Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 3: Monthly Sales Forecast

Ana Carolina Ottavi, number: 20220541
Carolina Bezerra, number: 20220392
Duarte Girão, number: 20220670
João Pólvora, number: 20221037
Luca Loureiro, number: 20221750

Group Q: OptimaDataConsulting

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa
May, 2023

INDEX

1. EXECUTIVE SUMMARY

This report discusses the development and implementation of a sales forecasting model for a Siemens business unit in Germany. The purpose of the model is to overcome the limitations of manual forecasting, which can be time-consuming and subject to the biases of multiple stakeholders. Through the use of data-driven strategies, the organization aims to minimize opportunity costs, optimize working capital, and enhance customer satisfaction.

The study focuses on selected product groups within Siemens' Smart Infrastructure Division in the German market, using sales data from October 2018 to April 2022, as well as relevant macroeconomic indices. Following the CRISP-DM methodology, this report details the process of understanding, preparing, modeling, and deploying the sales forecasting model.

In conclusion, this report emphasizes the importance of leveraging data to make informed decisions and the value of implementing a successful sales forecasting model within an organization. By leveraging different techniques to examine historical sales and market data, data-driven sales forecasting seeks to increase the accuracy of sales predictions.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. Background

Siemens has a subsidiary called "Smart Infrastructure," and its primary mission is to collaborate with its clients in the development of ecosystems that bridge the gap between the digital and physical worlds. The clients are able to make decisions through the use of data and analytics based on the goods and solutions provided by smart infrastructure. These decisions allow the clients to make their energy systems, buildings, industries, or particular ecosystems more efficient and sustainable[1].

Siemens Advanta is a Siemens AG company with a focus on IoT. A project was proposed within the scope of Business Cases for Data Science in which our group was asked to generate a monthly sales estimate for the Smart Infrastructure Division in Germany described above. The sales data covers the time period from October 2018 to April 2022. Finally, a test period ranging from May 2022 to February 2023 will be utilized to quantitatively evaluate the score and confirm the model's accuracy.

2.2. Business Objectives

The main business objectives of the project are:

- Develop an accurate and reliable sales forecasting model for each Mapped_GCK using the available historical sales data and market indexes.
- Identify the main sales drivers for each Mapped_GCK, such as producer prices, raw material prices, shipment indexes, and other pertinent market indexes.
- Employ the forecasting model and market insights to generate actionable insights for the sales team.
- Determine patterns and trends in the sales data to obtain insight into customer preferences and behavior.
- Create a sales forecasting model with a low RMSE score for accurate and reliable predictions, hence enhancing decision-making processes and lowering expenses associated with wrong forecasts.

2.3. Business Success criteria

The success of the project will be determined by:

- Desirably, achieve a root mean square error (RMSE) of less than 5% for the sales forecast model.
- Create a reliable and scalable forecasting model that can be used to forecast sales across multiple product categories.
- Determine the important market indicators that have the biggest impact on sales and utilize them to increase the accuracy of the sales forecast model.
- The reduction in resource utilization and bias in the forecasting process.
- The positive impact of the model on working capital and customer satisfaction.

2.4. Situation assessment

The team will design a model that will anticipate future sales for each Mapped_GCK using historical sales data, market data, and other pertinent economic indices. Advanced statistical approaches and machine learning algorithms will be used to create the model.

However, there are some potential risks associated with the project. The quality of the data is critical to the model's effectiveness, and there may be data discrepancies that affect the forecasting model's accuracy. To avoid these risks, the team will apply a rigorous data cleaning and validation approach, as well as cross-validate the model using test data.

The project's advantages are obvious. The company's sales forecasting process will be more precise and data-driven, allowing it to make more educated business decisions. Furthermore, the forecasting model will assist the company in identifying new business opportunities and optimizing pricing strategies, resulting in increased profitability

2.5. Determine Data Mining goals

The primary data mining goals of the project are:

- Investigate the data and identify variables relevant to the sales forecasting process.
- Examine patterns and relationships within the sales data and macro-economic indices.
- Develop an AI-driven sales forecasting model that integrates the identified variables and patterns.

3. METHODOLOGY

3.1. Data understanding

The primary objective of this section was to examine the datasets referring to the Siemens business unit in Germany, including sales data, test data, and market data. Sales data consists of 9,802 rows and three columns, where were identified the date, amount of sales in euros per day at each category. Test data comprises 140 rows and three columns, without target values. Market data has 221 rows and 48 columns, featuring macroeconomic indices. Most variables in the market dataset are related to machinery and electrical equipment production, shipments and producer prices indices for a few countries. To better understand the variables and their unique characteristics, the summary table for each dataset was created, providing preliminary insights for each of the features.

Sales Data:	
Features	Takeaways
DATE	9,802 entries, 1,216 unique values, most common date is 16.04.2021, appearing 14 times.

Mapped_GCK	9,802 entries, 14 unique values, most common value is #1, appearing
Sales_EUR	9,802 entries, 2,609 unique values, most common value is 0, appearing 7,134 times

Market Data:

Indicators	Takeaways
Duplicates	Columns MAB_ELE_PRO156 (China Production Index Machinery & Electricals) and MAB_ELE_SHP156 (China Shipments Index Machinery & Electricals) shown as duplicates.
Missing Values	MAB_ELE_SHP826 (United Kingdom), PRI27826_org (United Kingdom: Electrical equipment) have 18 missing values. PRI27250_org (France: Electrical equipment) has 35 missing values. PRI27156_org (China: Electrical equipment) has 23 missing values. PRO271000_org (World: Electrical equipment) has 11 missing values
Mean	The lowest mean at 34.213427 (MAB_ELE_PRO380 Italy's Production Index Machinery & Electricals) and the highest at 141.269730 (PRO27250_org France Electrical equipment). This indicates that the variables might be on different scales and could potentially benefit from normalization or standardization.
Standard Deviation	The standard deviation also varies across the variables, with the lowest value being 8.444573 (United States: Electrical equipment) and the highest value being 78.883209 (China's Production Index Machinery & Electricals and Shipments Index Machinery & Electricals). This suggests that the data points are spread differently across the variables.
Range	The minimum and maximum values for each variable vary significantly, further supporting the idea that these variables may be on different scales. The minimum values range from 16.940704 (China's Production Index Machinery & Electricals and Shipments Index Machinery & Electricals) to 83.310173 (United States' Production Index Machinery & Electricals), and the maximum values range from 121.495483 (Germany: Electrical equipment) to 329.413367 (MAB_ELE_PRO156 China's Production Index Machinery & Electricals and Shipments Index Machinery & Electricals).

3.2. Data Exploration

The goal of the data exploration phase of the CRISP-DM process is to comprehend and clean the dataset, visualize and summarize the data, and uncover relationships and correlations between the variables in order to get the data prepared to be used in modeling.

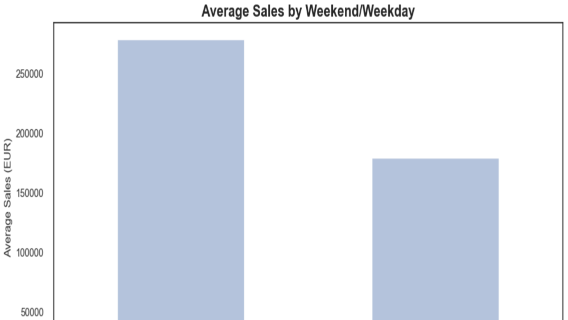


Figure 1 - "Bar plot for average sales by weekend/weekday ".

The graph ***(Figure 1)*** shows us the average sales' amount in week and weekend days. On the weekdays the average sale value is bigger than 250,000 and on the weekends the value is below 200,000. So that the week average sales is 25% percent bigger than the weekend one.

