MAY 2023

Sales Forecasting

Business Cases for Data Science

Lukas Gross
20221363

Tomás Domingos
20221370

Beatriz Carmo
20220685

Tomás Vicente
20221355

Karim Miladi
20220720

Group H

Table of Contents

1. Executive Summary

As any big company that operates in the global market, Siemens deals with many different areas of business and with many products and it realised that, to have a superior performance overall, it is important to be an organized company in every area.

Following this goal, data was collected by the company in order to accomplish monthly sales predictions. These predictions (made using several algorithms depending on the product in question) will help the company be prepared for future changes as they will highlight moments of sales decrease and increase. This will make it easier for the company to take advantage of "good times" and grow while shedding light on more difficult months, allowing the staff to be prepared. These forecasting will also be useful to control inventory and personnel inside the company.

With the analysis made it was possible to touch these points and others that will alert Siemens to more turbulent and also more shiny times in the next 10 months.

2. Business Needs and Requirements

Siemens is a global technology company that operates in a variety of industries and, due to the dynamic of that environment, it can sometimes be challenging to adjust resources, inventory, etc. to that always changing world. And in this kind of company, it is important to have control and to know what to expect when it comes to performance, finances, sales, inventory, resources and so on.

This is when it becomes crucial to have a good management strategy that is able to focus on each one of these matters. It is essential to know how the market will change in order to re-allocate resources and modify practices so the sales can stay inside a comfortable margin and also to get intelligence on how the company is doing overall (if it is performing well or not).

An effective way to evaluate all these factors is by doing a monthly sales forecast. With this, Siemens can have a deeper insight in the company's overall performance, but it can also help developing strategic plans to maintain its success and overcome difficulties that are now a constant in every companies' life since the War in Ukraine.

It is key to have a good budgeting and financial strategy and management as well as supervise how the inventories and sales will behave. And monthly forecasting can help dealing with these matters too. Having an accurate sales forecast is crucial for creating budgets and financial plans and manage inventory levels to ensure that there is enough product to meet the demand levels.

Overall, performing a monthly sales forecast can help a company improve its financial planning, inventory management, performance evaluation, and customer relationships and also identify growth opportunities. It is an essential tool for success, something that all companies desire.

3. Methodology

3.1. Business Understanding and Research[1]

Since Siemens is a very large company and therefore complex, it is important to first understand its dynamics fully. For this, much information on the company was collected, analysed, and

[1] All the information in this section was taken from: Siemens AG AnnualReports.com; available at: https://www.annualreports.com/Company/siemens-ag (Accessed: April 28, 2023)

interpreted including some annual reports. Market intelligence was also analysed as it has a significant impact on companies like Siemens.

After this analysis, some conclusions regarding the functionality of this company were made as well as notes on its yearly performances that will be presented below.

When it comes to channels, the Operating Company serves its customers through a broad variety of them: global product and systems sales organization, distributors, panel builders, original equipment manufacturers (OEM), value added resellers and installers, direct sales through the branch offices of its regional solutions and services units worldwide.

This makes it clear that Siemens has a very broad and diverse customer base. This base ranges through infrastructure developers, construction companies and contractors to owners, operators, and tenants of both public and commercial buildings (like hospitals, campuses, airports, and data centres) without forgetting utilities and power distribution network operators; companies in heavy industries such as oil and gas, mining and chemicals; and companies in discrete manufacturing industries (automotive and machine building).

As expected, this variety makes the demand be just as diverse and volatile so that while customer demand in discrete manufacturing industries is very reactive to macroeconomic cycles, demand in heavy industries and the utilities sector changes more slowly and not as strongly to economic cycles. It is also important to note that smart infrastructure is affected differently by changes in the construction markets, changing only two to four quarters after the actual change in the market.

On top of all this diversity inside these key points, there are also trends that shape companies and force them to adapt. These causal trends can include the rising of the population and of the urbanization that will force a company to produce more and, in this case, will increase the need for safe and sustainable environments with comfortable spaces that have low energy, operating and maintenance costs.

These trends lead to cross-sector coupling, such as electrification of heat and transportation to optimize energy efficiency; an example of that is decarbonization that is changing the energy generation mix towards renewable energy sources, which fluctuate with time of day and weather conditions. Another example is that, in fiscal 2021, Smart Infrastructure acquired C&S Electric Limited (C&S Electric), India, a provider of electrical and electronic equipment for infrastructure, power generation, transmission and distribution to strengthen its position in India as a supplier of low-voltage power distribution and electrical installation technology.

However, not all years are like 2021. Thus, it is also essential to understand how things were in the past and how every year is different. It is crucial to identify the differences between years and what may have initiated them.

3.1.1. Fiscal Year of 2019

During 2019, the order growth was the strongest in the solutions and services business with Europe, Russia and Ex-Soviet Countries, Africa, Middle East and the Americas to be the leading regions when it came to volume of orders. It is also important o note that the revenue rose the most in the Americas region with not only the services mentioned above but also in systems and software.

Revenue growth in the product business was due to low voltage products, while revenue in the other products businesses came in close to prior-year levels due partly to less favourable conditions in short-cycle markets.

In macro economies, the grid markets benefited from the need for intelligent and flexible energy networks and for automation. This was most noticeable in Asia, Australia, and the Americas. Heavy industries and the infrastructure industry also developed favourably during fiscal 2019, thanks to investments in oil and gas markets, data centres and in transportation infrastructure (e-mobility).

Discrete industries, which started strong in 2019, had a downturn in the second half of the year. However, construction markets continued their stable growth during the fiscal year, particularly in the U.S. and China and in the non-residential construction market overall. The growth in building electrification and automation market was driven by demand for building performance and sustainability offerings, including strong demand for energy efficiency and digital services.

3.1.2. Fiscal Year of 2020

During this year, orders in the products business decreased only moderately, despite an adverse market environment for the short-cycle activities at the beginning of fiscal 2020 that rapidly became significantly worse due to COVID-19.

