Master's Degree Program in

MDSAA

Data Science and Advanced Analytics

Business Cases with Data Science

Case 2: SIEMENS Sales Forecast

Ana Caleiro, 20240696
Duarte Marques, 20240522
Moeko Mitani, 20240670
Oumayma Ben Hfaiedh, 20240699
Sarah Leuthner, 20240581

Group Q

1. INTRODUCTION

Forecasting is a very important part of any multinational company. The ability to have an idea of how to prepare for the future helps guarantee that the company will not face problems related to over- or understocking. To achieve this goal, it is necessary to look not only at the company's data, but also at external factors that will have an impact on the final figures.

Given the difficulty of the task, humans are no longer the go-to to make these predictions: the complexity and the sheer amount of data makes it borderline impossible for a human to correctly analyze the patterns in the data in sufficient time. To reduce the probability of mistakes in these predictions, and to avoid the introduction of bias, Machine Learning models are used to help with the forecast.

In this project, following the CRISP-DM framework, and utilizing algorithms like XGBoost and ARIMA, we aim to predict the sales for specific product groups for the next 10 months of the company SIEMENS. With these results, we can build a business strategy, highlighting the implementation of the model, maintenance and how to improve it over time.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

The accurate forecasting of sales is a crucial aspect of business operations, especially in industries that deal with electrical components like switchgear products in smart power distribution. These products, which range from medium-voltage to low-voltage, are essential in ensuring efficient and safe power distribution in modern electrical grids.

Traditional sales forecasting methods often rely on manual processes that are time-consuming, resource-intensive, and prone to biases. Additionally, these methods require consolidating data from multiple sources, making them susceptible to inefficiencies and inaccuracies. Given the dynamic nature of market demand and the influence of macroeconomic factors, businesses must adopt more advanced forecasting techniques to remain competitive.

The project aims to predict monthly sales forecasts from May 2022 to February 2023 using historical sales data spanning from October 2018 to April 2022, along with significant macroeconomic indices. Leveraging AI-driven forecasting models will enable more precise and data-driven sales predictions, ultimately enhancing business decision-making and operational efficiency.

2.2. BUSINESS OBJECTIVES

The primary objective of this project is to develop an AI-driven sales forecasting model that provides accurate and reliable monthly sales predictions. This will help the organization achieve the following goals:

1. **Optimize Resource Allocation**: Reduce reliance on manual forecasting, saving time, money, and human effort.
2. **Minimize Bias**: Eliminate subjective influences by leveraging data-driven decision-making.

3. **Enhance Data Utilization**: Integrate and analyze data from multiple sources, ensuring a more comprehensive and dynamic forecasting process.
4. **Improve Forecasting Accuracy**: Reduce errors in sales predictions, leading to better inventory management and production planning.
5. **Support Digital Transformation**: Enable the adoption of AI-based tools to enhance operational efficiency and agility in response to market changes.

2.3. BUSINESS SUCCESS CRITERIA

The success of this project will be evaluated based on the following criteria:

- **Forecasting Accuracy** (metric: RMSE): Achieving a high level of accuracy in monthly sales predictions compared to actual sales figures.
- **Reduction in Forecasting Time**: Decreasing the time required for sales forecasting through automation.
- **Scalability and Flexibility**: The ability of the AI-driven forecasting model to adapt to future changes in market conditions and business needs.

By implementing an AI-driven sales forecasting approach, the company can significantly improve its decision-making capabilities, reduce inefficiencies, and gain a competitive advantage in the smart power distribution industry.

3. METHODOLOGY

In this project, the CRISP-DM model is utilized to understand business problems, to deploy the model, and to ensure the successful execution of this project.

3.1. DATA UNDERSTANDING

We used three datasets for this project: a csv file *Case2_Sales data* that contains past sales data (October 2018 to April 2022 - 43 months), an excel file *Case2_Market data* that contains important macro-economic indices (18 years 3 months- 218 months), and a csv file *Case2_Test Set Template* (May 2022 to February 2023 - 10 months) containing the template we want to fulfill with the final results used to test and evaluate our model by the stakeholders.