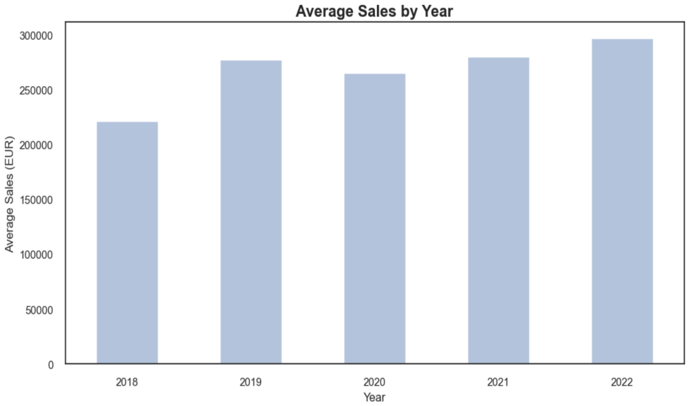


***Figure 2* -** "Bar plot for the average sales by year ".

The graph (***Figure 2)*** shows the average sales throughout the years presented. The biggest average sale happens in 2022 with a value of almost 300,000 euros and the smallest in 2018 with a value of 200,000 euros. There seems to be a positive trend on the average sales as well.

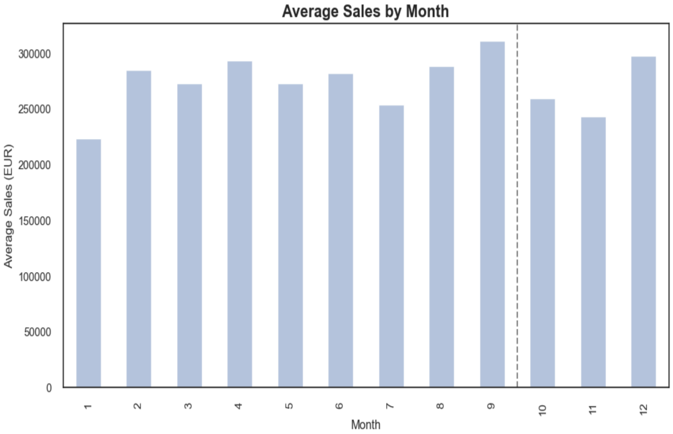
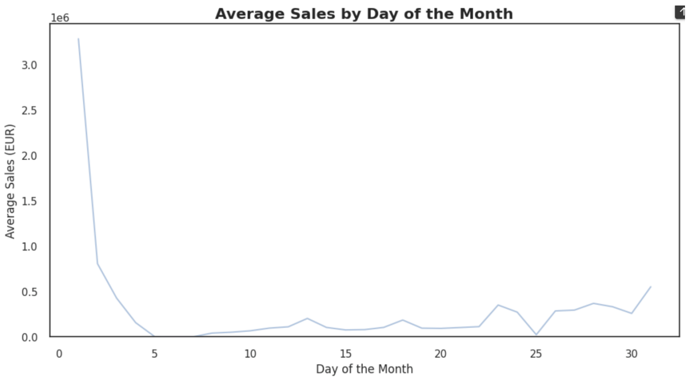


***Figure 3* -** "Bar plot for the average sales by month".

Figure 3 reveals the average monthly sales. September had the highest average sales revenue, while January had the lowest. Nonetheless, monthly average sales seem to be rather constant.

***Figure 4* -** "Line plot for the average sales by day".

The graph indicates that the average sales exhibit a peak during the initial days of the month, followed by a significant decrease, and then it remains relatively constant until the end of the month. Towards the end of the month, there is a slight increase in the sales value.



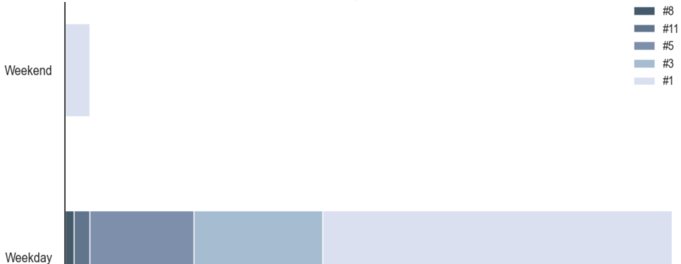


***Figure 5* -** "Stacked bar chart that displays the total sales by Mapped_GCK on weekdays and weekends".

The illustration displays sales for the top 5 Mapped_GCK categories in ascending order on weekdays and weekends. Sales during the weekend are mainly related to product category 1, while the categories with the highest sales values are #1, #3, and #5.

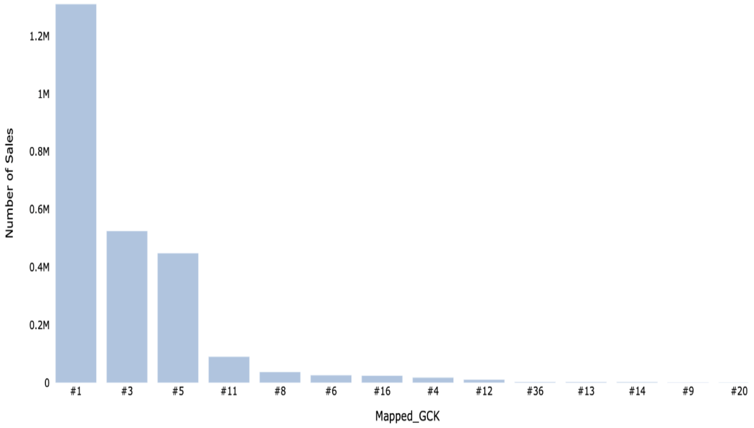
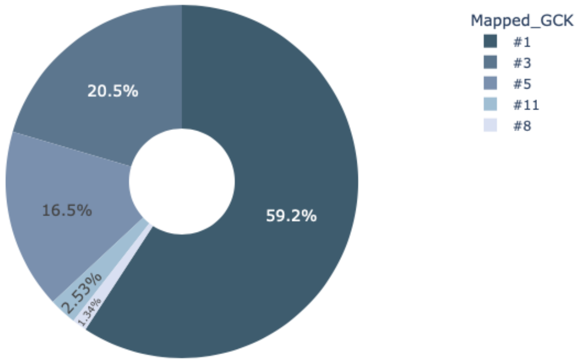


Figure 6 - "Bar chart for the average sales distribution per Mapped_GCK".

The chart enables a comparison of the performance of different categories. From the plot, it is evident that product category 1 has the highest number of sales, followed by the third and fifth categories.

Figure 7 - "Pie chart that displays the distribution of sales for the top 5 Mapped_GCK with the highest mean sales".

The chart reveals that product categories 1, 3, and 5 contribute to 94.8% of the total sales, making them the leading product categories in terms of overall sales performance.



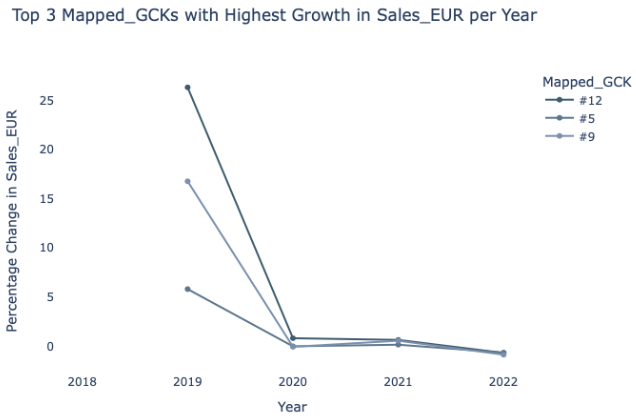
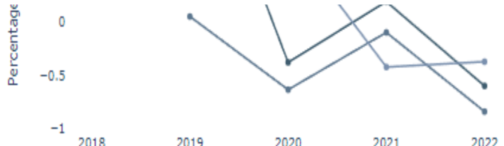


Figure 8 - "Top 3 Mapped_GCK with the biggest fluctuations in Sales_EUR per year"***.***

Categories 5, 9, and 12 experienced the highest sales fluctuations over the years, as shown in ***Figure 8.*** However, between 2019 and 2020, there was a perceptible decline in sales, which may be attributed to the COVID-19 pandemic, which caused significant disruptions in supply chains worldwide. The range of the y-axis is approximately 25%, indicating a significant decline in sales during this period.