On a geographic basis, the decline in orders and revenue was due to the regions Europe, C.I.S., Africa, Middle East and Asia, Australia, while volume in the Americas remained largely stable, due to different COVID-19 politics that made European markets the most strongly impacted by followed by the markets in the Asia, Australia region, while impacts on the U.S. markets were less severe.

Demand in the markets served by Smart Infrastructure declined moderately in 2024, as expected, due primarily to effects related to COVID-19. The strongest declines in market volume came, however, from the automotive, oil and gas and machine-building industries. Demand declined moderately in the chemicals industry as well.

On the other hand, grid markets remained relatively stable, as utilities continued to prioritize investment in making legacy networks more automated, intelligent, flexible, and reliable. Despite a moderate decline in demand in the construction market overall, the important segment of that market for Smart Infrastructure continued to grow, benefiting from persistent demand for energy efficiency and digital services (once again fuelled by the pandemic that forced companies to turn to online meetings, etc.).

Overall, there were declines in automotive, oil and gas, machine-building, chemicals and legacy construction market industries; while grid markets/utilities and sustainable construction markets were able to remain stable.

3.1.3. Fiscal Year of 2021

This fiscal year was still very much impacted by the COVID-19 pandemic. However, there was some recovery overall as the world learned how to deal with the new virus.

The strongest growth contributions came from the products business. Here was a clear recovery in demand from industrial customers, and from the systems business, which won a number of significant contracts. These included orders from semiconductor manufacturers in the U.S. Orders in the solutions and services business grew slightly as the business saw first signs of recovery in relevant markets towards the end of the fiscal year.

As a consequence of the growth in the products and systems businesses, there was a growth in revenue. And, despite the challenging supply conditions imposed by several COVID-19 politics, Smart Infrastructure maintained its delivery capacity by successfully avoiding major supply chain disruptions.

When it came to geographic analysis, orders and revenue were up in all regions, with double-digit volume growth in Asia (particularly from Chine), Australia and the Americas which included strong demand from residential markets in the U.S.

Industrial markets developed well, with robust growth in the machine building and pharmaceutical industries (once again influences by the pandemic), followed by the automotive, food and beverage, oil and gas and chemicals industries. Grid markets grew clearly as utilities continued to prioritize investments in making legacy networks more automated, intelligent, flexible, and reliable once more.

In general, machine building, pharmaceutical industries, automotive, food and beverage, oil and gas and chemicals industries experienced a significant growth as well as grid markets/Utilities.

3.1.4. Fiscal Year of 2022

With the world now healing from the pandemic effects on social life and economy, orders rose by double-digits in all businesses, led by the electrical products business and the electrification business including a number of larger contracts wins in fiscal year 2022. This growth was highlighted by strong demand from industrial customers.

Revenue rose in all businesses as well. As before this rise was led by the electrical products business, which operated in strong customer markets. Smart Infrastructure successfully avoided major disruptions from challenging supply chain conditions.

On a geographic basis, orders and revenue rose in all reporting regions. The strongest growth contribution came from the Americas region, driven by the U.S., as usual. The growth in the Asia, Australia region was still held back by impacts related to COVID-19 restrictions in China. Both order and revenue development included positive currency translation effects Profit and profitability rose in all businesses, with the strongest increases coming from the electrical products business and the buildings business. The increases were due mainly to higher capacity utilization, pricing measures to offset cost inflation and cost savings.

Market dynamics were influenced by a further recovery from COVID-19 related effects, severe supply chain and logistics constraints, strong price inflation and effects from the new war in Ukraine.

However, all reporting regions contributed to growth. But price inflation affected all regions and came in particularly high in the U.S. In China, growth was held back by lockdown measures, which also impacted growth dynamics in other countries, while Europe was most strongly affected by the war in Ukraine.

Grid markets grew above average with market growth driven by demand for integration of energy from renewable resources. Industrial markets grew nearly as fast as grid markets, driven by growth in the automotive industry among other factors. Growth in the buildings market came in somewhat lower mainly due to weaker growth momentum in commercial building markets.

All in all, buildings market fell bellow expectations while sustainable grid markets and Industrial markets/automotive industries experienced growth.

3.1.5. Fiscal Year of 2023

So far, in fiscal 2023, markets served by Smart Infrastructure are expected to grow slightly slower than in fiscal 2022. While growth in residential and commercial building markets and some industrial markets is expected to slow down somewhat, demand for data centres and power distribution is expected to be strong. Price inflation is also expected to contribute to market growth this year.

Overall, market development in fiscal 2023 is expected to continue to be influenced by supply chain constraints and effects from the war in Ukraine, including on energy prices. Further impacts could arise from potential lockdown measures in China and geopolitical tensions.

There are some key points to be highlighted and monitored during this year:

- Slowdown in Real Estate.
- Nominal gross Market growth.
- Supply chains could be a determining factor.
- War on Ukraine: **How can Smart Infrastructures gain (and lose) from it?**

3.2. Data Understanding and Pre-Processing[2]

For this project, Siemens provided datasets with a lot of information regarding its sales in the last few years and also about the markets. However, this information is not enough to preform a deep analysis of the problem. To overcome this, many other datasets were used for the study[3]:

- World Interest Rates
- World CPI
- World BI
- World FX
- World IIP
- US: Total Construction Spending: Manufacturing in the United States
- Manufacturers' New Orders: Construction Machinery Manufacturing
- Manufacturing BI: Russia
- Manufacturing BI: China
- Manufacturing BI: Brazil
- Manufacturing BI: US
- Manufacturing BI: India
- Manufacturing BI: EU
- Manufacturing BI: UK

All these additional datasets were selected following a brief analysis of the industries and regions that were of higher importance to the problem presented, meaning that the determining industries were:

- Real-Estate
- Sustainable Grid markets
- Industrial markets/automotive industry
- Machine building
- pharmaceutical industries

[2] These two steps were done together as some changes on the data were done as it was being explored.
[3] All the links to these additional datasets are mentioned in the References section.

- automotive
- food and beverage (retail?)
- oil and gas and chemicals industries
- Grid markets/Utilities
- Public sector

While the regions selected were:

- Americas
- Europe
- Russia and ex-Soviet Countries
- Africa
- Middle East (Saudi Arabia)
- India
- China

In order to have data able to produce good models, the datasets were changed: some features were created and missing values were checked.

