapt to evolving market conditions, economic shifts, and changing customer behaviour patterns.

2.3. Business Success criteria

Our success criteria will be defined by the effectiveness of the forecasting model in improving demand prediction, optimizing inventory and production, ensuring adaptability, and providing valuable insights for strategic decision-making.

The key criteria include:

1. **Accuracy of Demand Forecasting:** The model should reliably predict low-demand periods and align with actual sales trends.

2. **Inventory Optimization:** Enhanced coordination with logistics to minimize overstocking and understocking.

3. **Production Efficiency:** The ability to adjust production based on demand forecasts, optimizing resource allocation.

4. **Scalability & Adaptability:** The solution should handle increasing data volumes and adjust to changing sales patterns or market conditions.

5. **Strategic Impact:** Providing actionable insights to improve sales strategies, marketing campaigns, and investment decisions.

2.4. Situation assessment

This project is being conducted within a two-week timeframe by a team of five, leveraging computing resources with a maximum of 32GB RAM.

Available Data:

The data set consists of **historical sales records and macroeconomic indicators**, providing a foundation for building an AI-driven forecasting model.

Sales Train Data**:** Covers daily sales from October 2018 to April 2022.

Macroeconomic Indicators**:** Include external factors like energy prices**,** industrial production, and commodity prices, which influence Siemens' sales performance.

Sales Test Data**:** The group must submit the test file with the corresponding values, with the main evaluation metric for the model being RMSE (Root Mean Square Deviation).

This data will serve as the basis for feature engineering, model training, and performance evaluation.

Constraints and Limitations:

Throughout the project, we encountered several constraints that posed challenges in understanding and working with the provided data. Firstly, our limited industry knowledge made it difficult to immediately grasp the context and significance of certain data points, requiring additional effort to bridge this gap. Secondly, the **handover documentation provided by the client lacked sufficient detail and clarity**, which led to potential ambiguities in the project scope and required additional communication to ensure alignment.

Lastly, **the absence of comprehensive metadata** further complicated our initial exploration of the dataset. Without clear descriptions of variables, categories, and data structures, significant time was spent deciphering the data before meaningful analysis could begin. These constraints collectively added complexity to the early stages of the project, requiring proactive communication and deeper investigation to ensure accurate insights and effective forecasting.

2.5. Determine Data Mining goals

For this project, the evaluation criteria was defined by the client: results on the test dataset will be evaluated based on the score of Root of Mean Squared Errors (RMSE). This machine learning metric measures the square root of the residual errors, penalizing larger errors more.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}i - yi)^2}{n}}$$

Additional data mining goals for this project are:

- Identify potential improvements, specifically features or data to include on the model for optimized performance

3. METHODOLOGY

3.1.1 Data Overview

We began our analysis with two main datasets. The first, ***Case2_market data***, consists of **219 records and 48 columns**, capturing a broad range of indicators such as production and shipment indices for machinery and electrical equipment across various countries—namely, China, France, Germany, Italy, Japan, Switzerland, the United Kingdom, and the United States—along with aggregated European and global figures. It also includes important macroeconomic variables such as commodity prices (e.g., base metals, energy, crude oil, and copper) and producer prices, which provide valuable context for market trends.

The second dataset, ***Case2_sales data***, contains **602 records and 3 columns**, displaying daily records of sales in euros by group product sold, from October 2018 until April 2022.

3.1.2 Exploratory Data Analysis (EDA) Insights

To kickstart our EDA, we began by standardizing both datasets, cleaning out any non-descriptive text or rows, aggregating sales data into monthly records and formatting datetime features. Next, we plotted both datasets to identify patterns, trends, and seasonality. Market data did not display any noticeable cyclical pattern - data evolves quite irregularly over the years and months.

As for sales data, we can see significant fluctuations, with a more cyclical trend over the time span, as shown in the figure below. We can observe that:

- September and March-April are likely peak seasons
- December-January are likely low seasons

Overall, there is a somewhat constant fluctuation throughout the considered time span.

