NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 3: Monthly sales forecast

Carolina, Costa, number: 20220715

Martim, Santos, number: 20220540

Pedro, Pereira, number: 20220684

Rodrigo, Silva, number: 20221360

Rúben, Serpa, number: 20221284

Group C: Young Talent Consulting Group

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

May, 2023

INDEX

1. EXECUTIVE SUMMARY

Through the power of data, the Young Talent Consulting Group made an understanding of the existing sales records and identified unique sales patterns among different product groups, designing a ten-month sales forecast for the Siemens Business Unit, in Germany, to thrive. The outputs, including recommendations and timelines, are provided in detail at the end of the report.

In sum, the sales and market data shared by Siemens comprised a great foundation for the success of this project although a strong dataset restructure was needed for the macro-economic data. The team also captured external data referring to the evolution of Covid-19 cases and Germany school holidays during the years of analysis judging that they could be an impactful driver for the business unit sales and inventory fluctuations. Proper visualizations such as heatmaps and line charts were also a key factor in understanding sales composition and various factors influencing its quantities.

By conducting a detailed analysis with the support of statistical modelling, XGBoost and others, an automated price recommendation based on the history of sales offers was created in accordance with the findings and a list of business applications were meticulously thought out based on outcomes, with the purpose of improving sales margin and ultimately create a more efficient and transparent business process.

It is believed that by implementing the given recommendations and with appropriate model maintenance to keep the efficiency of the given predictions, the business unit will be able to anticipate demand prior to the time and adjust its supply chain accordingly.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. INDUSTRY OVERVIEW

On an industry level, the global smart infrastructure market is projected to grow from USD 97.20 billion in 2021 to USD 434.16 billion in 2028 at a CAGR (Compound Annual Growth Rate) of 23.8% during the 2021-2028 period (*Smart Infrastructure Market Size, Share & Forecast [2021-2028]*, n.d.). Various growth drivers play a role in the positive evolution of this market as can be seen in Table 1. The rapid development of new and improved technologies in areas such as artificial intelligence and big data allows for rapid improvement and innovation breakthroughs in this industry. Adding to this, the enormous population increase and further urbanization developments in the past decades created more pressure for an efficient and sustainable resource usage. Therefore, it is vital that organizations adapt to the usage of energy-efficient power distribution infrastructures. The increasing awareness on climate change pleads for a need of greenhouse gas emissions reduction and the implementation of sustainable infrastructures provided by this market.

Growth Drivers	Advancements in technology
	Increasing Focus on Sustainability & Switching to Green Energy Resources
	Increasing urbanization

Table 1: Growth Drivers for the Smart Infrastructure Market

In terms of market composition, the market can be segmented into five mains segments: Smart Grid, Intelligent Buildings, Intelligent Transportation Network, Small Water Network, Others. The business

unit offers products under the smart grid and intelligent building segments. As it can be seen in Figure 1, these ones accounted for the biggest portion of the market in 2020 with smart grid having 37.9% of the total market share.

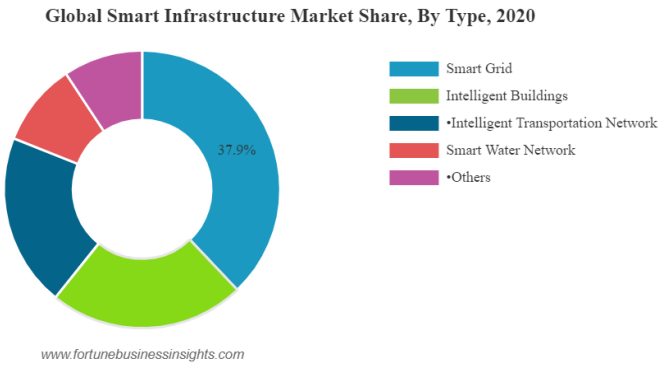

Figure 1- Smart Infrastructure Market Share in 2020

The Siemens Smart Infrastructure Division, specifically one of their business units in Germany, focuses on the development and implementation of smart power distribution solutions in different organizations and transversal to different industries for a more efficient, digital-driven and sustainable infrastructure.

The progressive evolution expected provides Siemens a great opportunity to continue to thrive in this market as one of its biggest players. Like in practically every market, Siemens is not alone. Some of the top manufacturers and competitors in the Smart Infrastructure market includes the following: ABB, Eaton, Mitsubishi Electric Power Products, Schneider Electric, General Electric.

The business unit product-offering ranges mainly from low-voltage to medium-voltage systems and solutions. Usually the first is commonly used for residential, commercial and light industrial applications while the second one focuses on larger industrial applications such as data centers and manufacturing facilities with the support of smart grids.

2.2. Business Objectives

The Smart Infrastructure industry is at a fast-growing rate, with various factors sustaining this growth. Even though this is the case, sales are extremely sensitive to external factors and therefore companies should be capable of containing such impacts on their businesses.

Based on this, the Siemens business unit based in Germany hopes to prevent future events from affecting their business sales performance and supply chain by utilizing past sales data from selected product groups of their Smart Infrastructure Division.

To do so, the company intends to gain data-driven insights from a sales forecast to achieve the following business objectives:

- **Increase sales margins**: It is clear that the industry is rapidly expanding and with that, the opportunity to increase sales becomes clearer than ever. But it is not as simple as that. With this project outcomes, the company would be able to predict demand of certain product categories during certain periods and based on that to properly adequate its supply chain, increasing production and inventory efficiency reducing efficiency costs and ultimately

increasing sales margins. Over the next 10 months of prediction, the aim would be to increase sales margins by a total average of 20% according to present margins.

- **Improve consistency and transparency of quotes**: Transparency within a business, both internally and externally, can help an organization promote growth, increase sales and maintain efficiency (Indeed Editorial Team, 2022). Production and material prices fluctuate at the slightest change. This change could come in various ends and in various forms, but the main point is that with the right anticipated knowledge, the company can act in a prompt way.

2.3. Business Success criteria

Business success is going to depend on the ability to create a 10-month sales forecast, in a data-driven and concise way. The Young Talent Consulting Group has defined the following objectives with the purpose of achieving success:

- Primarily, the focus goes on generating an **automated price recommendation** based on the sales history. With the support of machine learning algorithms and techniques, the success comes down to the ability of creating a good forecasting model that neither is highly accurate, not considering sudden factors not imputed in the model, or has a low accuracy, not taking advantage of the data provided and potentially losing valuable opportunities for the company to improve.
- In order to achieve the **20% increase over past total average sales margins**, the company should use the information obtained to **control supplier purchases and inventory levels**. This supply chain control will make it more productive and reduce total costs.
- To **improve consistency and transparency of quotes**, honesty and clarity should be nurtured from inside the company and further outside. From employees to suppliers and customers, either by **negotiating with suppliers** to contain price spikes, inform clients beforehand of market pressures or just create **detailed cost breakdowns**. In terms of price consistency, the success of achieving it passes through a **quote standardization** by creating a consistent and clear process. Further details about this can be found in the recommendation section of this report.