3.2.1. Sales Data

After collecting all the extra information, it was crucial to search for possible errors and anomalies. This process will be better explained further on. Firstly, it was key to review and explore the main dataset – the one that stored the sales information that was provided by Siemens.

The first step made was to convert the date columns observations into date format that could be then interpreted as so by the algorithms. After this, the original dataset was divided: one data frame for each ***Mapped_GCK*** value. Once this was done, the data frames were merged with the date as index along with each product percentage change over time. This will help on future operations with the sales data.

New variables were also created: every monthly sale for every product existing along with the total sales. Percentage changes were also calculated.

There were no missing values in this dataset. The percentage changes columns had all one missing value but wasn't interpreted as a problem as it was the result of the calculation done of get the values.

This information was checked along with the sales using plots to visualize the variables in an interactive way and be able to better understand the overall situation of the given sales data.

It is important to note that some products had a much higher total sales (**Sales_EUR**) value when compared to others (Figure 1). This was the case of the products **P1, P3** and **P5**. The same does not happen in percentage changes where **P14** was the product with the biggest change in sales followed by **P9** and **P20**.

The next step taken was to check for anomalies in the various products data. For this many graphs were plotted. Not all variables had bizarre values, however some still did, so the future models will be trained using only the regular values and therefore excluding the anomalies. This will result in better models since they won't be trained over odd values.

The anomalies were removed by creating a threshold (which values depended on product to product) and only using the values inside those limits. The products which data was subjected to this process were **P1** and **P3**. There were also some spikes in Total Sales, however, these were not removed since it could jeopardize the realness of the future results. Product **P16** had also an abnormal behaviour as there was a clear downtrend (Figure 2). To avoid training models on peculiar data patterns and singular episodes in time, some of the data regarding **P16** won't be used in the modelling phase.

Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) graphics were also plotted to analyse the behaviour of total sales data as a time series (Figure 3). In both plots, there was a spike in lag 0 followed by a sinusoidal pattern (this pattern may not be specifically present in PACF, however, the values in the plot are still within the boundaries). This suggests that the underlying signal has some periodicity and that its correlation with shifter versions decreases over time. This indicates a trend that can be the trigger for non-stationarity. However, after performing the Augmented Dickey-Fuller test on the data, the conclusions were that the series were stationary[4].

3.2.2. Suppliers Data

The suppliers' data was given by Siemens, and it did not have any missing or odd values (outliers, errors…) just like had happened in the previous datasets.

Several nations, including China, France, Germany, Italy, Japan, Switzerland, the United Kingdom, and the United States, are represented in the data files that are provided. They contain a wealth of historical information on various aspects of production and shipments of machinery and electrical goods in these nations. Additionally, these data files contain crucial details about the global prices of significant commodities, including copper, natural gas, crude oil, base metals, energy, metals, and minerals. Not only that, but the data also includes producer prices for electrical equipment in six different nations which will prove to be useful in the future.

Additionally, various country-specific production indices for electrical and mechanical goods are available. By choosing only the data from October 2017 to April 2022, the data is filtered for a particular time period to create the final data frame.

The only issue with the information inside this data frame was the far too long columns names. To overcome this, after creating additional columns to store the percentage changes, the columns were all renamed with abbreviations of their original names. A dictionary was created to keep track of these changes and also to help in understanding the columns information in case the abbreviation wasn't clear.

3.2.3. Additional Data

3.2.3.1. Market Data

After exploring the sales data, it was time to turn to the market information that was collected. As before, this data had no missing values.

One of the most important measures in economics that is always able to give a good insight in the world's economic situation is the CPI which is commonly used as a measure of inflation. The information on the original CPI dataset was filtered by country so only the countries inside the regions of interest listed above would be analysed.

[4] This test was done in *R Studio* with a p-value of 0.01.

Since this information will be used in prediction models, it was important to test the stationarity of the time series that consisted of this data. Fortunately, after running the Augmented Dickey-Fuller test on the data, the results indicated that it was indeed stationary and therefore of use in time series statistical models.

Since there were too many values in this data set and that would pose a problem when interpreting results, averages were created.

A similar approach was taken when exploring the **interest rates data.** As before, only counties listed above were taken into consideration and a graph was plotted to better understand the evolution of this measure over time (Figure 4). It was clear that the interest rate suffered many changes during the years being the biggest ones in April of 2020 (when there was a big decrease in interest) and in March of 2022 (when the contrary happened). These large changes can be explained by the COVID-19 pandemic as during April it was spreading fast while in March 2022 it was much more controlled, and the markets were returning to their normal functioning status.

After this, the same was done to the **Foreign Exchange data** (Figure 5)**.** As expected, this data was much more volatile and so the euro suffered many changes during the years, having many high and low values when it comes to percentage change. The most recent low values may be explained by the war in Ukraine that has been affecting the whole European region more than other parts of the world.

When it came to **Business Confidence Data,** there was a significant increase in July of 2020 possibly because, with the warm weather of the summer, people were more hopeful that the COVID-19 pandemic would be over soon. The biggest decrease happened in February of 2020, when the shadow of what the new pandemic could bring darkened everyone's hopes. (Figure 6)

3.2.3.2. US Exclusive Data

Once the United States of America hold the place of one of the most developed countries when it comes to economics and production, it was important to analyse its data separately. In this stop, the two datasets used (one with the information regarding Manufacturers' New Orders and the other Total Construction Spending) were joined into one data frame. The biggest and smallest changes in both types of information occurred before the original data time frame, in 2008 during a crisis in the U.S.

None of the mentioned datasets had missing values or outliers.

3.2.3.3. Manufacturing Business Confidence Data

Information on the manufacturing business confidence was also collected for the most important countries: Russia, U.S., Brazil, China, Europe, and India. After collecting the data, the various data frames were concatenated into one. The biggest average change on this data occurred in July of 2020, the first summer of the COVID-19 pandemic (Figure 7).