3.2. KEY FINDINGS AND TRENDS

The macroeconomic indices include various shipping and production indices per country as well as worldwide indices for resource prices. All indices were normalized with 2004 being the base year. We noticed that all countries except for China follow the same pattern across time in production and shipment indicators. However, China follows an upward trend which further proves its evolution until it became the largest market in terms of economic expansion.

For sales data, it was observed that product category 1 followed by product categories 3 and 5 were sold the most, however product category 14 sold the least in the dataset with some months registering 0 in terms of sales.

3.3. DATA PREPARATION

3.3.1. ADF Test

To ensure our dataset's suitability for time series forecasting, we conducted the Augmented Dickey-Fuller (ADF) test, a statistical test that determines whether a time series is stationary. This is a crucial step, as the stationarity of the data influences the choice of forecasting models. The test results confirmed that our dataset is stationary, allowing us to apply time series forecasting models such as ARIMA or XGBoost for time-series.

3.3.2. Feature Engineering

We have created four new features (*year, month*, *month_sin*, and *month_cos*). Since a feature like month has a cyclical pattern, it is important to encode them (in this case we used sine and cosine encoding) to make sure the algorithms capture that cyclical nature.

3.3.3. Granger Causality Test and Lagged Features

We tested Granger Causality Tests to determine whether certain macro-economic indices features can help predict sales. Then, we used the result of Granger Causality Tests to decide the best time lag for each feature for time series models like ARIMA.

3.3.4. Data Anomalies and Treatment

The data type of some features was changed from object to float.

3.3.5. Decomposition by Product Categories

We have decomposed the time series data of each product category into three components:

1. Trend: The overall direction of the data (upward, downward, or stable)
2. Seasonality: Repeating patterns at regular intervals
3. Residuals: The remaining part after removing trend and seasonality (random noise)

This helps us analyze patterns in the data and make better predictions.

3.4. MODELING AND MODEL SELECTION

3.4.1. Splitting Dataset

For the splitting of time series data, it is very important to avoid all possible data leakage problems; a model that uses data from the future will not be able to make predictions when deployed. To address this potential problem and guarantee data integrity, we are using the training data to predict a few months and then use this new data to predict farther ahead; repeating this process multiple times to maximize the likelihood of having a better prediction. Figure 1 shows a visual representation of the process.