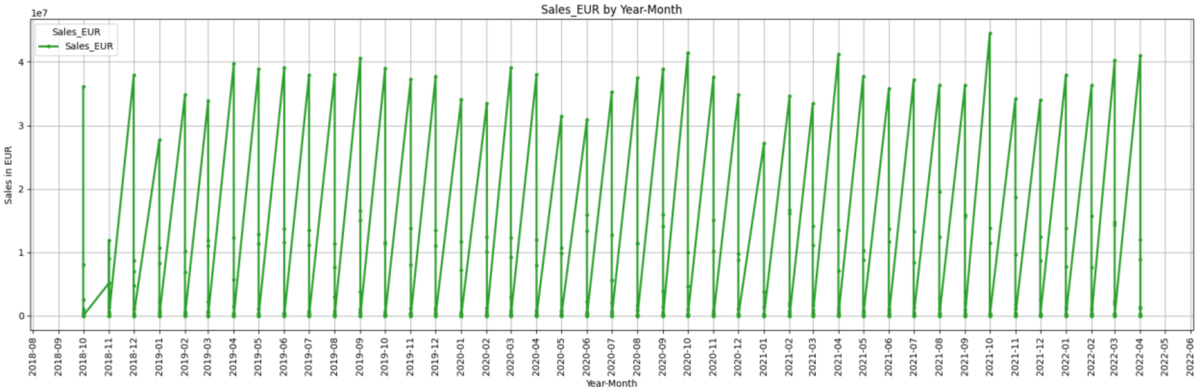

Figure 1 - Sales EUR by Year- Month

For a complete analysis, we plotted sales evolution by product group to better understand potential trends and patterns in spending preferences during the considered time-span. This uncovered clear trends and behavior within each product segment:

- Product group 1 is predominantly the most purchased
- Product group 3 and 5 display disputing demand, although group 3 has higher revenues
- All remaining product groups play a more residual part in turnover

Across product groups for the considered time-span, we can notice overlapping low and high seasons with overall sales performance.

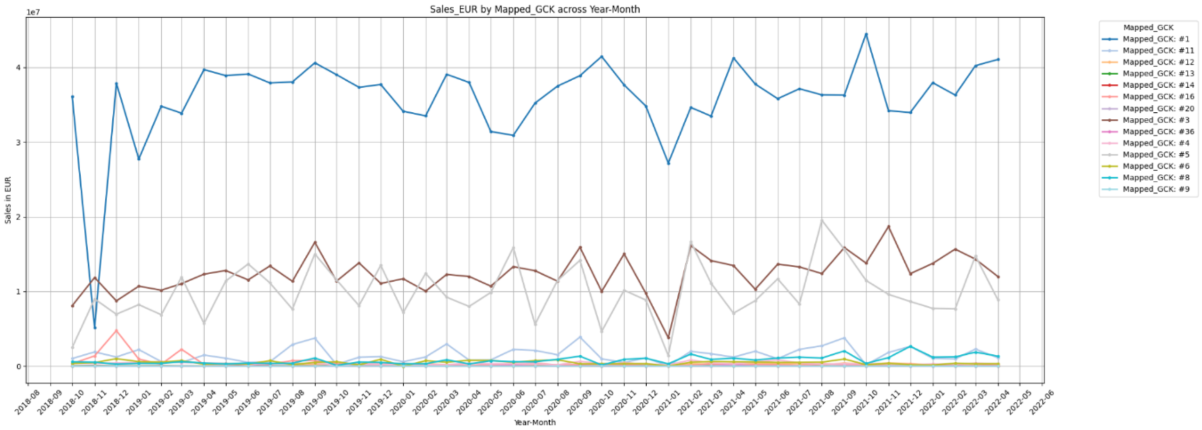

Figure 2- Sales EUR by Mapped GCK across Year Month

3.2. Data preparation

3.2.1 Handling Missing Values:

In this section, we identified multiple features with missing values, covering both historical data and the active sales period. We address the handling of missing values within the dataset, specifically focusing on market data, as sales records only spans three and a half years. The strategy adopted varies depending on the timeframe and pattern of the missing data.

- **Missing Values before 2018 - Pre-Sales Record:**

Some features exhibited missing values for periods predating the sales data (before 2018), notably between 2004-2006. Since these data points fall outside the relevant sales period, or any lag period projectable, they were deemed unnecessary for our analysis. Consequently, we excluded all data before 2006 from the dataset.