3.3. Modelling

After having cleaned, pre-processed, and analysed all the data, for it to be fully ready for modelling, all of the possible lag features were created. This will make models more effective and easier to interpret.

On top of this, and to make the modelling phase more pragmatic and efficient, a group of steps to be followed in every product forecasting was created. This was done in tune with the core testing metric, the forecasted RMSE. This process also allowed to have an interactive looping process, where it was possible to try different angles of approach to each product, check the

results and decide on what would be the best at each scenario. The process applied to each product goes as follows:

1. Scale the final dataset, with StandardScaler.
2. Set a bar (threshold) for the RMSE score that each one of the +1000 had gotten, so one could quickly remove features that aren't as helpful or that do not contribute as much to forecasting.
3. Remove lags from the same features that had higher RMSE Scores and, therefore, higher errors.
4. Plot feature selection to remove misleading RMSE score features (sometimes it is needed to trust more on intuition and practical results than in the theory behind the algorithms due to the nature of the problem).
5. From the final list of features, calculate all the possible selection options (features and best models).
6. For each possible selection, apply the grand majority of multivariate time series models available and pick the 5 best ones with the 5 respective best feature selections.
7. Retrain each model and plotting the results to pick the winning model.

This process was applied to all the products present in the dataset, each one at a time. This was done separately because each product has its own behaviour and set of features that affect its sales and therefore will determine its future predictions.

3.3.1. Product P1

After running the feature selection code, the features chosen for this product were ***Average Interest Rates Percentage Change with a lag of 10, the Production Index World: Machinery and Equipment n.e.c. Percentage Change with 1 month lag, and finally the Producer Prices Germany: Electrical Equipment Percentage Change.*** which were all related to the suppliers and additional market data.

Using these features, five different predictions were done with different algorithms. The best testing model was **Test 1** that corresponded to the MLPRegressor, which is a feedforward neural network that can learn complex non-linear relationships in the data, whereas other multivariate time series models, such as ARIMA or VAR, rely on linear or autoregressive assumptions. **Test 1** was the winning model since it was able to catch all of momentum swings, without overfitting on train. This was the model with the lowest RMSE as well.

*Following this. the model for the final prediction was the same used in **Test 1**. It is important to note how P1 has a decrease in sales in November of 2022 (according to the model) that happens right after it reached its peak (October 2022). The sales for this product are believed to stay between 33M and 39M as those are the thresholds observed in the plot (*

Figure 8).

3.3.2. Product P3

For this product, the winning model was not the same as above as it consisted of the Stochastic Gradient Descent (SGD) which is an optimization algorithm used to train a wide range of machine learning models, including regression models. It is different from other multivariate time series models in that it updates the model parameters based on a single training instance at a time, as opposed to batch or mini-batch gradient descent which updates the parameters using the average gradient across multiple instances.

On the other hand, on **Test 2**, the features used were still related to the supplier's data as they were ***India's Interest Rate Percentage Change lag 2, China's Manufacturing Business Confidence Percentage Change lag 11 and Saudi Arabia's Consumer Price Index lag 7.***

In the final prediction for this product, sales are forecasted to narrow down with swings that top at 15M and bottom at 13M (

Figure 9), reaching its high in July 2022.

3.3.3. Product P4

For this product, 12 features were selected. The set was a mix of suppliers' data related features and market data. After plotting the several tests against the target, it was clear that **Test 3** was the winning model since it had a better training prediction, as well as fine prediction test, given that all others didn't perform better.

This model used AdaBoostRegressor that is different from other multivariate time series models in that it uses an ensemble of weak learners that sequentially fit to the residuals of the previous learner, thus making it more robust against outliers and noise in the data.

Similar to what happened to the previous products, P4 also experiences a decrease in sales in the 10 months following the dates in the original data (Figure 10). Uptrend in spikes pre-covid is not forecasted within the next 10 months, instead, a flat trend of swings between low of 400k's and a high 200k's was forecasted.

3.3.4. Product P5

For P5, **Test 4** was the winning model. However, **Test 3** was very similar to **4**, it had a big error on the test range which indicated that **production index Germany: Machinery and equipment n.e.c. %Change** was a determining feature (only difference between the tests).

For this product, the algorithm used in each model was always the same (DecisionTreeRegressor) while the feature selection changed being composed **of Europe Shipments Index Machinery & Electricals Percentage Change, Germany Shipments Index Machinery & Electricals Percentage Change, Production Index Germany: Machinery and Equipment n.e.c. Percentage Change and Germany Production Index Machinery & Electricals Percentage Change.**

No lags were picked by the algorithms for this product, making it the product with the highest sensibility to current events.

In the next 10 months, the behaviour of P5 sales keeps on being very volatile and erratic with many highs and lows. Two swings, spiking at 13M in the first half of the 10 months, were forecasted. In the later stage a big spike in sales of 19M was predicted (Figure 11).

3.3.5. Product P6

In this case, **Test 1** is the winning model since it caches all of momentum swings, is not overfitted on train and has lowest RMSE. However, at a second iteration with other models, **Test 2** was

pricked. This model had a the high RMSE, and solid performance overall, given the intrinsic hard pattern detection of the P6 data.

The feature selection of the model consisted in only two features: **United Kingdom Shipments Index Machinery & Electricals Percentage Change lag 9 and United States Production Index Machinery & Electricals percentage Change lag 6.**

Therefore, the winning predictive multivariate model was the PassiveAggressiveRegressor which is a linear regression model that updates its weights in an online fashion, meaning it can learn from data as it arrives and does not require retraining on the entire dataset, making it more suitable for real-time applications in this multivariate time series analysis.

The forecasting results, suggest that that the sales for this product will keep their up trend that has been historically set, with regular swings. P6 keeps its usual behaviour and consistent behaviour (Figure 12).

3.3.6. Product P8

The feature selection for this product consisted of **Average Consumer Price Index Percentage Change lag 2, India's Business Confidence and Brazil's Interest Rate Percentage Change lag 4.**

The model chosen was the MLPRegressor, which was explained in P1 winning model since it was the one used in **Test 3**. This model learned the uptrend and on test predicted a spike. Although early ranges are bad, the model wasn't trained for it, so it didn't pose a problem.