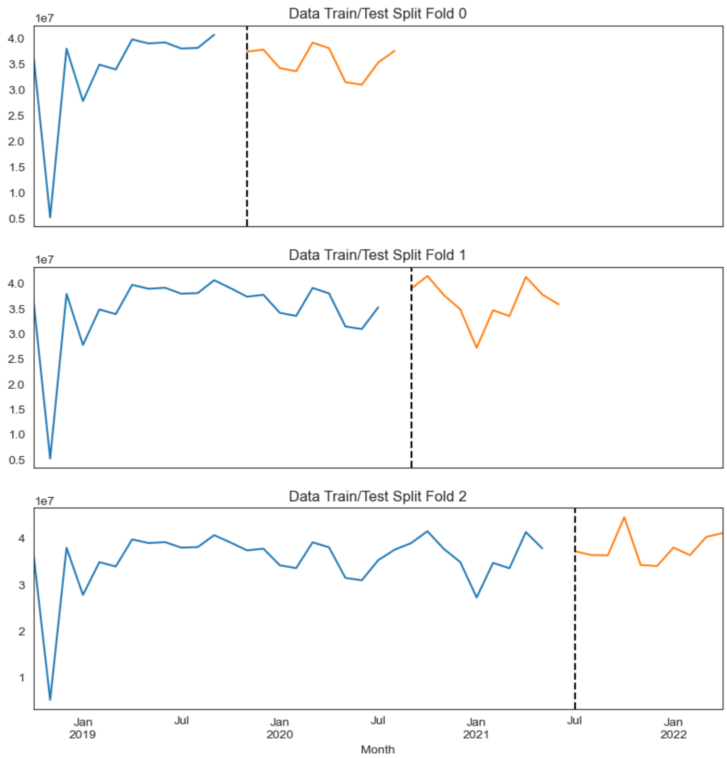

Figure 1: Splitting the Dataset

3.4.2. Important Features for Each Product Category

For each product category different features were selected for training the model. As every product group as individual features, possibly a different sales trend and most importantly also different customer segments possibly, it is important to select important features individually for each product group as different macroeconomic factors or different seasonality to different product groups apply. Some techniques for Feature Selection were used to access the most feature importance, like LASSO using regularizations techniques. However, their results were unreliable. During the testing phase, features penalized by LASSO were found important and the other way around. Therefore, other features were tested to get the highest results. A table with all the important features for training for each product category can be found in the appendix.

3.4.3. Testing different models

To have a wider choice of algorithms, in an effort to have the best results possible, we tried multiple algorithms: XGBoost, Linear Regression, AdaBoost, ARIMA and HistGradientBoosting. However, the latter three kept being outperformed by the first two. To keep the notebook and the report cleaner, we chose to just keep XGBoost and Linear Regression.

Table 1: Comparing Test Scores

Product Group	Scores (RMSPE)	
	XGBoost	Linear Regression
Product Category 1	**8.73%**	12.41%
Product Category 3	**26.80%**	31.61%

Product Category 4	**27.78%**	104.09%
Product Category 5	**51.69%**	119.90%
Product Category 6	**370.95%**	380.76%
Product Category 8	**42.71%**	660.55%
Product Category 9	**172.95%**	408.56%
Product Category 11	**499.97%**	810.30%
Product Category 12	**430.04%**	260.22%
Product Category 13	**61.24%**	2623.75%
Product Category 14	**60.71%**	244.00%
Product Category 16	**101.18%**	227.86%
Product Category 20	**41.41%**	403.57%
Product Category 36	**75.02%**	1903.87%

Disclaimer: despite the smaller percentage for product category 12 being smaller for linear regression, we used XGBoost for the actual prediction. We tried linear regression, but the results had 10 times more variance when compared with the historical data.

4. RESULTS AND EVALUATION

4.1. RESULTS AND EVALUATION

Our modeling strategy, and consequently its assessment, involved dividing the original sales data frame, already merged with market data, into 14 separate data frames, each corresponding to a specific sales group. This division was based on the distinct trends and overall behavior of each group. We recognized that different sales groups would not be affected in the same way by the same indicators, making it necessary to analyze and model them separately.