With the purpose of achieving success, an extensive data exploration analysis and proper modelling, by product category, is also going to be made in hopes of better understanding which factors have a stronger weight in product sales and to have concrete predictions depending of each product sales behavior. To evaluate model performance, some methods are going to be applied, including the R-squared, MAE and RMSE among others.

2.4. Situation assessment

For the project realization, various datasets were utilized including not only the ones handled but also two additional external datasets to help define sales fluctuations across time.

- Sales dataset: This dataset contains information about daily total sales for each of the 14 different product groups chosen. The dataset has 9802 registered sales and 3 variables.
- Market dataset: This dataset contains information regarding macro-economic indices and prices around some of the most important countries in Siemens business. This includes production indexes and raw material costs. The dataset has a total of 222 rows and 48 variables which needs a solid restructure due to lack of clear understanding.

- Covid-19 dataset: This dataset contains information about the pandemic evolution across countries around the globe and conditions such as hospitalizations, lockdowns and vaccinations. It is included to help identify any possible correlations between Covid-19 cases, lockdowns and sales patterns. The dataset has 306057 entries and 67 columns.
- Inflation dataset: This dataset contains information regarding consumer price indexes in Germany. It was retrieved by the Federal Statistical Office of Germany and helps to track inflation. The dataset has 231 rows containing the evolution of the consumer price index from January 2004 and March 2023.
- GDP dataset: This dataset shows the evolution of the Gross Domestic Product for Germany during 2004 and 2022. The dataset is structured where each input is the Germany GDP of a certain quarter of a year.
- Germany School Holidays and Suez Canal Obstruction dataset: This dataset includes information regarding school holidays in Germany and disrupted global trade days due to the 2021 Suez Canal Obstruction event. The dataset has a total of 3 columns and 690 dates or range of dates registered depending on the region of Germany since each region has its own holidays days.

During the development of this project, the Young Talent Consulting Group imported various libraries in order to achieve all the proposed objectives such as the following: matplotlib; scikit-learn; seaborn; scipy; numpy; pandas; network and warnings.

Some of the biggest risk factors involved are model limitations towards punctual and unforeseen events. Even though part of the model refinement comprehends the usage of data related to the sales, the inclusion of those does not eliminate the impact of these events but rather substantially reduces them compared to predicting sales amounts only using the sales data itself.

In order to tackle these risks, the business unit must conduct regular model maintenance and input updated data from time to time, for example, monthly.

2.5. Determine Data Mining goals

This project is going to follow a sequence of phases, industry known as CRISP-DM. To fulfil the proposed business objectives, three main technical checkpoints exist for each product group:

- **Model choice** according to stationarity and autocorrelation analysis.
- **Reduce Mean Absolute Error (MAE), Root Mean Square Error (RMSE)**: Both metrics calculate the difference between predicted and actual values with MAE measuring the average absolute difference and RMSE using the square root of the squared errors. Due to the high variability of the data achieving a low MAE and RMSE is highly unachievable so the main goal here is to reduce this metrics output to the maximum possible.
- **Reduce Mean Absolute Percentage Error (MAPE) and MAX Error**: MAPE measures the average percentage difference between predicted and actual sales while Max Error represents the maximum difference between predicted and actual values. The goal here is to reduce the MAPE to a specific percentage while reducing Max Error so the model even though may have some errors, and these are limited to a certain extent, not making large errors in its predictions.

- **Increase R-Squared**: Due to the importance of understanding sales fluctuation it's important to focus on obtaining a model with a strong R-squared value, meaning a model where most of the variation in the data is explained by it.

3. METHODOLOGY

3.1. DATA UNDERSTANDING

Before progressing to the data understanding phase, the three datasets were meticulously analysed ("*market_data*", "*sales data*", and "*covid_data*") using Excel. Furthermore, it was ensured that the data types were in their appropriate formats, which facilitated seamless processing and analysis. In addition, certain variables that did not contribute any value to the project were identified and eliminated, streamlining the datasets for improved intuitiveness and efficiency. This preliminary data cleaning and transformation allowed us to create a solid foundation for the subsequent data understanding phase.

3.1.1. Market Data

The "*market_data*" dataset contains historical data on various **economic indicators and indices**, focused on the **Electrical Equipment and Machinery sectors**. These variables are collected from different countries and are related to the price of commodities, production and shipments indices, and producer prices. This data will be valuable in understanding the factors affecting the sales of the electrical equipment and machinery sectors.

The dataset contains **47** columns, which includes global prices indexes for raw materials and commodities, EUR/USD exchange rate evolution. It also includes producer prices, production indexes, and shipments index for electrical equipment and machinery in various countries, such as China, France, Germany, Italy, Japan, Switzerland, the United Kingdom, and the United States.

A few columns have missing values, including *Producer Prices Electrical Equipment_United Kingdom* and *Shipments Index Machinery & Electricals United Kingdom* with **18** missing values, and *Production Index Electrical Equipment_Switzerland, Production Index Machinery & Electricals_Switzerland, Production Index Machinery and Equipment N.E.C._Switzerland, Shipments Index Machinery & Electricals_Switzerland, Shipments Index Machinery & Electricals_United States* with **1** missing value. These missing values will be addressed during the data pre-processing stage of the analysis.

Overall, this dataset provides a comprehensive overview of various factors that may influence sales. By analysing this data, it is possible to understand clearly how these factors contribute to sales forecasting for the target market.

Having a first look at the market data provided some insights could be taken. The team decided to substructure the data according to its typology. Therefore, the evolution of certain prices and indexes was plotted using different line charts.

Raw Material Costs

First, the evolution of raw materials prices around the globe was outlined in Figure 2. It could be observed a more or less similarity with price fluctuations between these materials pointing out some major decline during the initial Covid-19 pandemic spread-out especially on energy and crude oil

prices. From there, as countries tried to recover the demand for raw materials surged putting pressure on suppliers and consequently rising prices. Natural gas prices had an exponential rise due to disrupted production in the beginning of 2021 and later due to limited storage capacity.