Figure 9 - "Top 3 Mapped_GCK with the lowest fluctuations in Sales_EUR per year"***.***



Categories 13, 16, and 20 had the lowest fluctuations in sales per year, as shown in ***Figure 9***. The y-axis indicates that these Mapped_GCKs experienced the lowest percentage change in sales over the years, with a range of approximately 2%.

Time series analysis allows us to see how variables shift over time. The time series graphs developed help us identify trends, patterns, or seasonal fluctuations in sales data over time. According to the plot in Figure 10, the lowest sales occurred in November 2018 and January 2021, while the biggest sales occurred in December 2020. Figure 11 depicts a minor upward trend over time. As seen in ***Figure 12***, the largest daily sales peaks occurred in January 2022 and October 2020.

***Figure 10* -** "Line plot with the average sales per month throughout the time series"

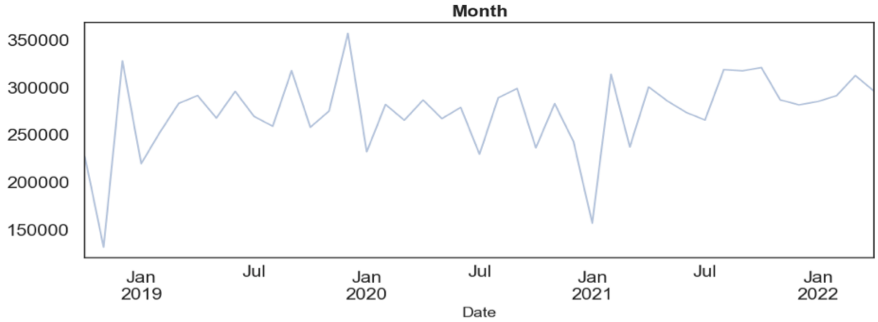


***Figure 11* -** "Line plot with the average sales per week throughout the time series"

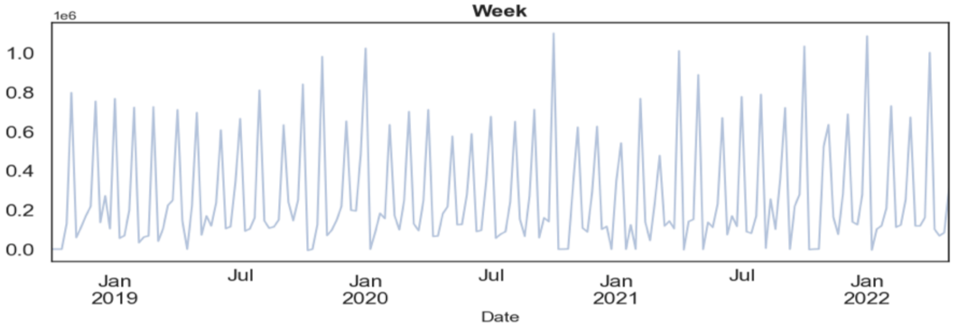
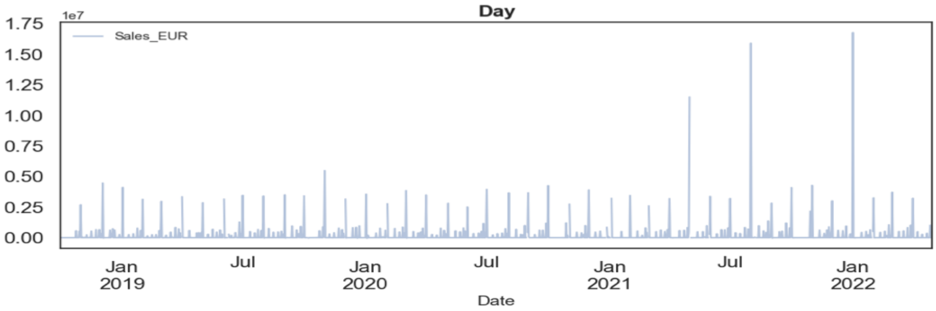


***Figure 12* -** "Line plot with the average sales per day throughout the time series"



The tendency plot depicts a dataset's central tendency and variability across time, revealing its overall trend. MAB_ELE_PRO392, MAB_ELE_SHP392, MAB_ELE_PRO756, and MAB_ELE_SHP756 have the same tendency line, showing that the manufacturing sector is meeting demand and shipping items on schedule. Since 2018, producer price indexes have risen while all raw materials fluctuate. Demand for commodities, supply chain disruptions, technical developments, geopolitical events, natural disasters, and other variables can make the production index unstable. November 2021, September 2019, and September 2020 had Production Index antimodes in multiple nations. The production index and other market indexes fluctuated due to 2018–2022 events in Europe, the US, China, and Japan. Siemens' Smart Infrastructure Division in Germany can use these market data patterns and their relationship with sales performance to construct a more accurate AI-driven sales forecasting model.

Later on, we investigated the relationship between holiday patterns and sales to understand how sales are affected during different holiday periods.

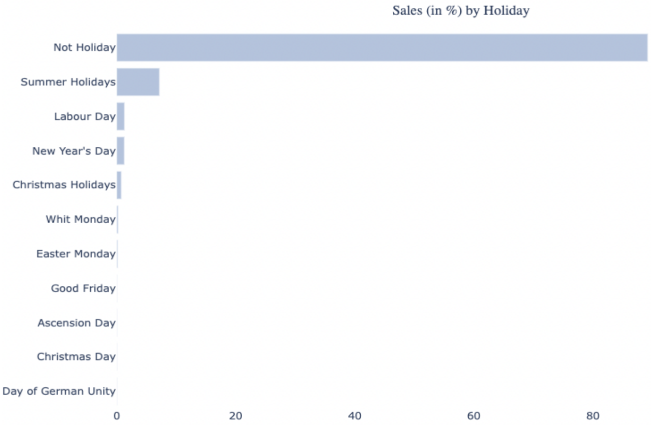


***Figure #* -** “The percentage of total sales for each holiday".

We compiled a list of significant holidays that we found relevant to explore their relationship with sales. The resulting dataframe, which combined holidays (or not) and sales, was sorted by the percentage of total sales. The bar plot in ***Figure #*** indicated that Labor Day and New Year's had higher sales compared with the average day, yet August had lower sales than the average month.

3.3. Data Pre-processing

Data pre-processing includes determining key features from market data and deleting unnecessary records in order to enhance data performance. Firstly, the 'Date' column in the market dataset was split into 'Year' and 'Month' columns. Both columns were then converted to integers to make manipulation easier. The "MAB_ELE_SHP156' column was removed to ensure data integrity, as it was a duplicate with the 'MAB_ELE_PRO156' column. Finally, the resulting data frame was sorted in ascending order based on 'Month' and 'Year'. Moreso, the 'Sales_EUR' column was then converted to a float data type. These steps allowed for a clean and organized analysis of market trends.

Regarding the 'Date' column in the market dataset, it was split into Year and Month, converted to integers, and then discarded. At this stage, we analyzed three approaches[1] for dealing with missing values: zero, mean, and KNN imputer in correlation (Spearman), lasso, and RFECV feature selection. KNN Imputer performs well with small datasets, enabling us to generate accurate estimates.

Based on the analysis of the market data, it can be concluded that the majority of histograms are not skewed and follow a normal distribution. Additionally, variables such as 'Roh CRUDE PETRO1000 org', 'Roh ENERGY1000 org', 'WKLWEUR840 org', 'PRI27826 org' and 'PRO27380 org' display multimodal tendencies, indicating the presence of multiple modes or peaks in the data.