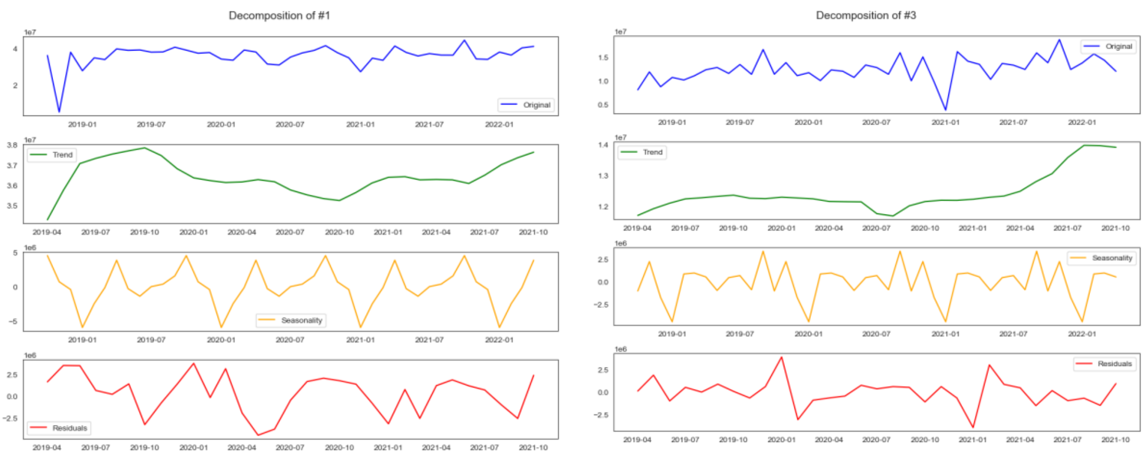

Figure 2: Decomposition Graphs

Overall, we observed a consistent pattern where the XGBoost model outperformed all other models across all sales groups so, we chose to forecast the sales groups exclusively using XGBoost.

The metrics used to assess the model were Root Mean Squared Error (RMSE) and Root Mean Squared Percentage Error (RMSPE), for a more intuitive display of the results. The decision to use these metrics was straightforward, as they were specifically requested by the stakeholder, SIEMENS. However, how we applied the formula needs to be handled carefully, as a small detail significantly impacted the results, and different sources gave different formulas. In the end, we computed the formula in a way that remained as true as possible to its original mathematical definition.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2}$$

Equation 1: Root Mean Square Error (RSME)

As seen in Figure 3, we observe a wide range of RMSPE values. We believe this variation occurs because some groups have higher sales volumes and, consequently, fewer rows with zero sales. This allows for a better understanding of the past, present, and future behavior of those groups. Although we considered using synthetic data, we ultimately decided against it to avoid introducing additional bias into our results. This idea is further supported by our findings, as the groups with the best scores #1, #3, and #4, have also the highest sales volumes.

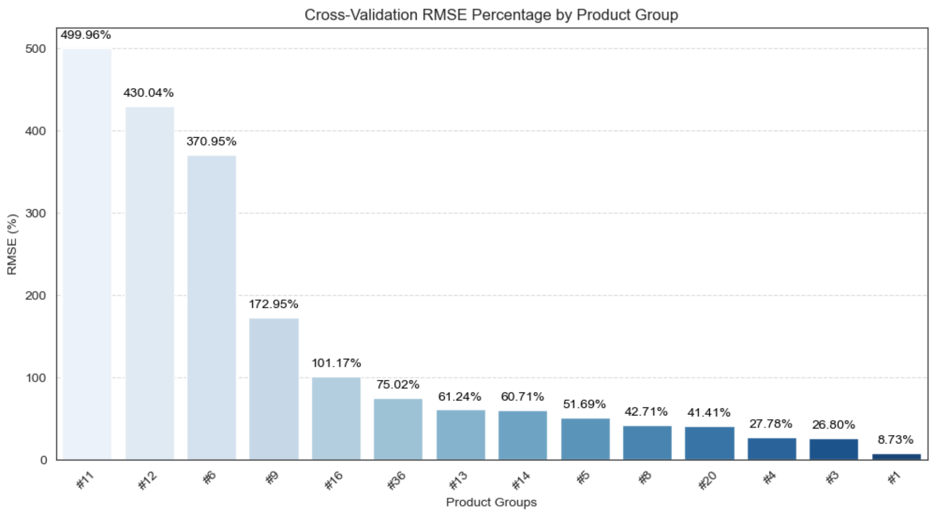

Figure 3: Cross-Validation RSME Percentage by Product Group

4.2. Interpretation of Sales Forecasts

Table 2 gives a short explanation of each sales forecast per product group:

Table 2: Interpretation of sales forecasts

Product	Description
#1	- Fluctuates between ~35M and ~40M EUR - No clear trend, relatively stable with slight variations = consistent
#3	- Ranges between ~10.3M and ~12.6M EUR - Slight fluctuations, some potential seasonality (repeating values at different periods)
#4	- Ranges between ~247K and ~348K EUR