- **Missing Values during Sales Record period - 2018-2022:**

Single-Period missing:

Proceeding to features with missing values during sales records period, we started by addressing features who had missing values for a single-month period - coincidentally, these five features all had missing data in April 2022, our last record of sales data. For that reason, simple interpolation techniques like averaging surrounding points were not feasible. Instead, we employed a **Polynomial Regression Model** to predict and impute the missing data based on recent trends, given that older data is less likely to capture the evolving market conditions.

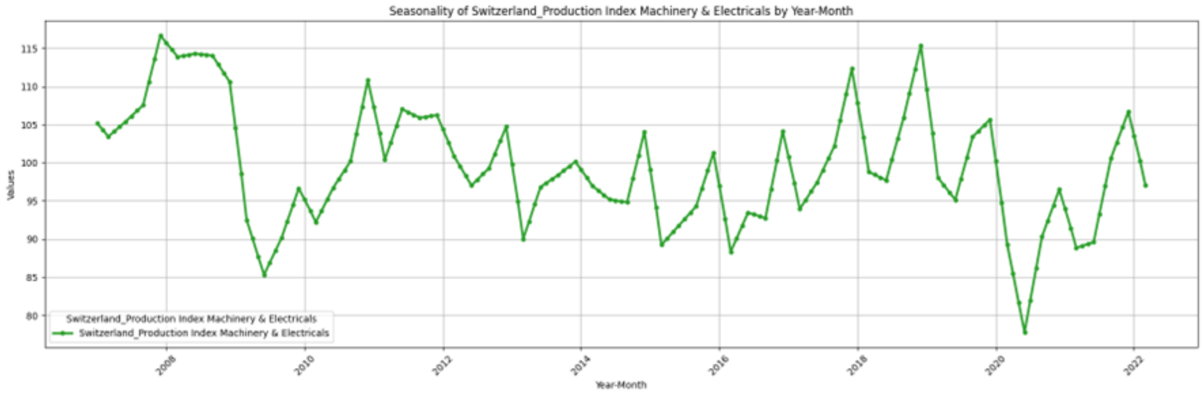

Figure 3- Seasonality of Switzerland production index machinery & electricals by year-month

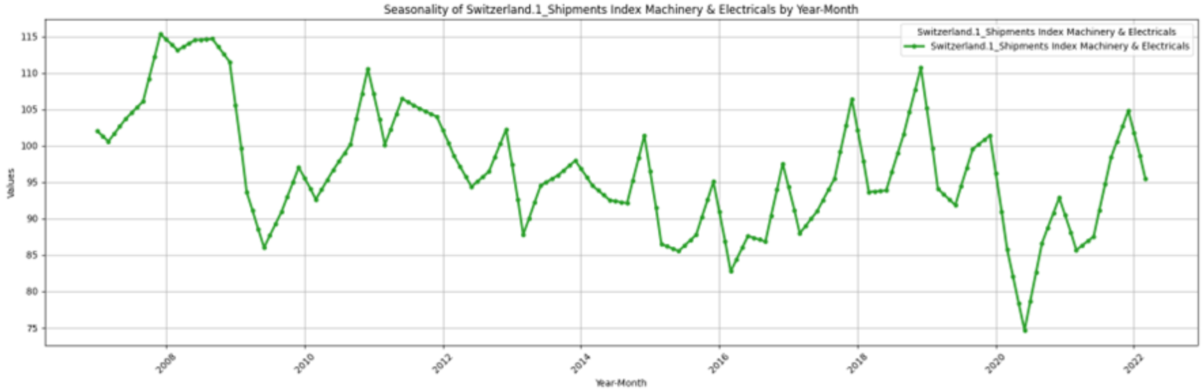

Figure 4- Seasonality of Switzerland.1 Shipments Index Machinery & Electricals by Year-Month

Longer Consecutive Periods Missing

Lastly, we identified prolonged gaps in two United Kingdom-related features between November 2020 and April 2022. Upon examining feature correlations, we identified strong relationships with some Germany-based indicators.