We tested a few different techniques, including the Interquartile Range (IQR) method, the Z-Score method, and manual elimination, to identify and manage outliers. Only the 'Sales_EUR' feature used the IQR method to detect outliers, but they were not eliminated because they were considered significant for sales forecasting. Since the Z-Score method did not identify any outliers, the entire dataset was preserved.

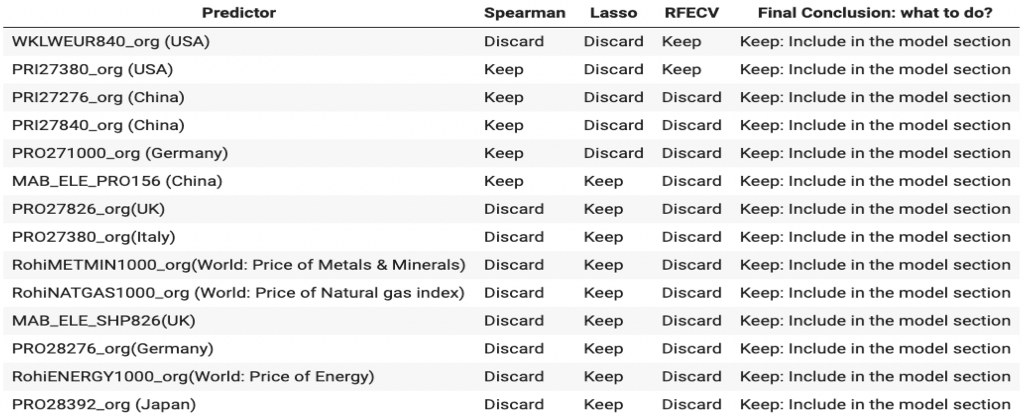
[1] REFERENCE Mandel J, S. P. (2015). A Comparison of Six Methods for Missing Data Imputation. Journal of Biometrics & Biostatistics, 06(01). https://doi.org/10.4172/2155-6180.1000224

The Spearman correlation was used on training data (target was merged). Our analysis led to several conclusions. First, no independent variable had a high correlation with the target (Sales_EUR), requiring other feature selection techniques for insights. Furthermore, the following variables had significant correlation between each other: PRI27276_org and PRI27840_org (0.643), PRI27380_org (0.618), MAB_ELE_PRO156 (0.595) and MAB_ELE_SHP156 (0.595), PRO271000_org (0.526), and PRI27826_org (0.518).

We decided to apply other feature selection techniques to determine which variables to keep. To identify the features most important for predicting sales, we filtered for correlations with 'Sales_EUR'. Sales correlations with an absolute value greater than 0.5 were considered significant: Year (0.644539), PRI27276_org (0.643627), PRI27840_org (0.643117), PRI27380_org (0.618803), MAB_ELE_PRO156 (0.595752), PRO271000_org (0.516377).

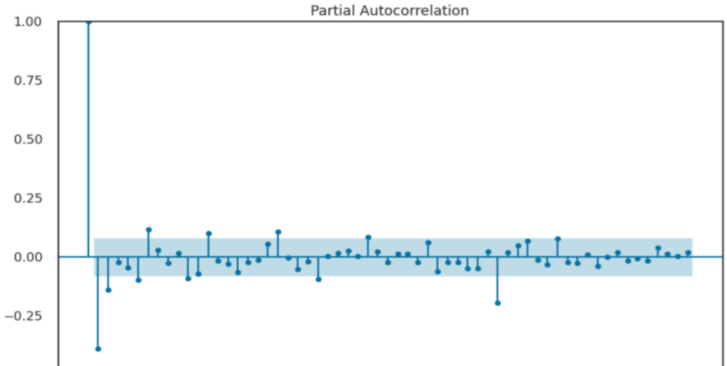
The following section discusses feature selection to identify the most significant features for the model's performance. We chose a subset of features in the market dataset using LASSO to improve model performance. Lasso distinguishes 9 variables (PRO27380_org, MAB_ELE_PRO156, PRO27826_org, RohiMETMIN1000_org, MAB_ELE_SHP826, RohiNATGAS1000_org, PRO28276_org, RohiENERGY1000_org, PRO28392_org). RFECV employed the same split as Lasso, providing a score for each feature after applying a Ridge regression model to identify the ideal number of features, which was 2: WKLWEUR840_org, PRI27380_org. Based on the KNN Imputer cleaning method, we decided to maintain all the features that appear in the correlation, LASSO and RFE. ***Figure 13*** gives an overview of feature selection using KNN Imputer to fill missing values.

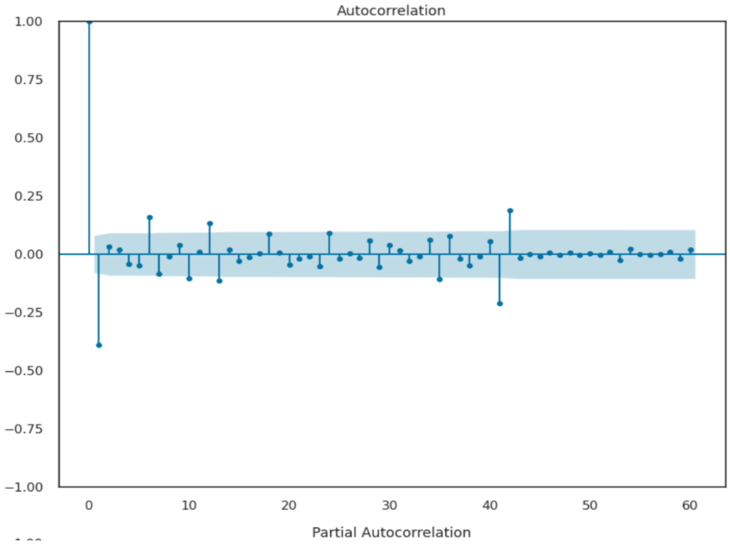
***Figure 13* -** "Overview KNN Imputer"



Predictor	Spearman	Lasso	RFECV	Final Conclusion: what to do?
WKLWEUR840_org (USA)	Discard	Discard	Keep	Keep: Include in the model section
PRI27380_org (USA)	Keep	Discard	Keep	Keep: Include in the model section
PRI27276_org (China)	Keep	Discard	Discard	Keep: Include in the model section
PRI27840_org (China)	Keep	Discard	Discard	Keep: Include in the model section
PRO271000_org (Germany)	Keep	Discard	Discard	Keep: Include in the model section
MAB_ELE_PRO156 (China)	Keep	Keep	Discard	Keep: Include in the model section
PRO27826_org(UK)	Discard	Keep	Discard	Keep: Include in the model section
PRO27380_org(Italy)	Discard	Keep	Discard	Keep: Include in the model section
RohiMETMIN1000_org(World: Price of Metals & Minerals)	Discard	Keep	Discard	Keep: Include in the model section
RohiNATGAS1000_org (World: Price of Natural gas index)	Discard	Keep	Discard	Keep: Include in the model section
MAB_ELE_SHP826(UK)	Discard	Keep	Discard	Keep: Include in the model section
PRO28276_org(Germany)	Discard	Keep	Discard	Keep: Include in the model section
RohiENERGY1000_org(World: Price of Energy)	Discard	Keep	Discard	Keep: Include in the model section
PRO28392_org (Japan)	Discard	Keep	Discard	Keep: Include in the model section

The objective of the analysis was to determine the optimal lag month for predicting time series with an ARIMA model. Creating a range of lagged versions of the dataset, fitting an ARIMA model to the training data, and making predictions on the validation data for each latency month in the range of 1 to 30 were the steps involved in this method. The month with the smallest root mean squared error (RMSE) was discovered to be 23. The autocorrelation function (ACF) and partial autocorrelation function (PACF) were then plotted (Figures # and #) to determine whether or not there was a correlation between observations at various time delays. The analysis revealed that the time series is either stationary or a white-noise series, in which values are random and unpredictable and there is no underlying pattern or trend.