While forecasting for this product, the results given by the model suggested a similar behaviour to the previous product (P6) as P8 also ha a very regular uptrend and swings (Figure 13).

3.3.7. Product P9

In this case, **Test 1** was the best one overall, catching most of the momentum on all ranges. The algorithm used was the KNeighborsRegressor which is different from all other multivariate time series models in that it is a non-parametric model that predicts the target value based on the average of the K-nearest neighbours, where the distance between the neighbours is calculated based on the Euclidean distance between the predictor variables, and it doesn't consider the temporal relationship between the samples. The features included in this were **World: Price of Natural Gas index Percentage Change lag 2 and Great Britain's Business Confidence Percentage Change.**

Since the KNeighborsRegressor needs a higher number of training data in order to have solid results, the lags were upped on the forecast phase. The results showed a continuation of the downtrend in spikes of sales post Covid, picking just under 15k, and flattening out at the near median range of 5k (Figure 14).

3.3.8. Product P11

The test picked for this product was **Test 3 Error! Reference source not found.**due to marginal a ccuracy increase on the test data, more realistic spikes, and momentum swings. This consisted of the XGBRegressor which is an optimized implementation of gradient boosting that can handle complex datasets with high-dimensional features, making it particularly useful for multivariate time series problems with a large number of covariates.

It is interesting to note that all features share the same lag and are not specifically related to the EU: **Japan Shipments Index Machinery & Electricals Percentage Change lag 6, Japan Production Index Machinery & Electricals Percentage Change lag 6, Production Index World: Machinery**

and Equipment n.e.c. Percentage Change lag 6 and Production Index World Electrical Equipment Percentage Change lag 6.

The forecast showed an uptrend for the spikes, from 2.8M to 3.5M, and bottoms of 1M, 35k, and 670k through the 10 months of prediction (Figure 15). This shows that P11 is keeping its swings tradition and its somehow erratic behaviour in sales.

3.3.9. Product P12

The results presented in **Test 5** seemed to be the best to predict further, even thought, the fit on train isn't the best, it was capable to predict the swings on test. The algorithm of this model was the ExtraTreesRegressor which differs has it is a tree-based ensemble model that can capture complex non-linear relationships in multivariate time series data while avoiding overfitting, by randomly selecting subsets of features and samples during the training process.

The features were related to the U.S. and China (both in electrical equipment): **Producer Prices United States: Electrical Equipment Percentage Change lag 1 and Producer Prices China: Electrical Equipment Percentage Change lag 6.**

During forecast, P12 seemed to stabilize its sales inside a specific range. There are still tight swings in the majority of the 10 months with 350k highs and 250k lows, with a spike downwards in the 9th month hitting a 230k bottom (Figure 16).

3.3.10. Product P13

For product P13, **Test 1** was the winner model since it had the best prediction out of all. However, it is important to make sure to reduce the overfitting tendency of the model itself with the whole train set. The algorithm for this was the ExtraTreesRegressor, the same has product P12.

The features used were very different from each other, not allowing to identify a pattern: **India's Interest Rate Percentage Change lag 3, Production Index United States: Electrical Equipment Percentage Change lag 5 and Saudi Arabia's Consumer Price Index lag 8.**

The forecast results for P13 sales showed two spikes in different directions, with an upwards spike at 45.5k and a downwards spike at 12.5k. However, this product sales seem to be following its usual patterns (Figure 17).

3.3.11. Product P14

For this product, **Test 3** had really well test results and was the winning model – a LinearRegression algorithm which that assumes a linear relationship between the dependent variable and the independent variables.

There was an attempt to clip the Covid spike and the x = 2. After changing the values of df['14'] the spikes on test are predicted successfully on most models.

The features in this model were related to production and also to market data: **Producer Prices Germany: Electrical Equipment Percentage Change lag 5, Production Index United States: Electrical Equipment Percentage Change lag 9, Producer Prices United States: Electrical Equipment Percentage Change lag 12, US's Interest Rates Percentage Change and Average Consumer Price Index lag 12.**

This product sales in the forecasted months, didn't show any spikes bigger than 40k, instead it forecasted swings in the 16 to 3k range (Figure 18). This contrasts with the previous behaviour of P14 sales.

3.3.12. Product P16

For this product, like before, some of the features selected were related to production (**Production Index Germany: Electrical Equipment Percentage Change lag 6 and Producer Prices United Kingdom: Electrical Equipment Percentage Change**). It is also interesting to note that these features were mostly related to the European region: **US's Manufacturing Business Confindence Percentage, Breat Britain's Manufacturing Business Confindence Percentage and Euro area (19 countries)'s Interest Rates Percentage Change lag 9.**

Test 5 was the closest to making a prediction, although this product set to be a hard one to predict correctly. The winning predictive multivariate model was the HuberRegressor a regression model that is less sensitive to anomalies and can handle multivariate time series data, making it a robust choice for time series forecasting tasks. Making sense why it performed better in this setting.

Sales of this product are forecasted to go flat, in the 155 to 150k range, which is not surprising since the existing data showed P16 sales were not as dynamic as the other products' (Figure 19).

3.3.13. Product P20

For P20, **Test 2** was the winner, predicting the big spike at the start of test, something all the other models didn't do as well. The results weren't the best, but to improve them, the data would have to be manipulated.

The wining model was the LinearSVR, a machine learning algorithm used for robust regression that can handle data with prominent level of noise and outliers; and the features were **Producer Prices Germany: Electrical Equipment Percentage Change lag 5, Producer Prices US: Electrical Equipment Percentage Change lag 12, United States Interest Rates Percentage Change and Average Consumer Price Index lag 12.**

For P20, it was forecasted a slow down on big swings, most likely due to exogenous variables which were proven to provide good test results. It is important to note that even thought there are no big swings anymore, there are still some small ones that seem to follow old patterns (Figure 20).