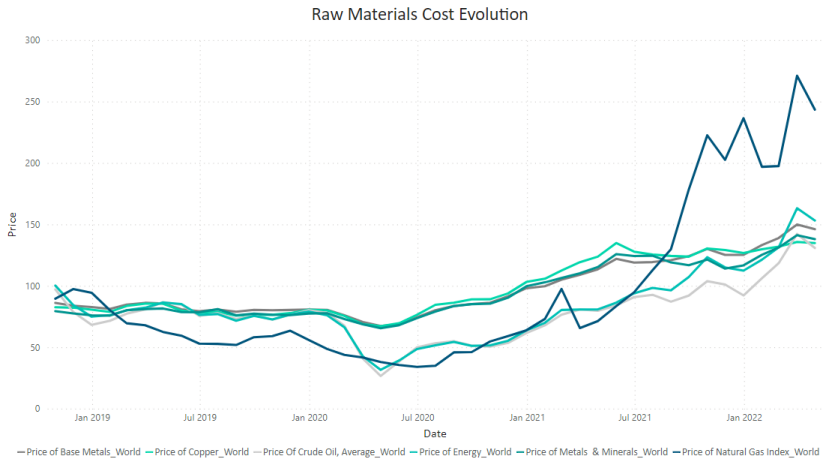

Figure 2- Raw Materials Cost Evolution around the Globe

Producer Costs

Looking at the evolution of producer costs in Figure 3, between the initial and final period of analysis prices had an overall increase. The important thing to highlight is in producer costs of United Kingdom manufacturers. The last entry data available for this country is in the last quarter of 2020. After some research to understand what could have happened, it could be concluded that this was majorly due to Brexit negotiations reaching a consensus with the UK-EU Trade and Cooperation Agreement on December 24, 2020, which came into effect on January 1, 2021. It establishes arrangements for future co-operation across a range of areas including trade, aviation, road haulage, fisheries, police and security, health insurance and continued UK participation in some EU programmes (*What Is the Trade and Cooperation Agreement? - UK in a Changing Europe*, 2021). The changes, uncertainty around regulations and concerns for future trade relations could have led the business unit to cut supply chains of electric equipment from the UK. The graphic also supports this idea and therefore from this point, the United Kingdom manufacturer indicators are going not going to be consider for the analysis.

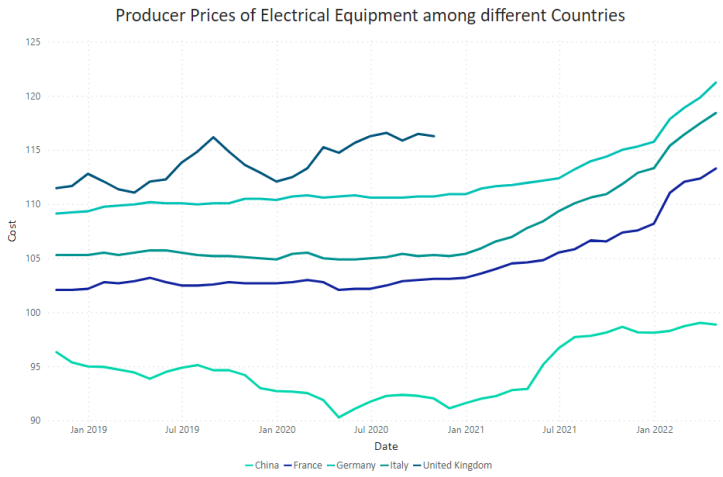

Figure 3- Evolution of Producer Costs in China, France, Germany, Italy and UK

Production Performance & Shipment Performance

Industry performance indexes were also analysed. Both production and shipment indexes had similar behaviours with seasonal performance drops during the third quarter of each year of analysis. This could be due to several factors like slower industrial activity during those months.

3.1.2. Covid Data

The "*covid_data*" dataset provides information on new COVID-19 cases, deaths and other factors across various countries. The dataset contains 306057 entries and 67 columns. Additionally, the dataset has been normalised by the **number of cases per 100,000 inhabitants** for each location. This step was taken to enable a better comparison of the impact of the pandemic in each location. A dictionary named '*populations_100k*' containing the count of 100k habitants for each country or region was created. This normalization will aid in the visual exploration and analysis of the dataset.

The foremost industry that is withering away because of coronavirus (COVID-19) is Manufacturing and logistics. Even though mobile solutions have played a huge role in empowering the manufacturing industries, the virus outbreak has slowed down the industry performance (Appinventiv, 2022). It is also important to include pandemic data in the analysis to understand better the pandemic spread across important countries and the possibility of it influencing sales amount. Having a general look at the data it can be seen different waves of new cases rising across specific time periods and is important to further down see if any of them relate to sales numbers.

3.1.3. Sales Data

Upon examining the "*sales_data*" dataset, it is possible to understand the sales trends for 14 products. The dataset contains **43** records, each representing monthly sales figures from 2018-10-31 until 2022-04-30, and is well-structured with no missing values. After some reshaping, the data is well-structured, and the varying sales figures among products can be used to inform future business strategies and decision-making.

Stationarity and ACF/PACF analysis

To ensure a rigorous approach to Modelling, it was necessary to assess the stationarity of each product over time and examine the total and partial autocorrelation in the data. This analysis was undertaken to identify any temporal dependencies present in the data and to inform the selection of appropriate models, such as autoregressive (AR) or moving average (MA) components.

As Shumway and Stoffer (2017) explain, stationarity is a crucial concept in time series analysis, which refers to a property of the data where statistical properties such as mean, variance and covariance remain constant over time. This characteristic is important because a stationary time series is generally easier to model and can produce more accurate predictions than a non-stationary one.

After analysing the stationarity plots in the section of the notebook labelled "Stationarity and ACF and PACF Analysis", it was observed that all of the products were stationary with the exception of 8, 12, 13 and 20. This means that the statistical properties such as the mean, variance, and covariance of these three products were changing over time, which can make them more difficult to model and analyse. Based on Shumway and Stoffer (2017) this technique is a common approach for handling non-stationary data in time series analysis and can help to ensure that any modelling and predictions based on the data are more accurate and reliable.

According to Chatfield (2019), autocorrelation is a statistical concept that refers to the correlation between values in a time series at different lags. Total autocorrelation (ACF) measures the correlation between any two observations in the time series, while partial autocorrelation (PACF) measures the correlation between two observations after taking into account the influence of other observations in between. These measures are important in determining the number of lags to include in time series models such as autoregressive (AR) or moving average (MA) models, and can aid in selecting appropriate models that accurately capture the temporal dependencies present in the data.