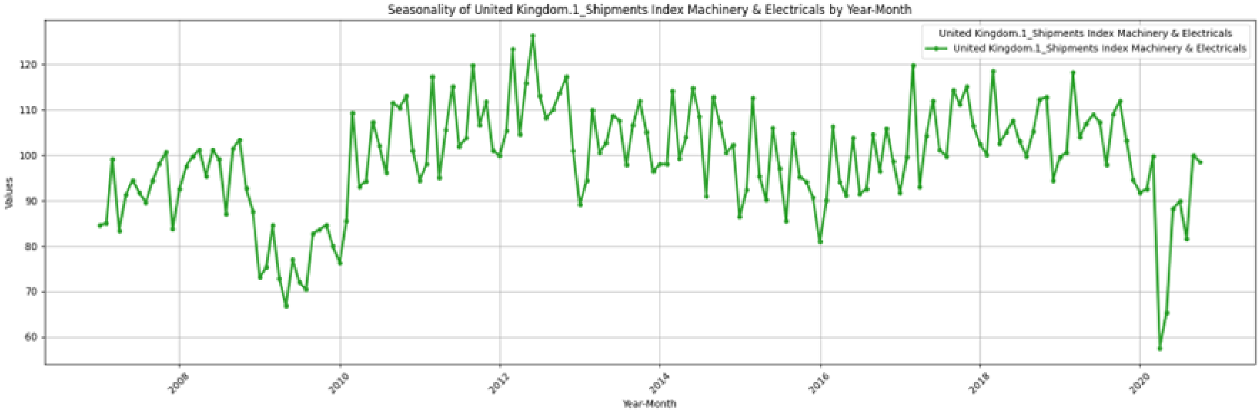

Figure 5- Seasonality of United Kingdom.1 Shipments Index Machinery & Electricals by Year-Month

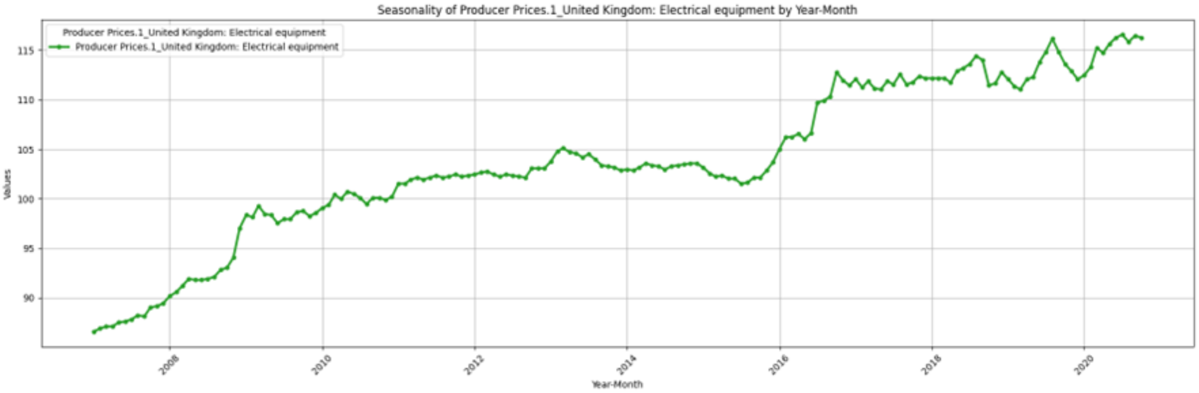

Figure 6- Seasonality of Producer Prices.1 United Kingdom: Electrical equipment by Year-Month

Given the strong correlation, we projected missing values using the corresponding German features as proxies. This method ensured that the imputed values retained meaningful trends and patterns reflective of real market behaviour.