	- Appears volatile, no clear directional trend, some fluctuation suggests seasonality
#5	- Forecast varies between ~7.6M and ~12.3M EUR - Strong fluctuation, significant increases and decreases - Possible seasonality or external factors affecting demand
#6	- Mostly stable around ~341K EUR (one spike) - Static trend, minor variations
#9	- Sales range from ~2.3K to ~8.2K EUR - Likely a stable product with minor demand variations (fluctuation)
#11	- Broad range from ~987K to 3.45M EUR - Large fluctuations indicate possible seasonality or market shifts - Peaks at certain times suggesting periods of high demand
#12	- Ranges between ~111K and ~328K EUR - Some sharp drops and spikes indicate volatility / seasonality
#13	- Ranges from ~8.2K to ~23.6K EUR - Gradual increase over time (some fluctuations and seasonality)
#14	- Sales between ~4.5K and ~29.8K EUR - No clear trend, generally decreasing with fluctuations
#16	- Most stable around ~166K EUR - Static product demand with minor dips
#20	- Small values from ~1.8K to ~3.2K EUR - Consistent but low demand, no major spikes or trends
#36	- Ranging from ~14K to ~131K EUR - Highly volatile, sudden spikes (seasonal event or anomaly)

As a summary, the following table shows key observations probably relevant for the sales team and the according product groups:

Table 3: Summary of Key Observations

Key Observations	**Product groups**
(Mostly) Stable Products: consistent forecasts, only slight fluctuations	#1, #3, #6, #9, #16, #20
(Highly) Volatile Products: significant fluctuations	(#4), #8, #11, #12, #36
Potential Seasonality: large periodic changes, cyclical demand	#5, #8, #11, #12, #36
Declining Trends	#14
Inclining Trends	#13

5. DEPLOYMENT - DEVELOPMENT AND MAINTENANCE PLANS

5.1. Digital Transformation and AI Forecasting Importance

The integration of AI-driven sales forecasting is a crucial component of digital transformation, particularly for industries dealing with complex supply chains and dynamic market demands. Traditional forecasting methods, which often rely on manual data entry and heuristic-based predictions, are prone to errors and inefficiencies. AI-based forecasting enhances decision-making by automating data analysis, reducing human bias, and improving predictive accuracy.

We are currently face following three challenges:

1. **Manual Forecasting Inefficiencies**: Conventional methods require extensive human effort and are time-consuming.
2. **Data Fragmentation**: Businesses must consolidate data from multiple sources, leading to potential inconsistencies.
3. **Lack of Real-time Adaptability**: Manual sales forecasting struggles to adapt quickly to market fluctuations and unexpected changes.

The following are the benefits of implementing AI-driven sales forecasting:

1. **Improved Accuracy**: AI models leverage large datasets and identify hidden patterns in sales trends.
2. **Automation**: Reduces reliance on manual forecasting, allowing managers to focus on strategic initiatives.
3. **Better Resource Allocation**: Enhances decision-making in inventory management and workforce planning.

5.2. SUCCESSFUL CASES

Here are three examples of successful digital transformation through AI prediction:

Case Study 1: PepsiCo's AI Transformation

According to Asia Growth Partners (n.d.), PepsiCo leveraged AI-based forecasting to enhance operational efficiency. Analysts reduced reporting time by 70%, built dashboards 90% faster, and improved overall data quality. The adoption of AI minimized forecasting errors and optimized decision-making in supply chain management.

Case Study 2: Sporting Goods Retailer

As claimed by Flevy (n.d.), a sporting goods retailer integrated AI-based forecasting to improve sales and operations planning. The initiative led to a 25% increase in online sales and an 18% reduction in operational costs, showcasing the value of AI in enhancing digital transformation efforts.

Case Study 3: Global Multinational Retailer

As stated by Flevy (n.d.), by implementing advanced analytics and machine learning, the company improved demand forecasting accuracy by 35%, minimized stockouts and overstocks, and enhanced customer satisfaction and profitability, highlighting the impact of technology on operational efficiency.