Furthermore, the total and partial autocorrelation plots were also analysed, and the following observations were made Product 1 exhibited no autocorrelation, indicating that lag variables cannot be used to predict this product. Similarly, product 3 had no total autocorrelation or partial autocorrelation (except for one lag value - 12), while product 4 exhibited no total autocorrelation or partial autocorrelation except for six lag values (6,16,17,18,19,20) in the partial autocorrelation plot. Product 5 also had no total autocorrelation or partial autocorrelation except for one lag value for both (6), while product 6 exhibited no total autocorrelation or partial autocorrelation except for one lag value (17). In contrast, product 8 showed significant total and partial autocorrelation, indicating that lag 3 can be used to predict this product, which also appears to have seasonality. For product 9, there was no total autocorrelation or partial autocorrelation (except for five lag values (9,10,12,19,20) in the PACF plot and only one lag for ACF (12)). Similarly, product 11 exhibited no autocorrelation, meaning that lag variables cannot be used to predict this product. However, product 12 showed a significant pattern, suggesting the possible use of an ARMA(3,3) model. Product 13 also exhibited significant total and partial autocorrelation, indicating that lag 3 can be used to predict this product, which also appears to have seasonality. Product 14 showed no total autocorrelation or partial autocorrelation, except for one lag value (6) in the partial autocorrelation plot. For product 16, there was no total autocorrelation or partial autocorrelation except for one lag value for both the PACF and ACF plots (3). Finally, product 20 exhibited no total autocorrelation or partial autocorrelation, except for four lag values (8,10,13,20) in the PACF plot.

STL Decomposition

A Seasonal-Trend Decomposition Procedure Based on Loess was also applied to each product group. Unlike additive or multiplicative decomposition, this procedure is based on the Loess regression to estimate trend and seasonal components. According to Cleveland (1990), one of its main advantages towards the above mentioned include robust estimates of the trend and seasonal components that are not distorted by aberrant behaviour in the data, allowing for a better performance in capturing nonlinear patterns, such patterns clearly demonstrated in the data given.

Sales Analysis

A key insight gained from the dataset is the **varying sales performance** across 14 products. While some products have sales in the millions, others average only a few thousand. The dataset also reveals differences in sales volatility, with some products exhibiting relatively high standard deviations compared to their mean sales figures.

Firstly, it is important to have a look at global monthly sales pictured in Figure X. There were some major sales declines months, especially in January 2021, a decline affecting all business product groups. Comparing this to the market data, on an initial analysis it can be related to Brexit negotiations concluded and implemented in January 2021 or due to the material cost rises or even due to the

second Covid-19 wave in Europe that started on the last quarter of 2020 and extended until the end of the first quarter of 2021. There were also some major sales rises, around September of each year.

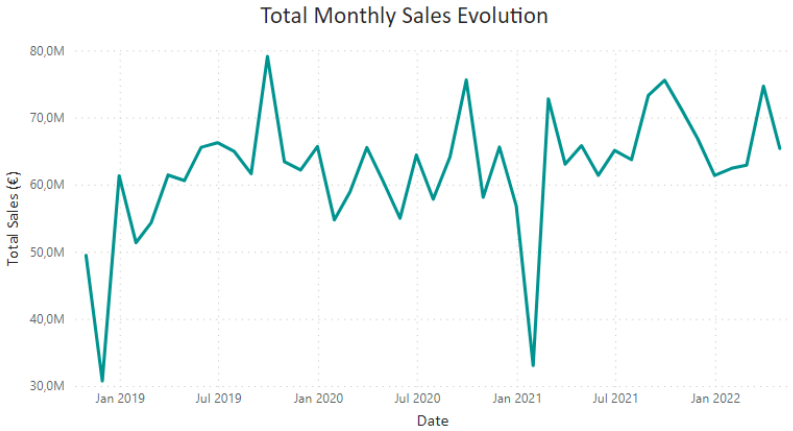

Figure 4-Global Monthly Sales

Figure 5 show a heatmap containing an aggregation of monthly sales values by product type. This visualization allows for a quick and easy to capture way of understanding which products contribute to larger sales, which products have had zero or negative sale months and the evolution of their sales. From looking at it, it can be concluded that Product category 1, 3 and 5 have the largest sales out of the 14 with product 1 clearly being the company's biggest product, sales-wise. Products 6, 9, 14 and 20 had negative or zero sale months with Products 9, 13, 14 and 20 being the company's smallest products, sales-wise.

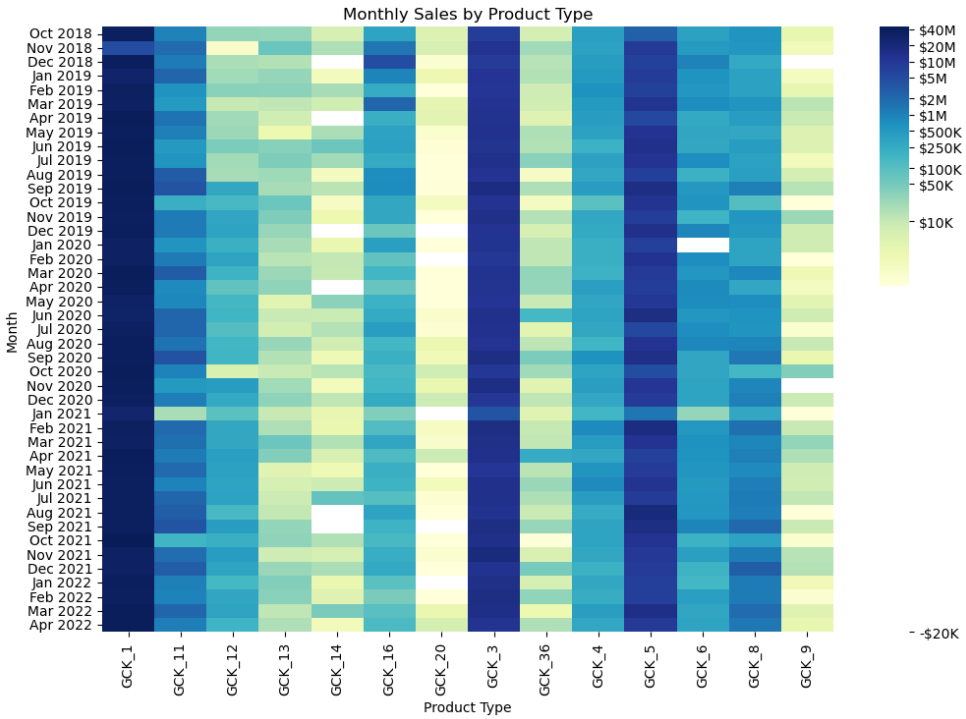

Figure 5- Heatmap illustrating Monthly Sales by Product Type

Analysing possible trends, certain products showed interesting patterns. Figure 6 various line plots for the evolution of sales by product group. Product 1 and 3 showed overall constant sales values with a strong breakdown in one of the months.