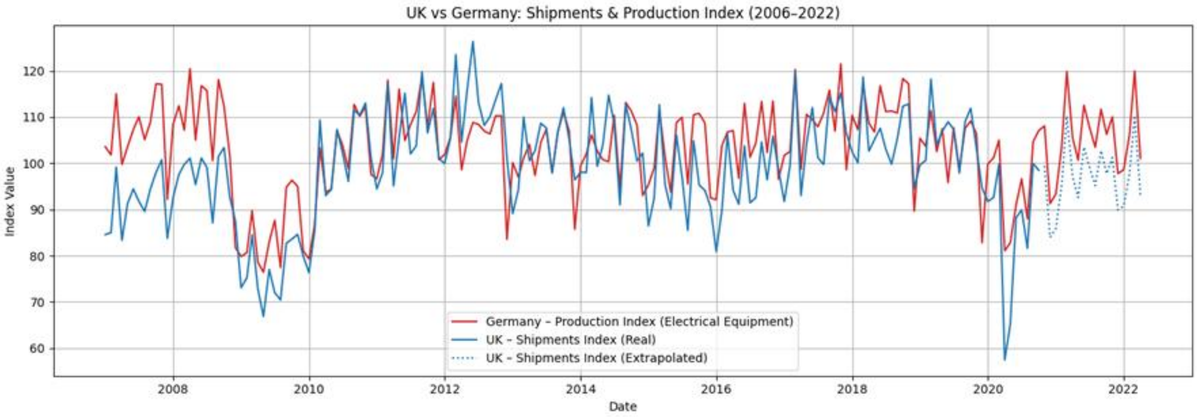

Figure 7- UK vs German: Shipments & Production Index (2006-2022)

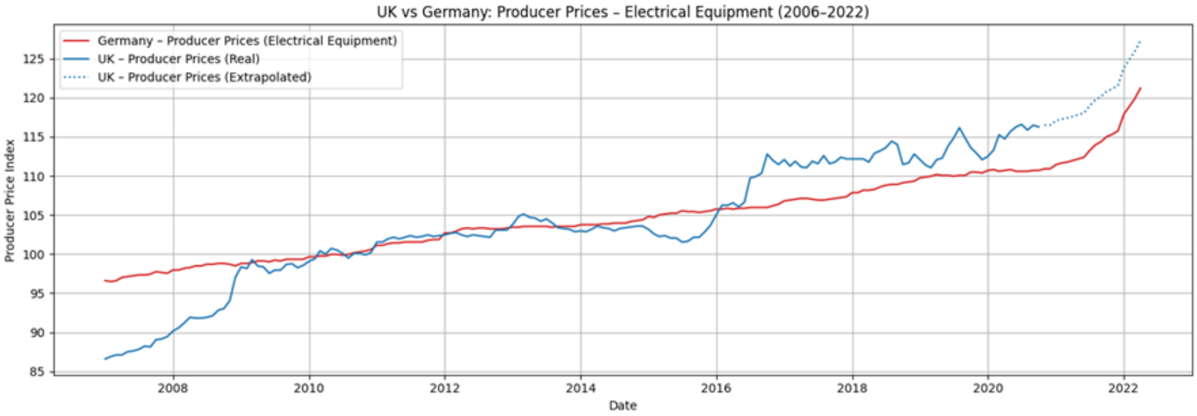

Figure 8- UK vs Germany: Producer Prices - Electrical Equipment (2006-2022)

By strategically removing irrelevant data, employing regression-based predictions for isolated gaps, and utilizing correlated features for longer gaps, we effectively mitigated the impact of missing values on our analysis. This comprehensive approach ensures data consistency and strengthens the reliability of our sales forecasting model.

3.2.2 Feature Engineering:

We identified several external market features that could improve the model's ability to interpret and predict sales trends. The selected features were chosen for their potential to influence Siemens' sales performance. By incorporating indicators such as stock prices, interest rates, and inflation, we aimed to capture broader economic patterns that may impact customer purchasing behavior and industry trends. These features are sourced from reputable platforms such as **Fred**, **Investing.com**, and the **People's Bank of China (PBoC)** to ensure data reliability and accuracy.

Below is a detailed explanation of each feature:

Feature	**Description**	**Source**
Siemens AG Monthly Stock Prices	Tracks Siemens AG's stock performance	Investing.com - Stock Market Quotes & Financial News
SPY ETF Stock Price History	Represents the S&P 500 ETF	Investing.com - Stock Market Quotes & Financial News
Shanghai Shenzhen CSI 300 Historical Data	Tracks the performance of the top 300 stocks traded on the Shanghai and Shenzhen stock exchanges, reflecting China's market conditions.	Investing.com - Stock Market Quotes & Financial News
US Unemployment	Measures the percentage of unemployed individuals in the US	FRED
FED Balance Sheet Size	Represents the total assets held by the Federal Reserve	FRED