3.3.14. Product P36

For this product, most models seem to not predict well compared to the test data. **Test 3** was the one that predicted the direction swings the best, with the added benefit that it didn't overfit massively on the train data like the others did. On the other hand, **Test 5** was chosen as the winning model as it only missed the downtrend on x = 30, which can be corrected by training the model with the whole data and had an accurate long prediction accuracy.

This model was GradientBoostingRegressor, a supervised machine learning algorithm that uses an ensemble of decision trees to make predictions based on past data and other features.

For this model, the feature selected were **World: Price of Metals & Minerals Percentage Change lag 9**, **World: Price of Copper Percentage Change lag 9, World: Price of Energy Percentage Change lag 4, World: Price of Base Metals Percentage Change lag 9 and China Production Index Machinery & Electricals Percentage Change lag 6.**

Similarly, to the tests, the model has struggled to predict the spikes, predicting only small ones in the range of 10 to 18k (Figure 21).

3.3.15. Total Sales

The same process was applied to the Total Sales. The same algorithm was tested (MLPRegressor) with different feature sets. In the end, the winning model was the one featuring: **Saudi Arabia CPI lag7, CHN BI%Change lag11, US_MNO BI%Change lag9, United States IR%Change lag8, PI_US_MEN_PC lag6.**

3.4. Evaluation

After analysing the forecast results for each product, it became clear that they all have very different behaviours during the predicted 10 months. This indicates that the company might have to focus on each one separately and develop different strategies for each problem.

Lags played a vital role too as many of the changes in the market only have consequences in sales months after. This highlighted the importance of paying attention to those lags in predicting and preparing for the future.

Some products were very sensitive to "live" changes having no lag features. This was the case of **P5** that will be harder to predict because of that. Products influenced by more lag features are easier to predict and allow one to develop strategies with more time since the consequences of the change will only appear later on.

It was also clear that some products were more influence by changes related to the supply data (micro economy) while others were sensitive to macro economy changes. This allowed a separation of the products into three groups:

- Micro-driven products: P3, P5, P8, P11, P12 and P36
- Macro-driven products: P4, P6 and P13
- Ambiguous products (driven by both): P1, P9, P14, P16 and P20

The products inside the Micro-driven group are expected to have a growth in sales of 5% compared to all the data. However, they will suffer a decrease of 16% when compared to the data of the 10 months prior to the prediction.

When it comes to the macro-driven products a 18% growth compared to the prior 10 months is expected while a 1.28% increase in sales was calculated for the post-covid era (10 predicts months vs all time data).

For ambiguous products, there was an increase of 13% in sales when compared to the 10 months prior to the predictions and a 3.29% decrease in the post-covid era, probably boosted by the war on Ukraine and the constant fear of some industries regarding the return of the pandemic (most of the located in China due to their strict COVID-19 politics).

4. Results Evaluation

Most of the models developed during the modelling part of the project offered good and satisfying results during both training and testing phases. It is important to notice that even when the predictions are not as accurate as one would wish, they are all still able to catch and represent the most significant moments and changes in the business. This alone is very positive since it is crucial to have a general idea of these most sensitive episodes in an industry.

Overall, the models and the forecasts made were satisfying since they were able to capture the biggest changes for each product behaviour. Even if some models didn't preform as well, they

still shed a light on the important phases in sales which is extremely helpful when it comes to organizing processes inside a company and inventory.

5. Deployment and Maintenance Plans

Global companies should all have clear and strategic plans keep up with market changes. From what was shown in the results of the various forecasts preformed, Siemens ought to evaluate the market frequently to be able to predict recession times and take the most of positive moments to grow as a company.

Following this and taking into consideration the results obtained, when it comes to product **P1**, the company should be prepared to meet the demand during July and November when sales go up and develop resistance plans for the November where **P1** sales aren't as high. It is also crucial to stay alert to changes in Average Interest Rates, Production Index World: Machinery, Equipment n.e.c. and Producer Prices Germany: Electrical Equipment since these variables affected the predictions and might still affect sales of this product in the future. All these should be done while keeping in mind the month difference between the changes in the macro economy and their result in sales.

The same happens with **P3** that is also highly influenced by macro features that have a lag. Therefore, Siemens should stay alert when it comes to the U.S. and Germany Production Index in Machinery & Electricals as well as Electrical Equipment. The world price of copper and the U.S. producer prices for electrical gear is believed to have influence in **P3** sales as well and should be looked at with awareness.

It is also important to focus on the demand this product is expected to have in the months its sales peaked. This can, of course, be a result of the increase in the material prices, but its nonetheless crucial to have a plan to keep up with all those adjustments. For example, focusing more on **P3** during November instead of working in **P1** could be a good strategy as **P3** is expected to generate more revenue compared to the low sales of **P1**.

For **P4**, one should keep an eye on India, China and Saudi Arabia's market as features related to these regions are very influential. This product sales seem to be more related with the interest and spending power in these countries.

It appears that in June there will be a rise in sales which in this case can indicate a rise in demand. Siemens should therefore be prepared to meet it without compromising the production of other products. In the future it is also important to keep an eye on these countries financial situation to be able to predict changes in sales (be them good or bad, but especially if they are bad).

It would be smart to organize production for both **P4** and **P6** in June since both have a peak in sales during that month. However, in contrast with the latest product, **P6** presents many peaks, having two in September and December. These up trends seem to follow a pattern that should not be looked at without care.

These rises in sales can be explained by the interest and purchase power of some countries, just like it happened with **P4**. But, in this case, the country present is Brazil along with average interest rates and average consumer price index. Product **P6** sales are deeply related with interest rates so, assuming they will stay faithful to that relationship, it is key to stay alert to those metrics in the future.

The next product has a similar behaviour to **P6**. **P8** also has many peaks during the 10 months predicted and is also related to Brazil and interest rates along with buying power. In this case, India's business confidence also plays a role, so the economy of both these countries should be looked at carefully so one can be prepared for changes in the market and in sales, focusing on meeting the demand in January, October, and June (when the sales reach their peaks during those 10 months).