The ARIMA model was used in this part to forecast monthly sales. The proper ARIMA model parameters were identified by plotting the ACF, PACF, and decomposition of the time series data with functions. To forecast sales for the following six weeks and illustrate the prediction graph, a predict_sales() method was constructed. The dataset was separated into two parts: training and validation, with 70% and 30% allocated to each. The model summary and diagnostics showed potential difficulties with unstable standard errors while applying the ARIMA model. A lagged version of the dataset was constructed and scaled using MinMaxScaler to improve the model. On the training data, the ARIMA model was fitted with the order (2, 0, 0), and predictions were made on the validation data. The RMSE was roughly 6101757.90, indicating that more research and testing with alternative model parameters were required to attain better results.

4. MODELING

Time series forecasting and machine learning with RNN algorithms are two modeling approaches for sales forecasting. Moving averages, exponential smoothing, and ARIMA models are common time series forecasting strategies. For sales forecasting, machine learning methods such as decision trees, random forests, and neural networks are used. Our group developed ARIMA, Prophet, and ensemble machine learning models for this project.

When we associate a temporal or time component to the forecast, it becomes Time Series Forecasting and the data is called Time Series Data. In statistical terms, time series forecasting is the process of analyzing the time series data using statistics and modeling to make predictions and informed strategic decisions

Thinking about a time series is like understanding three components: trend, seasonal and residual. The decomposition calculates the components individually. Ideally, trend and seasonality should capture most of the time series patterns.

Therefore, the residuals represent what is left of the time series, after the trend and seasonality have been removed from the original signal[2]. So, we are going to look at our time series in this perspective.