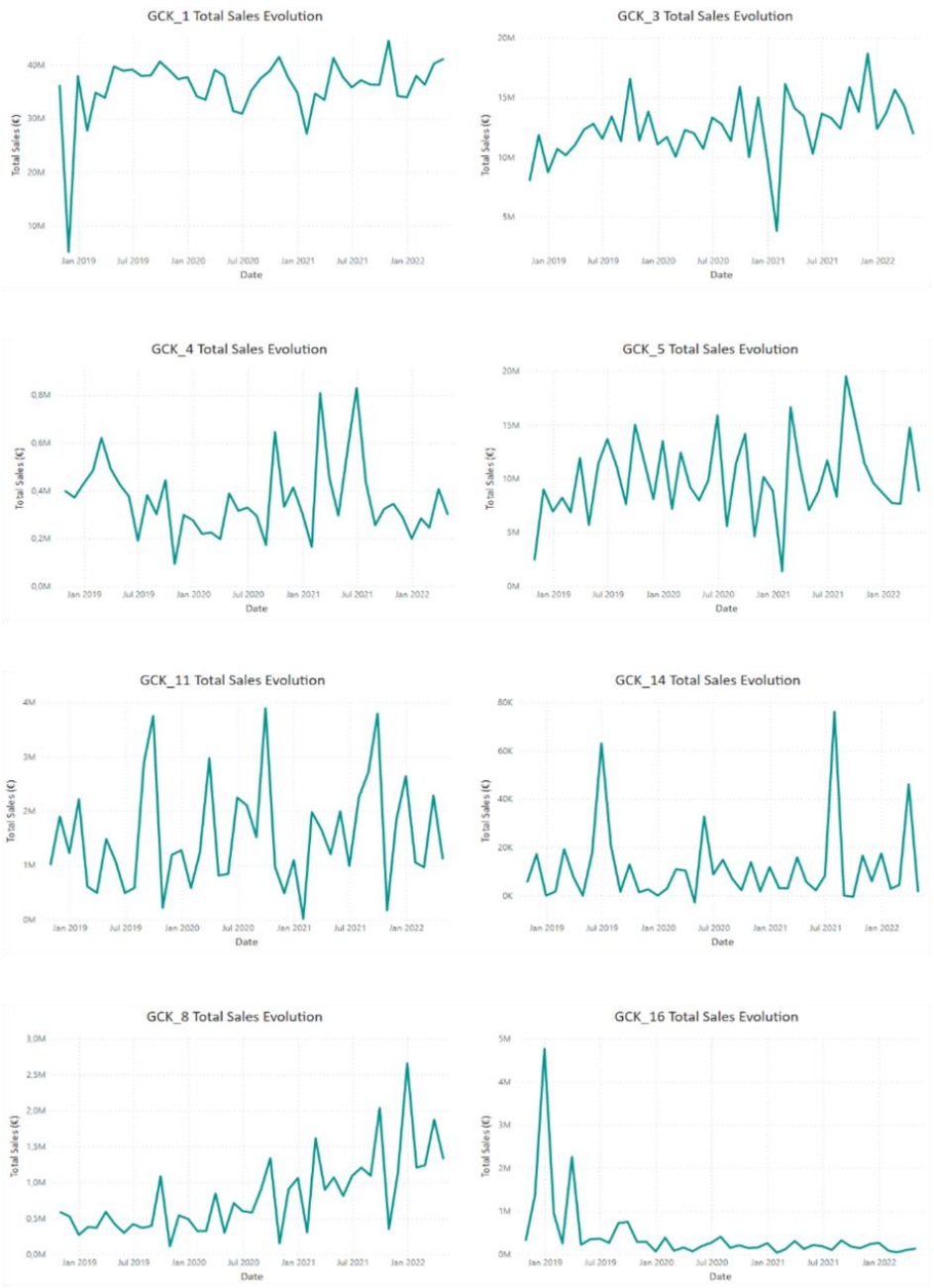

Figure 6- Products' Sales Overview

Across the timeline there were two products with a clear positive/negative evolution with product 8 growing in sales and product 16 having an unusual sales month and from there slowly decreasing its sales. Product 11 seems to have a cyclic behavior having exceptionally great sales every September and dropping sales values by at least more than two thirds right on the next month. Additionally, some trend similarities can be observed between product 1 and 3, and product 4 and 20, respectively, while product 4 and 5 showed contrary behavior. Product 14 sales had a stable trend during 2020 after a constant decline trend in the past year and with a positive trend after it. This is majorly affected by the two high sales months verified in 2021 and 2022.

Regarding COVID-19 impact on sales, overall, the pandemic adversely impacted sales, with GCK_5 and GCK _12 being the most prominent examples. However, products such as GCK _8 exhibited a moderate correlation with the rise in cases across Europe, resulting in a boost in sales for these items.

Furthermore, concerning the obstruction of the Suez Canal, for products GCK _3, 4, 5, 8 and 11, it had a negative impact on sales. In contrast, for products GCK _6, 9, 13, 14 and 16, there was an increase in sales during the peak of the Suez Canal crisis, suggesting that those products do not come from Asian countries.

3.2. DATA PREPROCESSING

3.2.1. Fixing Inconsistencies

The "*covid_data*" is reindexed to a monthly frequency using the sales data index. Missing values are filled with 0, and the index is renamed to 'date' for consistency purposes.

3.2.2. Missing Values Treatment

It was decided to drop the columns with more than one missing value after the date of 2016-12-31, as it is the latest information needed to obtain the lagged variables, thus preventing a higher bias when predicting with imputed values. Thus, 'Producer Prices Electrical Equipment_United Kingdom' and 'Shipments Index Machinery & Electricals_United Kingdom' were dropped. For the columns with only one missing value, these were filled with the previous value, as these indexes do not vary considerably from month to month, hence it is a good approximation.

3.2.3. Feature Engineering

The goal of Feature Engineering was to develop new features which encompass some external factors that could potentially impact SIEMENS' sales. Additionally, all the features were converted to lagged variables in order to capture the temporal dependence structure of the data.

Adding Exogenous Features:

One key aspect that the company has identified as influencing sales is vacation days. During the year, there are certain periods when sales are impacted by reduced work activity. Therefore, a variable **Holidays-to-Month Ratio** was created, which measures the number of holidays in a given month compared to the total number of days in that month. In March 2021, one of the busiest trade routes in the World, the **Suez Canal** was blocked for almost a week, thus disrupting global trade and impacting markets heavily. Since this was an important event which negatively affected the products' shipment, four days of holidays were added in the respective time frame.