US M2	Represents the total money supply in the US	FRED
EUR/CNY Exchange Rate	Tracks the exchange rate between the Euro and Chinese Yuan	People's Bank of China (PBoC)
Chinese Monthly Inflation	Measures inflation levels in China	People's Bank of China (PBoC)
Chinese Short-Term Interest Rate	Tracks short-term borrowing rates in China, reflecting economic policy changes.	People's Bank of China (PBoC)
US 10Y Treasury Yields	Represents long-term US government bond yields	Investing.com - Stock Market Quotes & Financial News
US 1Y Treasury Yields	Represents short-term US government bond yields	Investing.com - Stock Market Quotes & Financial News
Siemens AG Monthly Stock Prices	Tracks Siemens AG's stock performance.	Investing.com - Stock Market Quotes & Financial News

Also, we introduced the count of holidays per month of each country on the dataset since they significantly impact consumer behavior and can help explain seasonal demand patterns throughout the year. Finally, we identified and removed 19 redundant features to reduces dimensionality, improves model generalization, and minimizes the risk of overfitting.

3.2.3 Lag Analysis:

Recognizing that market indicators often have delayed effects on sales, we conducted a lag analysis to determine the optimal lag period for each indicator. This accounts for factors such as economic cycles, contract timelines, and inventory to production delays.

By examining correlations between sales and lagged versions of market indicators, we identified the most influential lag periods for each GCK. Below, we present the top three correlations for the three most significant GCKs, with their respective lag periods and correlation values:

GCK #3 – Interest Rates and Production Index Impact:

➞ Chinese Short-Term Interest Rates_norm | Lag: 11 | Corr: 0.62
➞ China Production Index: Machinery & Electricals | Lag: 11 | Corr: 0.59
➞ United Kingdom Production Index: Electrical equipment | Lag: 8 | Corr: 0.55

GCK #8 – Global Production and Financial Indicators:

➞ China Production Index: Machinery & Electricals | Lag: 45 | Corr: 0.74
➞ Average 1Y Yields_norm | Lag: 57 | Corr: 0.74
➞ SPY Average Stock Price_norm | Lag: 47 | Corr: 0.72

GCK #12 – Producer Prices and Stock Market Correlations:

➞ Producer Prices in Italy: Electrical equipment | Lag: 33 | Corr: 0.73
➞ Producer Prices in France: Electrical equipment | Lag: 32 | Corr: 0.72
➞ SPY Average Stock Price_norm | Lag: 22 | Corr: 0.72

By aligning market indicators with the appropriate lag periods, we significantly enhanced the predictive power of the model. The results highlight the importance of conducting lag analysis on a GCK basis, as evidenced by the strong correlations observed across key features. This customized approach improves our model's ability to capture the delayed effects of market changes on Siemens' sales performance. At the end of this section, a dictionary was created containing 14 data frames, each corresponding to a specific GCK, along with its respective explanatory variables and their corresponding lags.

3.2.4 Feature Selection:

To improve the accuracy and efficiency of our sales forecasting model, we implemented a structured feature selection process. This approach incorporated multiple statistical and machine learning techniques to identify the most relevant predictors of sales.

To ensure robustness, we applied a voting mechanism across the methods used (Lasso, Ridge, Decision Tree, and Variance). Each method ranked the top 10 features based on its selection criteria, and only the features that appeared in at least two methods were retained. This strategy helped to leverage the strengths of various selection techniques while reducing the biases of individual methods.

The final feature selection was applied on a per-GCK basis, and new data frames were created containing only the selected features, along with "Year-Month" and "Sales_EUR." This process ensured that each dataset retained only the most relevant predictors, improving both the model's interpretability and efficiency. At the end of this section, the existing dictionary was updated, and the data frames for each GCK contained only the features resulting from the feature selection process.

3.3. Modeling

For the sales forecasting task, we adopted an ensemble approach, combining predictions from two distinct models: Neural Prophet and LSTM (Long Short-Term Memory). This strategy aimed to leverage the strengths of both models, increasing the robustness and accuracy of the forecasts.