Product **P9** is one to be careful with as its sales seem to be suffering a down trend. This can be related to the Price of Natural Gas index since with the War in Ukraine its value has rose and sometimes it became nearly impossible to get gas at an admittable price.

Great Britain's Business Confidence is also something to consider. The UK has suffered a lot sometimes since their departure from the EU and the country is still experiencing some problems because of that and now with the war as well.

If the sales of **P9** follow the observed trend, it would be interesting to start developing a resilience plan to make more profit from this product: making it more expensive or making it more attractive to customers so they tend to buy it more.

P11 is a product which sales seem to be dependent on Japan's shipments and production indexes as well as production indexes for electrical equipment and machinery around the world. It is expected for it to have a big decrease in October. To avoid losing profit and fall into debt because of this, it is important to develop a strategy ahead that is able to overcome this difficult month. One idea would be just carrying on and hope that the increases in sales in the months prior and after can make up for that, however that is not always a good strategy as it relies on the future and the future is highly unpredictable.

To maintain the sales of this product under control one should also keep the indexes mentioned above under control so that it is easier to predict and avoid damages.

Product **P12** seems to be deeply related with electrical equipment production prices. Following this, it is key to always know if these indexes changed so, one could be ready for the consequences in sales. Overall, the forecast for this product showed that its sales will keep inside a certain margin, however, it is possible that this changes in future due to the volatility of the economy at the moment, prompt by the war and also the political conflicts between various countries including China and the U.S. (the ones which prices are used in forecasting **P12).**

P13 is another product which sales are influenced by non-European countries for its sales are related to India's Interest Rate, Production Index United States: Electrical Equipment and Saudi Arabia's Consumer Price Index. It also interesting to note that Saudi Arabia is the country that seems to influence these sales the longest after it changes (having a lag of 8 months).

Overall, one should have prepared a plan for the decrease in sales during October and another one for the possible demand in August where the sales experience a rise.

One however, doesn't have to worry about the lags when it comes to the relationship between product **P14** and US interest rates as they have an immediate impact on sales. This can explain why in the 10 months predicted there was no peak as high as the ones experienced before – with the war, everyone around the world has felt their purchase power getting smaller and with that become less confident in the world's economy overall.

The sales of **P16** are not as dynamic as the other products. However, it seems to be a rise in sales which can be explained by a rise in Producer Prices United Kingdom: Electrical Equipment Manufacturing Business confidence and US's Manufacturing Business confidence which would have an immediate impact since there is no lag. The Interest rates inside the euro area are also important for **P16** sales so it would be interesting to stay alert to them so one can know how sales will change 9 months later.

As it happened with other products, **P20** experiences a low trend in sales in the 10 months predicted never recovering from it as it does not reach the previous peaks. The same happens with **P36** even thought they don't have explaining features in common.

Takeaways from the Total Sales best train and test combinations, for the train time range, the test forecast was predicted to have a swings with local tops of 1.7 and local bottoms in the -0.19 area, similar behaviour, although intended, to the train set. In terms of the feature combinations, a lot of the features used were given by the Siemens team. With both points considered, they lead one to believe that those data points help to forecast, with their respective lag, in normal market conditions.

On the other hand, the test dataset proves to be a Post-Covid, Lockdown Reopening, Supply Chain Shocks and Banking Liquidity Injections cocktail and a resulting Inflation hangover.

Why?

- Macro datapoints that we loaded to the model proved to be much better indicators to forecast the last 10 months of Sales%Change.
- The regular swings forecasted by the best forecasting features in the train dataset disappeared.
- Inflation rises, Rate hikes, changes in manufacturing business confidence, we're the main top features.

For the intended 10 Month forecast, it is hard to believe market conditions have come back to normal so it is best to assume that the same indicators that were important to forecast the last 10 months will be the same for the next 10.

It is also important to stay alert to changes in the market to organize the company in order to thrive when their consequences hit. One must not forget to do forecast regularly as well. Forecast help with any kind of management inside a company: inventory, personnel, funds, resources, etc. But, most importantly, it is key to have the capacity to overcome recession times and overall bad for business moments and forecasting can be a staggering tool to do so.

6. Conclusions

At the end of the analysis, it became clear how different each product is from one another and how they are affected by distinct factors.

Forecasting proved to be a challenging task as the environment in which one is forecasting is always changing and sometimes it becomes impossible for a model to predict sales or any other metric correctly. This happens because many times models learn specific patters in the training set that are not repeated in the future due to explicit conditions present at the time of training.

Fortunately, and although the data available was very much marked by the COVID-19 pandemic, the models used were able to provide good enough results that will be helpful identifying when the company needs to take a leap of faith and grow and when it should lay low.

Another lesson, for forecasting, is that one shouldn't be anecdotal when selecting features and forecasting to the Smart Infrastructure departments. Hindsight will always be 2020 of course, but if a team relied on what previously worked at 2022-04, they would forecast wrong data that would put Siemens in further jeopardy (assuming COVID slowed down their operations). The key here is to forecast, not only with the key features of the past, but also having in mind what could be the important feature for the next months. In this case the Macros were important.

Overall, it became apparent that sometimes in forecasting, the most important thing is not to get as precise results as possible but a general view of the future that helps develop strategies to overcome hard times and thrive in a highly competitive market where it's not always easy too keep up with the demands.