5.3. CHALLENGES AND SOLUTIONS

However, implementing AI-driven sales forecasting comes with its challenges. Below are three key obstacles along with potential solutions to address them effectively.

5.3.1. Resistance to Change

One of the biggest obstacles in digital transformation is organizational resistance to change. Employees and leadership may be hesitant to adopt AI-driven sales forecasting due to concerns about job security, lack of technical expertise, or comfort with traditional methods. A shift in mindset is essential to fostering innovation and continuous learning. To address this issue, enhance hands-on training and workshops to help employees gain confidence in AI tools. In addition, AI-based sales forecasting will be introduced gradually so that teams can adjust without sudden disruptions.

5.3.2. Model Accuracy and Trust

Managers may be skeptical of AI-generated forecasts, particularly when they differ from manual estimates. A lack of trust can lead to underutilization of AI insights, limiting the potential benefits. To solve this problem, benchmarks can be created using historical data and AI forecasts can be compared to historical trends to verify the accuracy of manual forecasts. As mentioned earlier, AI-driven sales forecasting may also be implemented in parallel with traditional methods before full-scale implementation to increase reliability. Furthermore, the model will be maintained every three months to check the quality of the results by comparing the traditional sales forecasts with the AI-based sales forecasts.

5.3.3. External Influences

Unstructured factors such as regional holidays and economic shifts can significantly impact sales forecasts. While AI can incorporate these variables, data availability and quality remain critical challenges. To resolve this issue, external APIs can be integrated to enhance data collection and improve the accuracy of forecasts for real-time holidays and economic indicators.

5.4. Future Plan

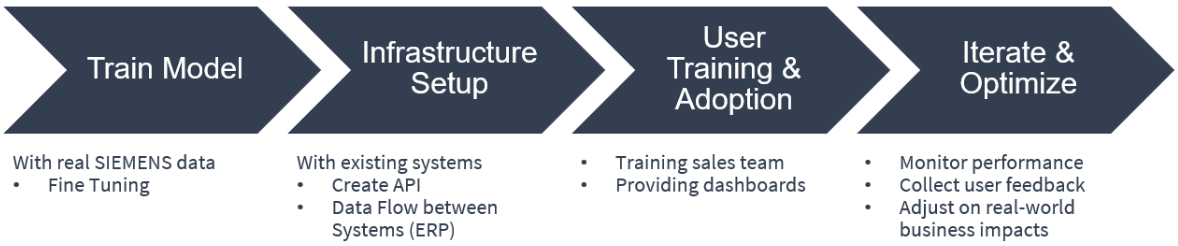

Figure 4: Future – Next steps

To successfully implement AI-driven sales forecasting, the following steps will be taken:

5.4.1. Model Training with Real SIEMENS Data

As a first step, the created model needs to be trained with real SIEMENS data. Some fine tuning according to the data will be performed to optimize model parameters and enhance predictive performance.

5.4.2. Infrastructure Setup and System Integration

Seamless integration with existing enterprise systems is essential for effective AI deployment. An API will be developed to facilitate smooth data exchange between the AI model and internal systems, such as the Enterprise Resource Planning (ERP) system. This ensures that sales forecasts are automatically updated and accessible in real time.

5.4.3. User Training and Adoption

To maximize the effectiveness of AI-driven sales forecasting, the sales team will undergo comprehensive training on how to interpret and leverage AI-generated insights. A user-friendly dashboard will be provided, enabling sales personnel to visualize forecasts, analyze trends, and make data-driven decisions efficiently.

5.4.4. Deployment with continuous Iteration and Optimization

AI performance will be continuously monitored to ensure accuracy and reliability. User feedback will be collected to assess the practical impact of the forecasts on business operations. Based on these insights, the model will be refined and adjusted to align with real-world business needs, ensuring long-term value and effectiveness.