Figure 15

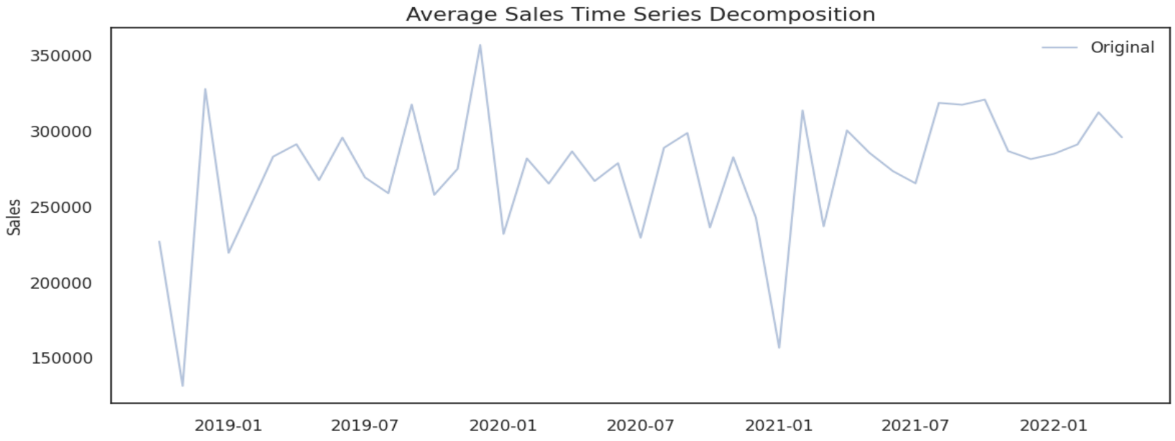


Figure 16

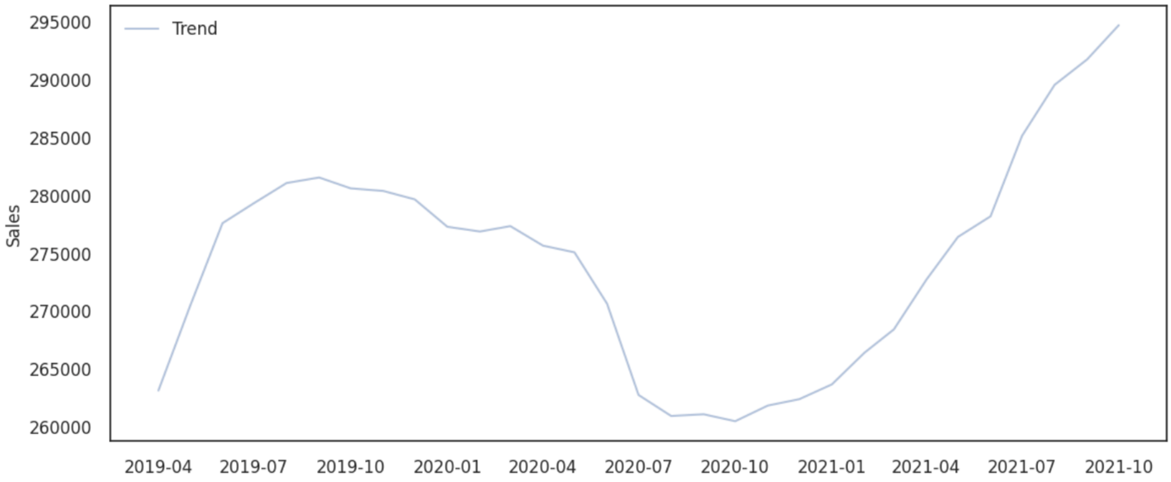


Figure 17

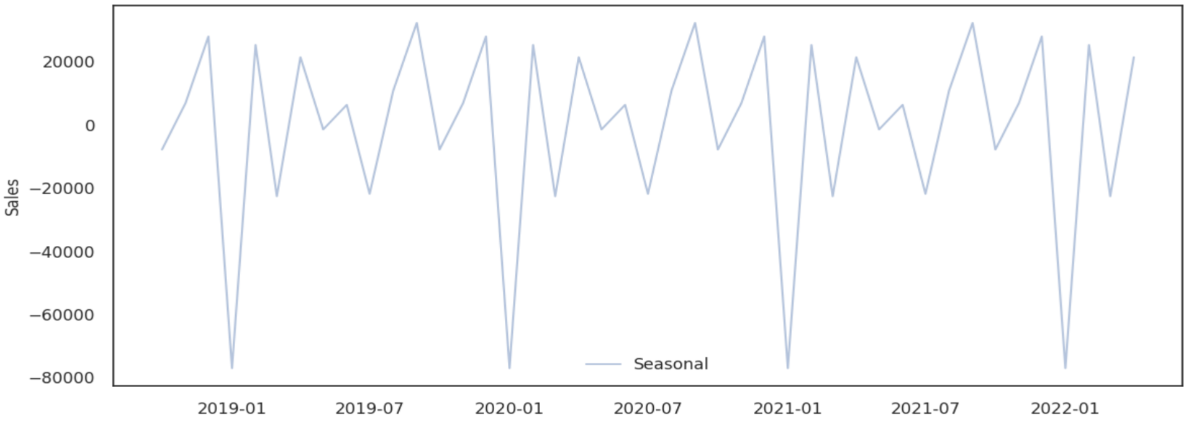


Figure 18

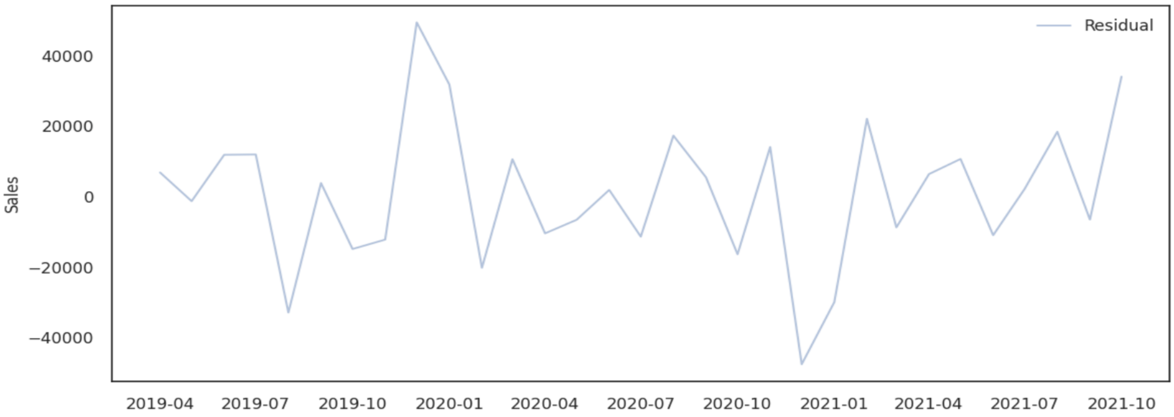


Figure 19

Decomposition shows the time series' long-term trend. In our decomposition (Figure 16), it decreases in the second and third quarter of 2020 before increasing again in October. Seasonality is data that repeats at regular intervals. Weather, holidays, and economic cycles can cause it. Our breakdown (Figure 17) shows a large decline over Christmas and likely school breaks. Residual (Figure 17) highlights external influences like economic changes or customer behavior that are not accounted for by trend and seasonal components.

In order to apply the ARIMA model, we need to validate the stationarity of the time series; once the time series is not stationary the properties of the estimators obtained from the data are unreliable. We used two tests. The ADF test assumes the null hypothesis that the time series has a unit root and the alternative hypothesis that it is stationary. Yet, the KPSS test assumes the null hypothesis that the time series is stationary and the alternative hypothesis that it has a unit root. The ADF test is better for time series predicted to be stationary around a linear trend, whereas the KPSS test is better for constant mean. The ADF test shows the time-series is stationary (p-value = 3.94 > 0.05), while the KPSS test shows it is not (p-value = 0.07 > 0.05). In time series that include trends , the KPSS test is more effective than the ADF test for stationarity testing. We then assumed stationarity of the time series based on ADF.

In the next step, we extracted the time series trend by applying Holt-Winters' exponential smoothing to the average sales data. We specified additive trend and 0.2 smoothing parameter. The findings show the original sales data (blue) and the fitted values (green) that estimate the trend. The fitted trend line captures the general sales increase over time. Until January 2020 and January 2021, seasonality was upward.

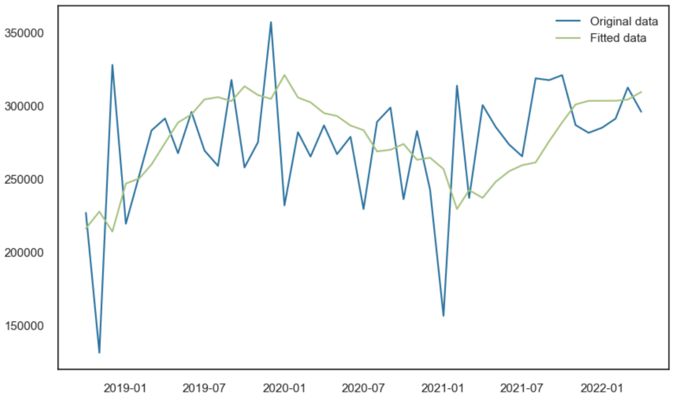
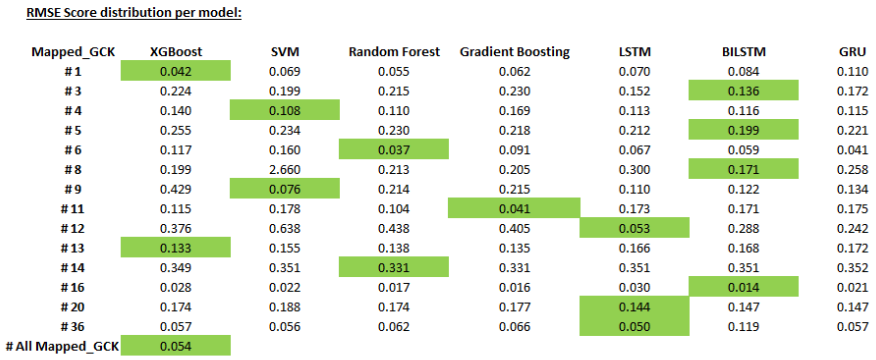


Figure 20

After forecasting using a more traditional method as ARIMA and a more recent and simpler one, as Prophet, we got to the conclusion that neither one of them got us satisfactory results. In this sense, we aimed to apply more sophisticated methods such as Machine Learning models (Random Forest, Gradient Boosting, XGBoost (which is an extreme application of the previous one), and SVM); and Recurrent Neural Networks, known for their powerful capacity to capture long-term dependencies and to forecast accurately, when dealing with large and complicated datasets, such as LSTM, BILSTM and GRU.

In our first approach, we tested each of the seven models above defined only on sales correspondent to Mapped_GCK #1, to obtain some sensibility of each model performance and to understand what were the best parameters for our dataset, through a Grid Search hyperparameter with an 10 k-folds on cross validation; then, we generalized our rationale for the remaining thirteen product categories, testing each model, with its best parameters previously discovered. After this process, we compared every model results, per product category, and assessed which one delivered us a lower RMSE. Based on this search, we built a function that aggregated every best model of the fourteen existent Mapped_GCK (***Figure 21***).

As we can see in the figure below, there is not one specific model that outperformed all the remaining ones. It really depends on the characteristics and on the patterns of each specific product category sales from 2018 to 2022. As we can see below, we obtained a final RMSE for the entire training dataset of 0.054, which is remarkable.



RMSE Score distribution per model:

Mapped_GCK	XGBoost	SVM	Random Forest	Gradient Boosting	LSTM	BILSTM	GRU
# 1	0.042	0.069	0.055	0.062	0.070	0.084	0.110
# 3	0.224	0.199	0.215	0.230	0.152	0.136	0.172
# 4	0.140	0.108	0.110	0.169	0.113	0.116	0.115
# 5	0.255	0.234	0.230	0.218	0.212	0.199	0.221
# 6	0.117	0.160	0.037	0.091	0.067	0.059	0.041
# 8	0.199	2.660	0.213	0.205	0.300	0.171	0.258
# 9	0.429	0.076	0.214	0.215	0.110	0.122	0.134
# 11	0.115	0.178	0.104	0.041	0.173	0.171	0.175
# 12	0.376	0.638	0.438	0.405	0.053	0.288	0.242
# 13	0.133	0.155	0.138	0.135	0.166	0.168	0.172
# 14	0.349	0.351	0.331	0.331	0.351	0.351	0.352
# 16	0.028	0.022	0.017	0.016	0.030	0.014	0.021
# 20	0.174	0.188	0.174	0.177	0.144	0.147	0.147
# 36	0.057	0.056	0.062	0.066	0.050	0.119	0.057
# All Mapped_GCK	0.054						

Taking into consideration Figure 19, we noticed that the most robust RNN models did not outperform significantly the remaining ones, which performed best in 7 categories. We believe that the dataset we are working with is big enough (has less than 10.000 records, representing five months) to store significantly historical sales trends. Additionally, we also noticed that BILSTM was the model that achieved most interesting results, mainly on the product categories with higher total sales amounts, as #3 and #5. These results are in line with our theoretical expectations, since processes sequence in both forward and backward directions, naturally increasing prediction complexity, when compared with the remaining RNN models. From the simpler machine learning models, we highlight XGBoost, which obtained the best score for three of the product categories, including the one that represents the bigger total sales amount, the first one, with an amazing RMSE score of 0.042.