Finally, two important economic indicators were added, the **Gross Domestic Product (GDP)** and the **Consumer Price Index (Inflation)** for Germany. A high inflation environment not only impacts business' costs, but also decreases consumer's purchasing power, which influences the sales. Regarding GDP, when it is growing, it can lead to increased consumer confidence and spending, thus resulting in higher sales for businesses.

Converting to Lagged Variables

Understanding the amount of time, it takes for sales to be affected by changes in market data variables is crucial in the forecasting process. However, when choosing the appropriate time frame of the exogenous variables, a relevant issue emerged regarding the availability of data. Since the goal is to predict sales for the next 10 months, assuming a lag lower than 10 would demand a forecast of the exogenous variables as well. Therefore, for the sake of simplicity, it was decided to identify the optimal lag within the range of 10 to 18 months. The 18 months choice considers the volatility and speed of

change of the markets nowadays. Concerning the sales data, multiple lags were tried as an attempt to improve model's accuracy. The final choice was 8 months.

3.2.4. Feature Selection

Regarding Feature Selection, it was decided to define the set of relevant features for the forecast for each product individually. Furthermore, to understand not only which features should be selected for the model, but also to understand what the optimal lag for the exogenous variables is, a correlation analysis was performed.

Firstly, the top 20 features were selected for each product by analysing the correlation between each market data variable and the target variable, this is each product sales in April 2022. Afterwards, redundant features, with a correlation higher than 0.8, were removed, maintaining the variable most correlated with the target. Finally, the **top 5 features** were selected as the ones having highest correlation. Table 1 summarises the features selected for each product.

Product	List of Features
GCK_1	'Producer Prices Electrical Equipment_Italy(t-16)', 'Producer Prices Electrical Equipment_Italy(t-15)', 'Producer Prices Electrical Equipment_Italy(t-14)', 'Producer Prices Electrical Equipment_United States(t-11)', 'Shipments Index Machinery & Electricals_Italy(t-10)'
GCK_3	'Production Index Machinery & Electricals_China(t-11)', 'Production Index Electrical Equipment_World(t-12)', 'Production Index Machinery And Equipment N.E.C._Germany(t-13)', 'Producer Prices Electrical Equipment_China(t-17)', 'Production Index Machinery And Equipment N.E.C._Italy(t-15)'
GCK_4	'Shipments Index Machinery & Electricals_Italy(t-13)', 'Production Index Electrical Equipment_United States(t-10)', 'Production Index Electrical Equipment_World(t-13)', 'Shipments Index Machinery & Electricals_Europe(t-13)', 'Production Index Machinery And Equipment N.E.C._World(t-13)'
GCK_5	'Production Index Machinery & Electricals_Japan(t-13)', 'Production Index Electrical Equipment_Japan(t-10)', 'Production Index Machinery & Electricals_China(t-15)', 'Shipments Index Machinery & Electricals_United States(t-12)', 'Shipments Index Machinery & Electricals_Japan(t-16)'
GCK_6	'Production Index Machinery & Electricals_United States(t-18)', 'Production Index Electrical Equipment_Switzerland(t-18)', 'Production Index Electrical Equipment_United States(t-18)', 'Price Of Natural Gas Index_World(t-18)', 'Producer Prices Electrical Equipment_Italy(t-17)'
GCK_8	'Producer Prices Electrical Equipment_China(t-17)', 'Producer Prices Electrical Equipment_China(t-15)', 'Production Index Machinery And Equipment N.E.C._Japan(t-16)', 'Producer Prices Electrical Equipment_Germany(t-10)', 'Producer Prices Electrical Equipment_China(t-13)'
GCK_9	'Production Index Machinery & Electricals_United Kingdom(t-11)', 'Production Index Electrical Equipment_Germany(t-10)', 'Production Index Machinery And Equipment N.E.C._United States(t-11)', 'Production Index Machinery And Equipment N.E.C._United Kingdom(t-11)', 'Shipments Index Machinery & Electricals_France(t-11)'
GCK_11	'Production Index Machinery And Equipment N.E.C._France(t-13)', 'Shipments Index Machinery & Electricals_Italy(t-13)', 'Production Index Machinery And Equipment N.E.C._Italy(t-13)', 'Shipments Index Machinery & Electricals_France(t-16)', 'Production Index Electrical Equipment_Switzerland(t-15)'
GCK_12	'Consumer Price Index(t-16)', 'Producer Prices Electrical Equipment_China(t-12)', 'Consumer Price Index(t-18)', 'Producer Prices Electrical Equipment_United States(t-14)', 'Producer Prices Electrical Equipment_Germany(t-16)'
GCK_13	'Production Index Machinery & Electricals_China(t-14)', 'CPMNACNSAB1GQDE(t-10)', 'Production Index Electrical Equipment_Switzerland(t-12)', 'Production Index Electrical Equipment_Switzerland(t-11)', 'Production Index Electrical Equipment_Switzerland(t-13)'

GCK_14	'Production Index Electrical Equipment_Switzerland(t-13)', 'Production Index Electrical Equipment_United Kingdom(t-14)', 'Production Index Machinery & Electricals_China(t-17)', 'Production Index Electrical Equipment_Switzerland(t-12)', 'Shipments Index Machinery & Electricals_Italy(t-12)'
GCK_16	'Producer Prices Electrical Equipment_Italy(t-17)', 'Producer Prices Electrical Equipment_Italy(t-16)', 'Producer Prices Electrical Equipment_United States(t-14)', 'Producer Prices Electrical Equipment_Germany(t-12)', 'Producer Prices Electrical Equipment_United States(t-17)'
GCK_20	'Production Index Machinery & Electricals_China(t-14)', 'Production Index Electrical Equipment_United Kingdom(t-17)', 'Production Index Electrical Equipment_United Kingdom(t-13)', 'Production Index Electrical Equipment_Japan(t-13)', 'Production Index Electrical Equipment_Italy(t-17)'
GCK_36	'Production Index Machinery & Electricals_United Kingdom(t-12)', 'Production Index Electrical Equipment_France(t-12)', 'Production Index Machinery And Equipment N.E.C._United Kingdom(t-12)', 'Production Index Electrical Equipment_Germany(t-12)', 'Price Of Crude Oil, Average_World(t-12)'

Table 1- Final Features Selection

3.3. MODELING

Regarding the modelling, two approaches have been defined:

- For products that exhibit **stationarity** before or after applying differencing and show **no significant total or partial autocorrelation**, the mean, median, and XGBoost models will be applied;
- For products that exhibit **stationarity** before or after differencing and show **significant total or partial autocorrelation**, the Facebook Prophet model will be applied, along with autoregressive integrated moving average (ARIMA) using the previously defined lag values, XGBoost, and an ensemble method.