6. CONCLUSION

6.1. BUSINESS IMPLICATIONS

By leveraging the AI-driven forecasting model, SIEMENS can unlock significant business advantages that enhance efficiency, optimize resources, and drive profitability. The following key benefits are:

- **Higher Accuracy**: Sales forecasting improves demand calculations, optimizes inventory levels and streamlines the supply chain to minimize waste and reduce costs
- **Time Efficiency**: Analysis of the forecast will make strategic decision-making easier by automating complex data analysis and forecasting processes which usually is resource intensive due to different data sources
- **Enhances Financial Planning & Budgeting**: Better cost control and financial predictability through more precise revenue and expenditure forecasts, less opportunity costs
- **Adaptability to Market Changes:** Quickly adjust business strategies to align with shifting market conditions, ensuring resilience and competitiveness

With AI-driven sales forecasting, businesses like SIEMENS no longer react to market trends - they anticipate them. This proactive approach will lead to a stronger competitive edge.

6.2. CONSIDERATIONS FOR MODEL IMPROVEMENT

To further improve the accuracy and reliability of the AI-driven sales forecast, continuous model evaluation and refinement in the process are essential. Several key areas for optimization have been identified to strengthen performance and ensure robustness.

Currently, sales forecasting is conducted by multiple models tailored to different product groups. By expanding the model testing with additional forecasting models, such as Prophet, their effectiveness and potential over existing approach can be assessed improved. Additionally, other Feature Selection methods and techniques should be implemented, as the ones used exhibit limitations. This shows as some features gave the model higher scores although not chosen as important models by the Feature Selection.

Furthermore, the quality of the data needs to be improved. Some product groups suffer from low-quality scores due to insufficient historical data. As they have low sales and some zero sales months a bigger timeline of data is needed for an accurate prediction. Sampling techniques in these cases would introduce noise and bias and therefore is no valid solution. The model's performance and reliability can not only be improved by expanding the data collection, but also with the supplement of further external data sets. Besides the macroeconomic indices, more consumer behavior data, competitor pricing and promotions as well as seasonal and environmental factors could be included to enhance predictive power.

Overall, a more comprehensive model evaluation pipeline should be developed to systematically test different models across each product group. This pipeline should analyze every combination of forecasting models as well as the given features to identify the best performing configuration for each scenario. Due to the time limit of this project and the expertise of the students the dataset was split according to the product group and individually handled. Ideally, this will not be necessary with an overall pipeline.

7. REFERENCES

Bridges, M. (2024, November 16). 30 Case Studies on Effective Sales & Operations Planning (S&OP). *Medium*. https://mark-bridges.medium.com/30-case-studies-on-effective-sales-operations-planning-s-op-50527b4a0dfe

Case Study Sales forecasting. (n.d.). *Siemens Advanta*. https://www.siemens-advanta.com/cases/sales-forecasting

Case Study - Pepsico's Transformation to Smarter Sales Forecasting with Designer Cloud (n.d.). *Asia Growth Partners.* https://asiagrowthpartners.com/case-study/pepsico-s-transformation-to-smarter-sales-forecasting-with-designer-cloud/c3054

Groene, N., & Zakharov, S. (2024). Introduction of AI-based sales forecasting: how to drive digital transformation in food and beverage outlets. *Discover Artificial Intelligence*, 4(1). https://doi.org/10.1007/s44163-023-00097-x

Joseph Robinson (n.d.) Flevy Management Insights Case Study Sales & Operations Planning Improvement for a Global Retailer. *Flevy.* https://flevy.com/topic/sales-and-operations-planning/case-sales-operations-planning-improvement-global-retailer

Joseph Robinson (n.d.). Flevy Management Insights Case Study Integrated Sales & Operations Strategy for Sporting Goods Retailer in North America. Flevy. https://flevy.com/topic/sales-and-operations/case-integrated-sales-operations-strategy-sporting-goods-retailer-north-america

10 Digital transformation Challenges to overcome. (n.d.). *UNIVIO*. https://www.univio.com/blog/10-critical-digital-transformation-challenges-to-overcome/