Neural Prophet

Nueral Prophet deep learning-based time series forecasting tool, was trained separately for each GCK (product group) to capture seasonal patterns, long-term trends, and include macroeconomic variables as external regressors. The model configuration included:

- Autoregressive lags on the target variable: 12 months
- Forecast horizon: 1 month
- Yearly seasonality: Enabled
- Weekly and daily seasonality: Disabled
- Macroeconomic regressors: Included based on GCK-specific feature selection
- Data preprocessing: MinMax scaling
- Dataset split: Fixed cutoff in December 2020

LSTM

In parallel, we implemented a forecasting model using LSTM neural networks, known for their ability to learn complex temporal dependencies, particularly in noisy and nonlinear time series. Each GCK was trained individually with the following setup:

- Autoregressive lags on the target variable: 12 months
- Forecast horizon: 1 month
- First LSTM layer: 128 units
- Second LSTM layer: 64 units
- Loss Function: Huber Loss
- Optimizer: Nadam with learning rate = 0.001
- Weekly and daily seasonality: Disabled
- Macroeconomic regressors: Included based on GCK-specific feature selection
- Data preprocessing: MinMax scaling
- Dataset split: Fixed cutoff in December 2020

Model Ensemble

After training both models, we combined their predictions using a simple average ensemble, assigning equal weights (50% each) to the LSTM and Neural Prophet models. This balanced approach allowed us to benefit from the strengths of both models—Neural Prophet's ability to capture seasonality and trends, and LSTM's effectiveness in modeling complex temporal dependencies.

This ensemble strategy was applied per GCK. The existing dictionary structure was updated so that each key (representing a GCK) contained the cleaned and lagged dataset, the trained Prophet and LSTM models, their individual forecasts, and the final ensemble prediction.

3.3. Evaluation

The performance of each model was evaluated using Root Mean Squared Error (RMSE), a standard metric for assessing forecast accuracy. RMSE values were calculated for each GCK based on the difference between the predicted and actual sales values.

In addition to the numerical evaluation, we also conducted a visual analysis by plotting the actual versus predicted values over time.

These comparison plots helped us qualitatively assess the alignment between forecasts and real sales, identify potential seasonal mismatches, and visually detect under- or over-prediction patterns. This dual approach—quantitative and visual—ensured a more comprehensive evaluation of the model's performance across different product groups.

Individual GCK Performance:

GCK´s	RMSE
1	0.079261
3	0.165651
4	0.091784
5	0.217897
6	0.225564
8	0.324373
9	0.125743
11	0.298398
12	0.244759
13	0.177309
14	0.310294
16	0.020046
20	0.097017
36	0.062167

$$\boldsymbol{Average\ RMSE = 0.1743}$$

4. Results Evaluation

The primary objective of the model is to generate accurate demand forecasts over a 10-month horizon to support business decision-making. The model demonstrated reasonable predictive accuracy, with an average RMSE of 0.1743. However, certain product segments (e.g., **GCK_11** and **GCK_8**) exhibited higher error rates, indicating areas for refinement.