After evaluating our current models, based on our validation dataset, we also predicted the forecasts for the next ten months, from May 2022 to February 2023, as part of our requirements. In this sense, we also obtained very interesting results, noticing a clear positive trend (with any particular month exception), which is in line with the past behavior for the demand of sales that could be seen from trend and seasonal components of the decomposition from the monthly average sales of previous years. In absolute values, we predict an increase from May 2022 of +77M Eur to February of +511M Eur, representing a total increase of +435M Eur (***Figure 22***).

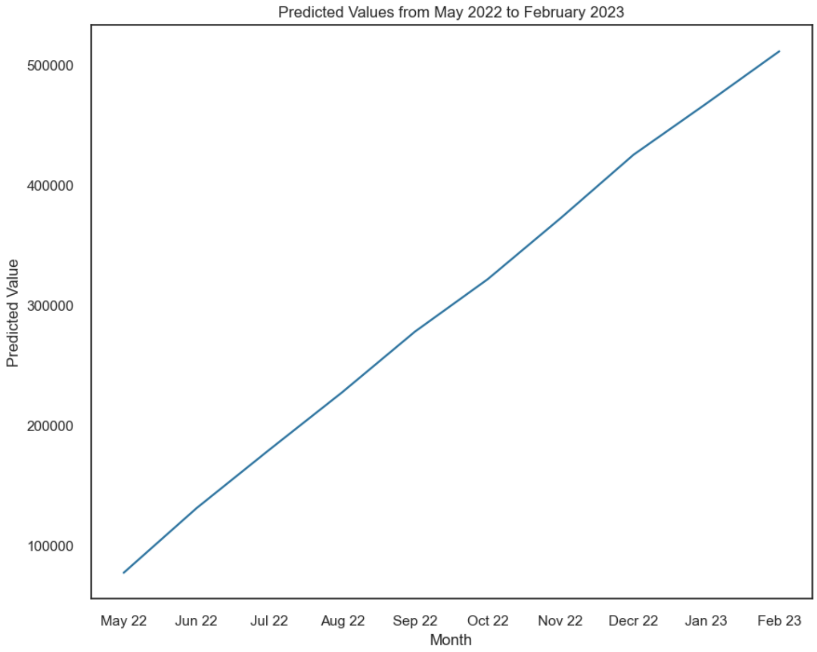


Figure 22

5. DEPLOYMENT AND MAINTENANCE PLANS

In a short term data analysts must implement an overview sales forecast dashboard addressing the 14 different product sales predictions. The dashboard would be a tool to be used in regular quarterly reviews for the decision makers such as financial and marketing directors. Not least, systematic documentation should be applied from technical to operational level to monitor forecast performance.

For the medium term, incorporate the prediction into sales and marketing strategies. Thus, the marketing team could promote the products through campaigns. In addition, the A/B testing should be applied in selected countries as a validation strategy to calculate the expectations based on real transactions.

Lastly, strong KPIs to monitor the model performance of sales. This allows Siemens to identify any discrepancies between the predicted and actual sales making possible adaptations in the forecasting methodology. The adjustments could be at an algorithms or statistical level to improve the accuracy or incorporating new data to the model.

6. CONCLUSIONS

[1] Smart infrastructure for a sustainable future. (n.d.). siemens.com Global Website.
https://www.siemens.com/global/en/company/topic-areas/smart-infrastructure.html

[2] https://medium.com/analytics-vidhya/time-series-forecasting-a-complete-guide-d963142da33f

https://github.com/SaraxSilva/Business-Cases-Projects-21-22/blob/main/BC4_crypto_forecasting/BC4%20Project.ipynb

https://github.com/jajokine/Business-Cases/blob/main/sales_forecasting.ipynb

https://github.com/jajokine/Business-Cases/blob/main/deep_learning_forecasting.ipynb

https://github.com/jajokine/Business-Cases/blob/main/advanced_deep_learning_forecasting.ipynb