8. APPENDIX

Important Features Used for Model Training per Product Group

Product Category 1	'trend_12m'; 'seasonal_rolling_std_3m'; 'seasonal_rolling_std_6m'
Product Category 3	'United Kingdom_PrIn_lag_15'
Product Category 4	'month_cos', 'China.1_ShIn_lag_13', 'France.1_ShIn_lag_15', 'Germany_PrIn_lag_15', 'Germany.1_ShIn_lag_15', 'United Kingdom.1_ShIn_lag_15', 'United States.1_ShIn_lag_13', 'Europe_PrIn_lag_15', 'Europe.1_ShIn_lag_15', 'Producer Prices.3_FrEl_lag_13', 'production index.9_WoEl_lag_15', 'production index.11_UnKi_lag_15', 'production index.13_JaEl_lag_15', 'production index.15_GeEl_lag_15', 'quarter', 'rolling_mean_10m', 'seasonal_rolling_mean_6m'
Product Category 5	'seasonal_rolling_mean_6m', 'seasonal_rolling_std_6m', 'Germany_PrIn_lag_15', 'Japan_PrIn_lag_15', 'United Kingdom.1_ShIn_lag_15', 'United States.1_ShIn_lag_13', 'Europe_PrIn_lag_15', 'World: Price of Base Metals_lag_13', 'World: Price of Copper_lag_13', 'Producer Prices_UnSt_lag_13', 'Producer Prices.2_ItEl_lag_13', 'Producer Prices.4_GeEl_lag_13', 'production index.1_WoMa_lag_15', 'production index.3_UnKi_lag_15'
Product Category 6	'rolling_mean_10m', 'rolling_std_10m', 'seasonal_rolling_mean_3m'
Product Category 8	'seasonal_rolling_mean_6m'
Product Category 9	'Germany.1_ShIn_lag_15', 'China_PrIn_lag_13'
Product Category 11	'Germany_PrIn_lag_15', 'China_PrIn_lag_13', 'seasonal_rolling_mean_6m', 'month_sin', 'production index.15_GeEl_lag_15'
Product Category 12	'China.1_ShIn_lag_13'
Product Category 13	'month','China_PrIn_lag_13', 'France.1_ShIn_lag_15','United Kingdom.1_ShIn_lag_15', 'Europe.1_ShIn_lag_15', 'Producer Prices.1_UnKi_lag_13','Producer Prices.4_GeEl_lag_13', 'production index.9_WoEl_lag_15', 'production index.10_SwEl_lag_13', 'production index.15_GeEl_lag_15', 'seasonal_rolling_mean_6m'
Product Category 14	'China.1_ShIn_lag_13', 'seasonal_rolling_std_12m', 'month', 'month_sin', 'month_cos','rolling_mean_10m'
Product Category 16	'seasonal_rolling_std_12m', 'rolling_mean_10m'
Product Category 20	'month_cos', 'China.1_ShIn_lag_13', 'France.1_ShIn_lag_15', 'Germany_PrIn_lag_15', 'Europe_PrIn_lag_15', 'World: Price of Copper_lag_13', 'Producer Prices_UnSt_lag_13', 'Producer Prices.1_UnKi_lag_13','production

	index.1_WoMa_lag_15', 'production index.3_UnKi_lag_15', 'rolling_std_10m','seasonal_rolling_std_6m','seasonal_rolling_mean_3m'
Product Category 36	'month', 'month_sin', 'month_cos','Germany.1_ShIn_lag_15','United Kingdom_PrIn_lag_15', 'United States.1_ShIn_lag_13','World: Price of Base Metals_lag_13', 'Producer Prices.1_UnKi_lag_13', 'Producer Prices.2_ItEl_lag_13','production index.11_UnKi_lag_15', 'production index.15_GeEl_lag_15','rolling_std_10m','seasonal_rolling_mean_3m'