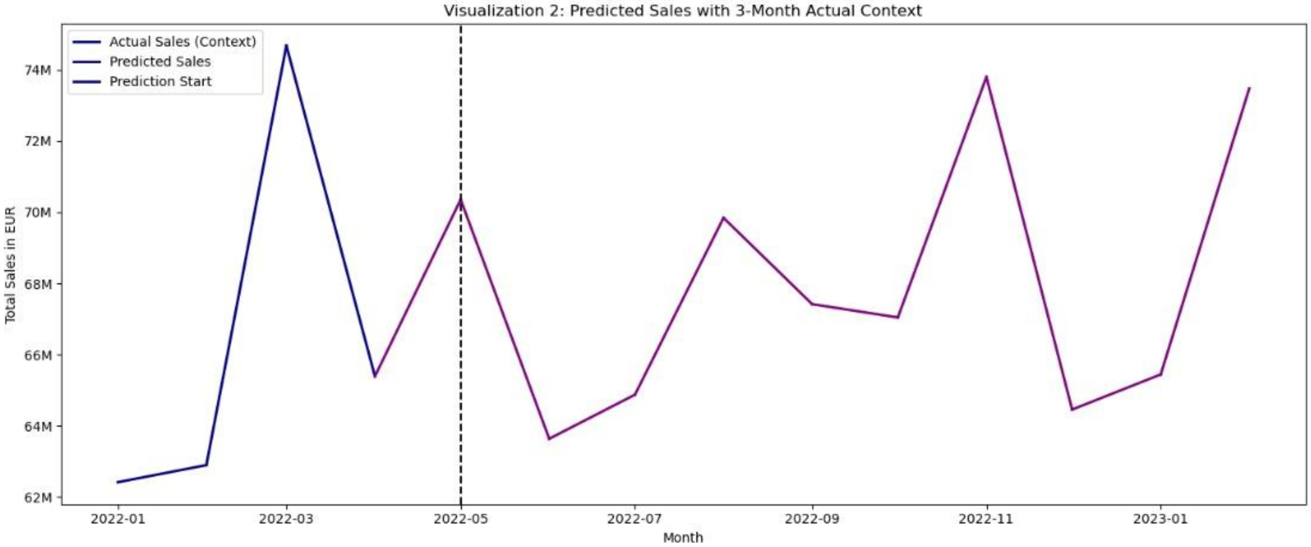

Figure 9 - Actual Sales vs Predictions 1

To fully evaluate the model's effectiveness, it should be deployed in a live forecasting environment, where predictions can be systematically compared against actual sales outcomes. This will validate its accuracy, identify areas for improvement, and assess its practical business impact.

Areas for Improvement:

- **Feature Selection Refinement:** Certain segments may require more targeted feature selection to enhance predictive accuracy.
- **Enhanced Data Inputs:** Incorporating additional historical data or external factors (e.g., macroeconomic indicators, market trends) may improve model performance.
- **Alternative Modeling Approaches:** Exploring hybrid models or incorporating exogenous variables could enhance forecast reliability, particularly for high-variance product segments.

Real-World Validation Approach:

To assess the model's business impact, a structured validation process is recommended:

- **Controlled Experiments:** Use forecast-driven decision-making to evaluate key business KPIs such as stock optimization and revenue impact.
- **Stakeholder Feedback:** Engage business teams to refine model assumptions and improve interpretability.

Despite limitations in certain segments, the model provides a strong foundation for data-driven sales forecasting. Future enhancements could focus on refining feature engineering, testing additional forecasting techniques, and integrating real-time market data for adaptive forecasting.

4. Deployment and Maintance plans

4.1 Deployment Strategy

The sales forecasting model will be deployed on Siemens' cloud or on-premises infrastructure, integrating with existing ERP and CRM systems. The deployment process includes infrastructure setup, system integration, stakeholder training, and validation through A/B testing. Key teams involved are Data Science, IT & DevOps, and Business & Sales teams. Once validated, the model will be documented and made accessible via APIs or automated reports.

4.2 Model Monitoring and Maintenance

Post-deployment, the model’s accuracy will be continuously monitored by comparing predictions with actual sales. Automated alerts will detect anomalies, and periodic retraining will ensure adaptability to market changes. IT teams will manage system maintenance, while business users will provide feedback for continuous improvement. This strategy ensures the model remains reliable and valuable for decision-making.

5. Conclusions

The implementation of the sales forecasting model for Siemens successfully addressed the core objectives defined at the outset of the project. By leveraging historical sales data alongside relevant macroeconomic indicators, the team developed a robust and product-specific forecasting solution, delivering more accurate and actionable insights aligned with Siemens' operational and strategic needs.

The primary objective of accurate demand forecasting was effectively achieved, **with the model attaining an average RMSE of 0.1743**. This demonstrates the model’s capability to reliably anticipate future sales trends over the defined 10-month horizon, thereby enhancing planning and alignment with business requirements.

Additionally, the automation of the forecasting process significantly reduced the reliance on manual, resource-intensive tasks. This improvement not only streamlined internal workflows but also alleviated the burden on staff, allowing teams to redirect efforts toward higher-value strategic activities.

Designed with user adoption in mind, the model offers clear and interpretable forecasts tailored by product group, directly supporting key stakeholders in making informed decisions regarding inventory management, workforce planning, and financial forecasting. Its data-driven nature ensures unbiased predictions, eliminating subjective judgment and promoting trust in its outputs.

The forecasting tool further enhances operational efficiency by enabling the anticipation of short-term resource demands and reducing opportunity costs through improved allocation. The incorporation of macroeconomic indicators—such as interest rates, inflation, industrial output, and exchange rates—grants the model adaptability to market fluctuations and evolving customer behavior.

In conclusion, **the forecasting solution provides a strong foundation for Siemens’ transition to data-driven sales planning**. **It not only meets the initial business goals but also establishes a scalable and adaptable framework for future forecasting initiatives**. With future refinement and validation in a live business environment, this solution has the potential to generate substantial value by improving the accuracy, efficiency, and reliability of sales forecasting at Siemens.