8. APPENDIX

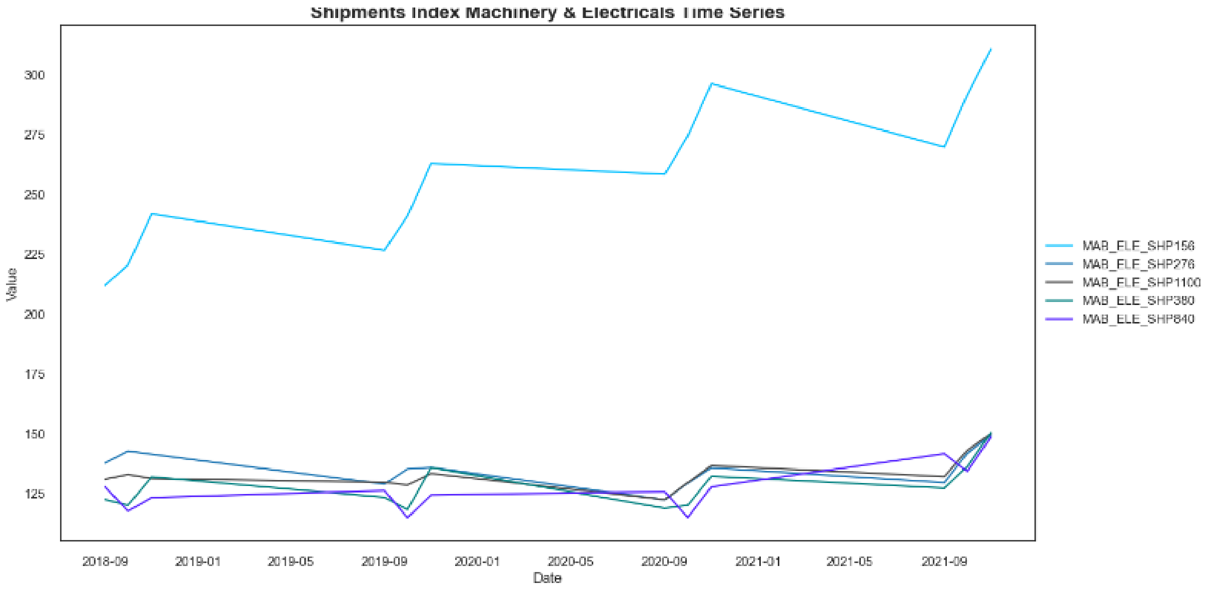


***Figure 12* -** "Shipments Index Machinery & Electricals Time Series"

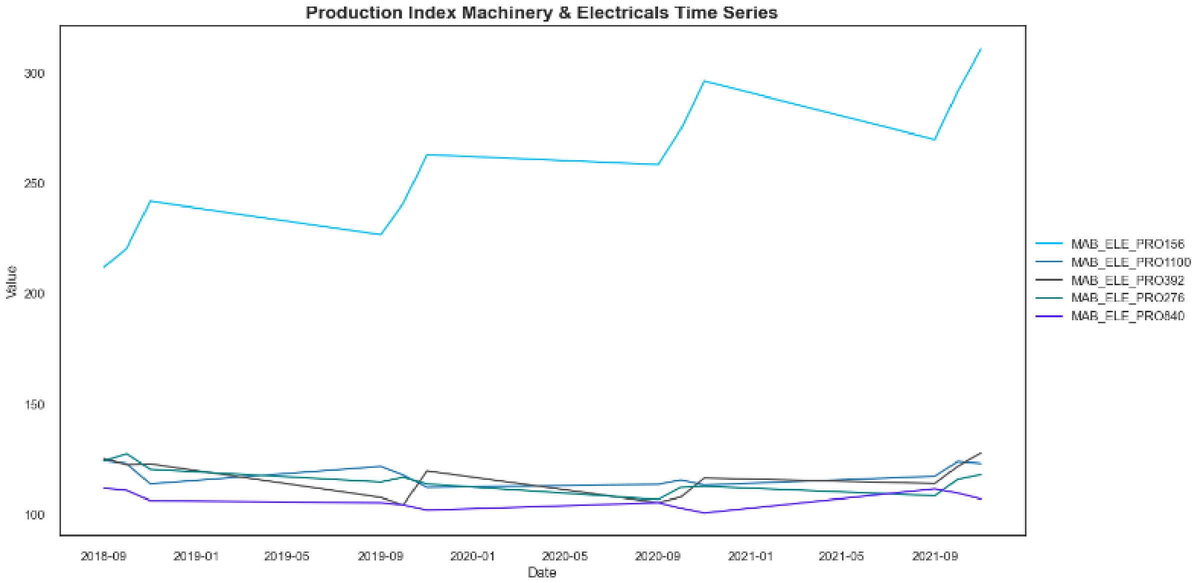


***Figure 13* -** "Production Index Machinery & Electricals Time Series"

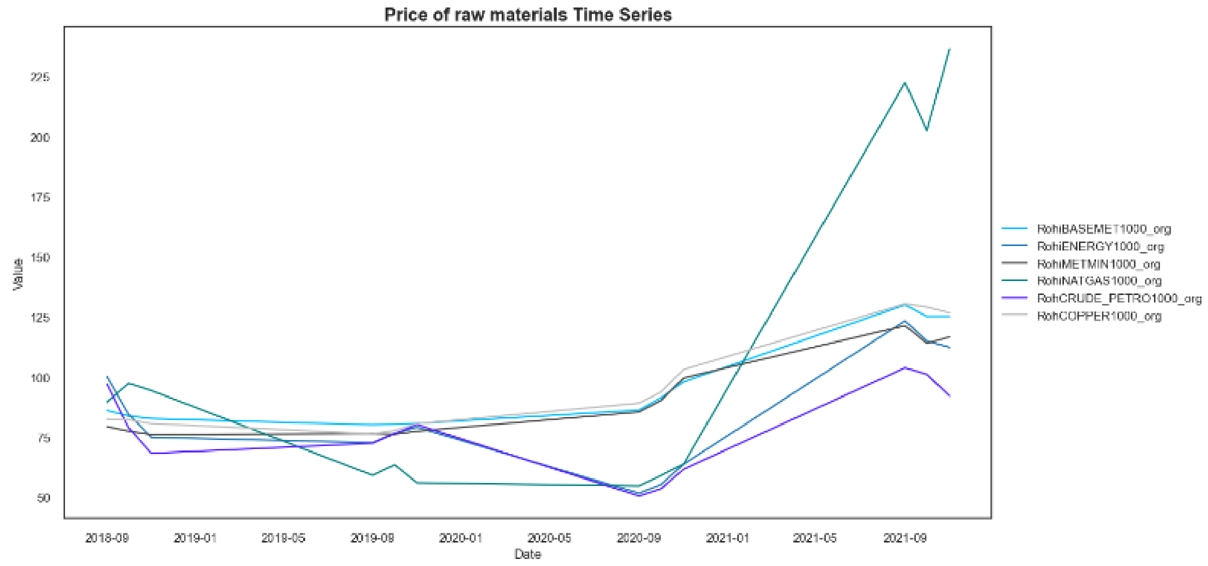


Figure 14 - "Price of raw materials Time Series"

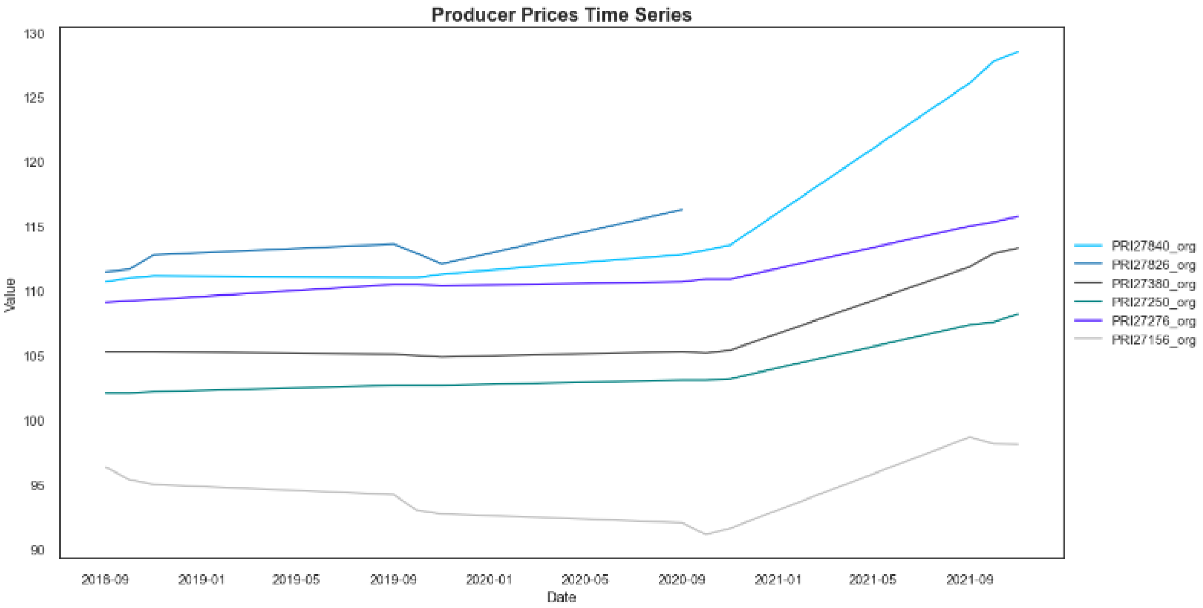


Figure 15 - "Producer Prices Time Series"

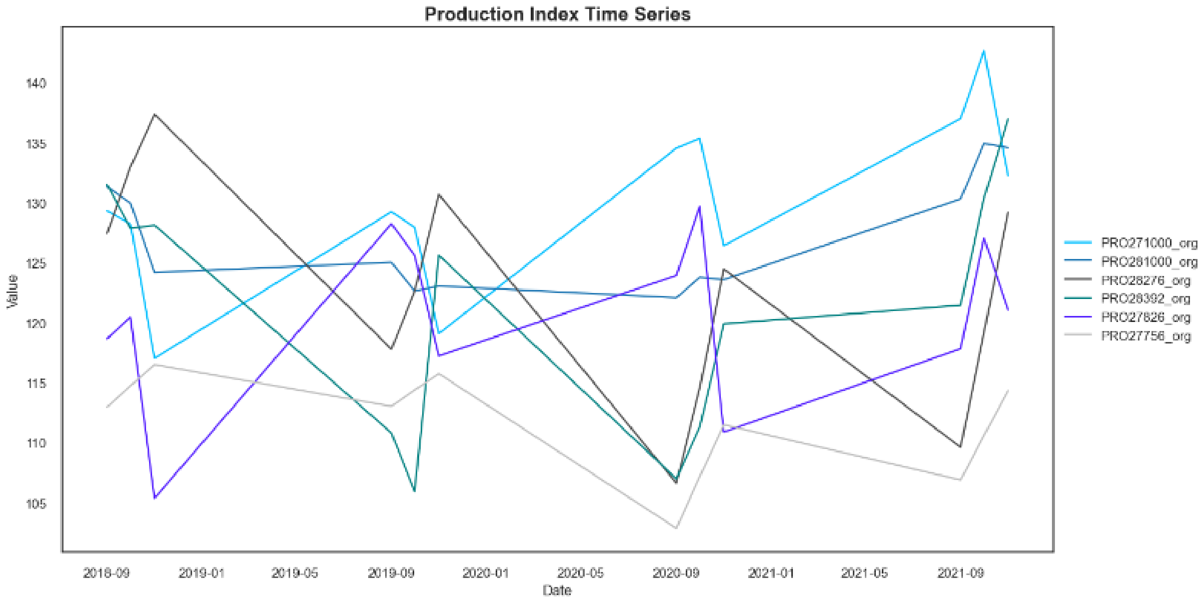


Figure 16 - “Production Index Time Series”

Figure 5 - “Dataset correlation Spearman Heatmap”