In the realm of sales forecasting, there are various techniques that can be employed to develop accurate predictions. As stated by Chatfield (2000), **ARIMA** is among the most widely used techniques. It is a time series analysis and forecasting model that takes into account the autoregressive, moving average, and differencing components of a time series to make predictions.

Taylor and Letham (2018) state that **Facebook Prophet** is capable of automatically handling seasonality, trend changes, and outliers. Additionally, Prophet can model multiple seasonality cycles, making it a versatile tool for time series analysis.

In addition to these advanced techniques, simple statistical measures like **Mean** and **Median** can also be utilized for sales forecasting. While these techniques may be less sophisticated than ARIMA and Prophet, they can still be effective in certain situations, particularly when the data is relatively stable and predictable.

Various other techniques can be employed for time series forecasting, among which are machine learning models that use historical data to learn the patterns and relationships in the data and make predictions about future values, and they provide a "*means to learn temporal dynamics in a purely data-driven manner*" (Lim, B, Zohren S. 2021). These models can capture complex patterns and can provide higher accuracy than traditional statistical methods, especially for non-linear time series data. Nevertheless, usually they require large amounts of data to train the model effectively, and they tend to overfit the training data, which may lead to poor generalization to new data.

XGBoost is a variation of Gradient Boosting that is designed to be highly scalable and efficient. As an ensemble technique, Gradient Boosting combines the result of several weak leaners, usually Decision Trees, to build a model which outperforms a conventional single machine learning model. In particular, XGBoost uses advanced regularization which improves model generalization and reduces overfitting, and it has an in-built capability to handle missing values.

For the purpose of producing an accurate and outstanding model, three main steps were taken during the Modelling stage:

1. Split the Data into train and validation/Define target and independent variables:

When splitting the data, the training set is used to train the model, and the validation set to test how the model performs when generalizing to new data. This is relevant to identify and address issues such as overfitting. Moreover, it allows hyperparameter tunning of the model, which is critical for the model's performance. In this case, the validation set includes the **last 10 months** of the dataset to be as equal as possible to the test set.

In summary, the train set includes data from October 2018 until June 2021 (**2 years and 8 months**), the validation set ranges from July 2021 until Abril 2022 (**10 months**), and the test set includes the time range from May 2022 until February 2023.

Furthermore, the target and independent variables were defined for each product according to the previously selected features.

2. Time Series Forecasting for the Validation set:

On a preliminary attempt, the models Mean, Median, Prophet and Arima were tested individually for each product using exclusively past sales information. Subsequentially, another model was applied for both Prophet and ARIMA, with the inclusion of exogenous variables.

Regarding the modelling phase using XGBoost, it was critical to proceed with hyperparameter tuning of the model, mainly focusing on the following parameters:

- Number of estimators: maximum number of boosting trees to use in the model.
- Maximum depth of a tree.
- Learning rate: parameter that controls the step size at which the algorithm makes updates to the model weights.

Additionally, the Walking Forward methodology was applied in the XGBoost forecasting model. This approach involves moving the time series one-time step a time, thus providing a robustness estimation. Nevertheless, it is computational expensive especially with larger amounts of data.

Secondly, all possible combinations of two or more models previously implemented were tried in an ensemble, including the best hyperparameter tuned XGboost. This process involves combining the strengths of both approaches to produce more accurate results. The predictions from each model were combined using the mean to produce a final forecast.

Finally, the optimal model for each product was chosen based on the lowest Root-mean-square error (RMSE) score.

3. Time Series Forecasting for the Test set:

Lastly, a unique ensemble was created which is able to predict future sales for each individual product based on the optimal model chosen in the validation step.

3.4. EVALUATION

In order to not only be able to select the final model, but also to evaluate the results of the time series forecasting for each product, the main measure used was the Root Mean Square Error (RMSE). Simultaneously, the R-square (R^2) and Mean Absolute Percentage Error (MAPE) were analysed.

The RMSE reveals how far predictions fall from measured true values using the Euclidean Distance. Since the errors are squared, this ratio is highly affected by inferior predictions. However, it has the advantage of representing the error in the same unit as the predicted column, thus having easier interpretability. The lower the value, the better the model fit.

The R^2 measures the amount of variance in the predictions explained by the dataset, the closer the value is to 1, the better models' performance. When the value is negative, it means that the predictions tend to be less accurate than the average value of the dataset over time. Nevertheless, it is important to highlight that one limitation of this measure is that adding more variables automatically increases its value.

The MAPE is calculated by taking the mean of the absolute difference between actuals and predicted values divided by the actuals. As it is scale-independent, it can be used to compare the outcome of multiple time series models with different scales. However, one shortcoming is that it will favour models that under-forecast rather than over-forecast.

Regarding the results for each product, the following aspects are important to emphasize:

- GCK_1, GCK_5, and GCK_11 have the highest RMSE values, which is reasonable due to the higher sales amount compared to the other products.
- On the other hand, GCK_20, GCK_9, and GCK_13 have the lowest RMSE values, which is expected since they are on the top 4 products with less total amount of sales.
- All the products apart from GCK_9 and GCK_36 have a MAPE lower than 1%, which is a very good result.
- R^2 are dissatisfactory. Not only GCK_1, GCK_4, GCK_9 have a negative R^2, but also the remaining products have larger scores, which indicates a poorly fitted model.
- Through visual analysis, it is possible to understand that GCK_3, GCK_5, GCK_6, GCK_11, GCK_12, GCK_16, GCK_20 predictions were approximately similar to the real values.

4. RESULTS EVALUATION

To minimize MAE, RMSE, MAPE and MAX error and maximize R-square, stationarity and autocorrelation were examined, feature selection and engineering were enhanced, certain model parameters were adjusted, and additional external datasets were incorporated.

According to the best-performing forecasting models, the following strategies were defined in order to increase the sales margin and increase the transparency towards customers:

- For products 1, 3, 4, 9, 13, 14, 16, and 20, where sales are predicted to remain approximately constant over time, a **co-marketing strategy** could be outlined. To achieve this, **promotions** would be created for products or services of both companies, in order to reach a wider and more diverse audience and benefit from the skills and resources of other companies. In addition, a **loyalty program** could be established that rewards customers who make purchases at the expected frequency for the specific product and encourages brand loyalty, for example, by providing early access to the launch of new products. Additionally, a **transparent communication** strategy could be implemented towards customers. Through the most appropriate media for each product, clear and precise information about their benefits would be provided to customers.
- Regarding products 5, 6, 8, 11, 12, and 36, where sales are predicted to be more or less seasonal over the months, a good approach is to **develop products** during the times of higher sales, for example, by **offering packages** of complementary products or improving quality to further enhance this increase. **Premium options** could also be offered that include the products with the highest sales growth to existing customers to incentivize them to choose higher value products. In months where sales are predicted to decline, **remarketing strategies** could be implemented to capitalize on the initial interest of customers. This could include email marketing campaigns, targeted advertising, or special offers to encourage customers to consider purchasing these products again.

5. DEPLOYMENT AND MAINTENANCE PLANS

After obtaining the optimal forecasting models for each product, it is crucial to deploy and maintain these models in a production environment to ensure that the sales forecasting system is running efficiently and providing accurate predictions. The following sections outline the deployment and maintenance plans for the sales forecasting models.

5.1. DEPLOYMENT PLAN

Integration (Week 1-3): Integrate forecasting models into existing infrastructure via APIs, database connections, or cloud platforms. Data pipeline (Week 3-5): Create an automated pipeline for data ingestion, cleaning, transformation, and storage. Model versioning/monitoring (Week 6): Implement version control, track model changes, and monitor performance using key metrics. UI development (Week 7-9): Create a user-friendly interface with visualization tools and dashboards for easy forecast access. Training/documentation (Week 10-11): Educate sales and data teams on using and interpreting forecasting models for decision-making.

5.2. MAINTENANCE PLAN

Data updates (Monthly): Update data, including exogenous variables and sales data, to keep forecasts current. Model retraining/evaluation (Quarterly): Retrain models using updated data and assess their performance for accuracy. Fine-tuning (6 months/needed): Adjust models based on new data, market conditions, or business requirements for better accuracy. System monitoring/troubleshooting (Ongoing): Monitor models and data pipeline, resolving issues to maintain smooth operation. Stakeholder communication (Ongoing): Gather feedback from stakeholders and use it to improve models and address concerns. By following these deployment and maintenance plans, the company can ensure that the sales forecasting models are effectively integrated into the business processes and continue to provide accurate and timely predictions to support decision-making and sales strategies.

The timeline provides a general outline of the tasks involved, depending on the actual time required for each task may vary.

6. CONCLUSIONS

The Young Talent Consulting Group conducted a comprehensive sales forecasting project for Siemens Business Unit in Germany by analysing historical sales records, macro-economic data, Covid-19 cases, holiday trends as well as other external events. By leveraging techniques such as XGBoost, ARIMA, and Facebook Prophet, the team managed to generate a ten-month sales forecast, helping the company anticipate demand and adjust its supply chain accordingly. The project methodology included data cleaning and pre-processing, including feature engineering and lagged variable creation, time series forecasting, while evaluation was conducted using RMSE, R2, and MAPE measures, with most products having low MAPE and a negative R2. The team provided recommendations and timelines to improve sales margin and increase efficiency. By implementing the provided recommendations and maintaining the predictive models, the Siemens Business Unit expected to optimize its supply chain, leading to overall business growth and success.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

Although the current models have demonstrated effectiveness in sales forecasting, there are opportunities for improvement.

Firstly, **further tuning of the model parameters**, particularly for XGBoost and Prophet, could lead to improved performance. Techniques like grid search, random search, or Bayesian optimization may identify optimal hyperparameters.

Secondly, **exploring alternative machine learning models**, such as Long Short-Term Memory (LSTM), could offer additional insights and enhancements in forecasting accuracy. These models are particularly effective in capturing complex patterns and long-term dependencies in time series data.

Thirdly, **combining multiple models via model ensembling** may yield superior results by leveraging their respective strengths. Experimenting with different ensemble techniques, such as stacking or voting, could enhance the overall forecasting accuracy.

Fourthly, **incorporating additional variables** like competitor information or promotional events may improve the models' ability to predict sales more accurately. However, this will require extra data collection, pre-processing, and feature engineering to ensure compatibility with the existing models. In addition, the treatment of outliers developed in the notebook could be applied, giving later input of these values through statistical models.

Finally, **establishing a system for ongoing monitoring and updating of the models** will ensure they remain effective and relevant. Regular evaluation of model performance, coupled with periodic retraining using new data, will help maintain accuracy and adapt to changes in the sales environment.

7. REFERENCES

Alhamid, M. (2022, May 27). Ensemble Models - Towards Data Science. *Medium*. https://towardsdatascience.com/ensemble-models-5a62d4f4cb0c

Alhazmi, N., Alghamdi, M., Almutairi, N., & Alhazmi, A. (2021). A comparison of the optimized LSTM, XGBOOST, and ARIMA in time series forecasting. IEEE Access, 9, 62047-62057. https://doi.org/10.1109/ACCESS.2021.3075469

Appinventiv. (2022, July 22). *An Analysis of Coronavirus impact on Industries (& Survival Measures)*. https://appinventiv.com/blog/coronavirus-impact-on-industries/

Chatfield, C. (2000). Time-series forecasting. Chapman & Hall/CRC.

Chatfield, C. (2019). The Analysis of Time Series: An Introduction. CRC Press.

Cleveland, R. B. (1990). STL : A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Office Statistics*.

Great Learning. (2021, April 26). RMSE: What does it mean? - Great Learning - Medium. Medium; Medium. https://medium.com/@mygreatlearning/rmse-what-does-it-mean-2d446c0b1d0e

Indeed Editorial Team. (2022). FAQ: Why Is Business Transparency Important? *Indeed Career Guide*. https://www.indeed.com/career-advice/career-development/why-is-business-transparency-important

Jordan, J. (2018). Hyperparameter tuning for machine learning models. *Jeremy Jordan*. https://www.jeremyjordan.me/hyperparameter-tuning/

Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379,2197, https://doi.org/10.1098/rsta.2020.0209

Lingyu Zhang et al (2021). Time series forecast of sales volume based on XGBoost .*Journal of Physics: Conference Series.* doi:10.1088/1742-6596/1873/1/012067

Shumway, R. H., & Stoffer, D. S. (2017). Time series analysis and its applications: With R examples. Springer.

Smart Infrastructure Market Size, Share & Forecast [2021-2028]. (n.d.). https://www.fortunebusinessinsights.com/smart-infrastructure-market-106346

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37-45 https://www.tandfonline.com/doi/abs/10.1080/00031305.2017.1380080

Wang, J., Zhang, Y., & Sun, X. (2021). Machine learning for time series forecasting: A review. *Journal of Manufacturing Systems*, 60, 144-156. https://doi.org/10.1016/j.jmsy.2021.04.004

What is the Trade and Cooperation Agreement? - UK in a changing Europe. (2021, July 20). UK In a Changing Europe. https://ukandeu.ac.uk/the-facts/what-is-the-trade-and-cooperation-agreement/