7. References

- Business tendency surveys for manufacturing: Confidence indicators: Composite indicators: OECD indicator for Brazil (2023) FRED. Available at: https://fred.stlouisfed.org/series/BSCICP03BRM665S (Accessed: April 24, 2023).
- Business tendency surveys for manufacturing: Confidence indicators: Composite indicators: OECD indicator for China (people's republic of) (2023) FRED. Available at: https://fred.stlouisfed.org/series/BSCICP03CNM665S (Accessed: April 24, 2023).
- Business tendency surveys for manufacturing: Confidence indicators: Composite indicators: OECD Indicator for India (2023) FRED. Available at: https://fred.stlouisfed.org/series/BSCICP03INM665S (Accessed: April 24, 2023).
- Business tendency surveys for manufacturing: Confidence indicators: Composite indicators: OECD indicator for the Russian Federation (2023) FRED. Available at: https://fred.stlouisfed.org/series/BSCICP03RUM665 S (Accessed: April 24, 2023).
- Business tendency surveys for manufacturing: Confidence indicators: Composite indicators: OECD indicator for the United States (2023) FRED. Available at: https://fred.stlouisfed.org/series/BSCICP03USM665S (Accessed: April 24, 2023).
- Manufacturers' new orders: Construction machinery manufacturing (2023) FRED. Available at: https://fred.stlouisfed.org/series/U33CNO (Accessed: April 24, 2023).
- Oecd, OECD statistics, OECD Statistics. Available at: https://stats.oecd.org/ (Accessed: April 24, 2023).
- Siemens AG AnnualReports.com. Available at: https://www.annualreports.com/Company/siemens-ag (Accessed: April 28, 2023).
- Total construction spending: Manufacturing in the United States (2023) FRED. Available at: https://fred.stlouisfed.org/series/TLMFGCONS (Accessed: April 24, 2023).
- UNIDO statistics data portal UNIDO Statistics Data Portal. Available at: https://stat.unido.org/ (Accessed: April 24, 2023).
- Business tendency surveys for manufacturing: Confidence indicators: Composite indicators: OECD indicator for the Euro Area (19 countries) (2023) FRED. Available at: https://fred.stlouisfed.org/series/BSCICP03EZM665S (Accessed: April 24, 2023).
- Business tendency surveys for manufacturing: Confidence indicators: Composite indicators: OECD indicator for the United Kingdom (2023) FRED. Available at: https://fred.stlouisfed.org/series/BSCICP03GBM665S (Accessed: April 24, 2023).
- Joaquin Amat Rodrigo and Javier Escobar Ortiz (no date) multi-time series forecaster, Multi-time series forecasting- Skforecast Docs. Available at: https://skforecast.org/0.5.0/user_guides/multi-time-series-forecasting.html# (Accessed: 03 May 2023).

8. Appendix

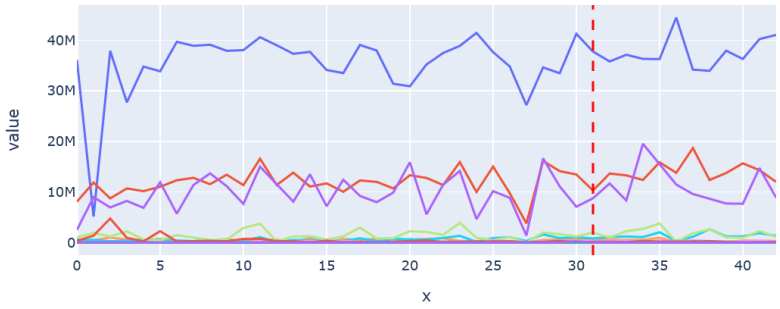

Figure 1 - Products value over time. P1 - first blue line, P3 - first red line, P5 - first purple line (from top to bottom)

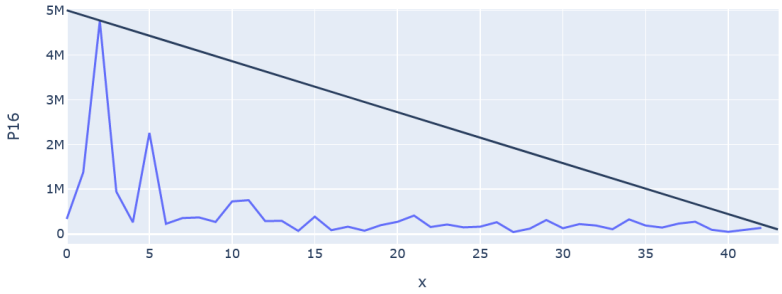

Figure 2 - Clear downtrend in product P16

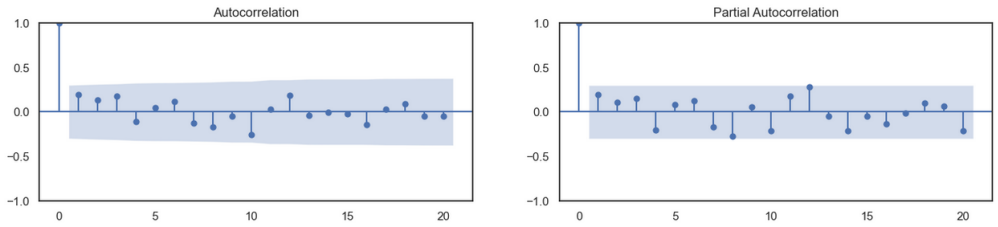

Figure 3 - AC and PAC plots of total sales

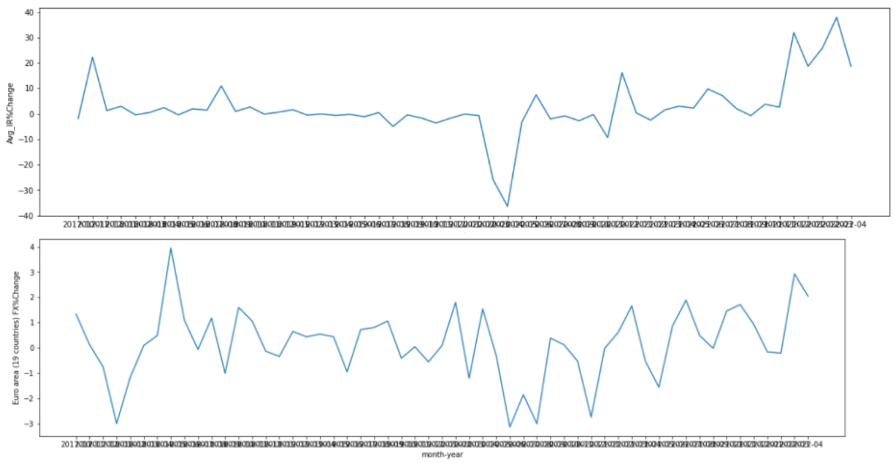

Figure 4 - Average Interest Rate change (%)

Figure 5 – Average Euro FX change (%)

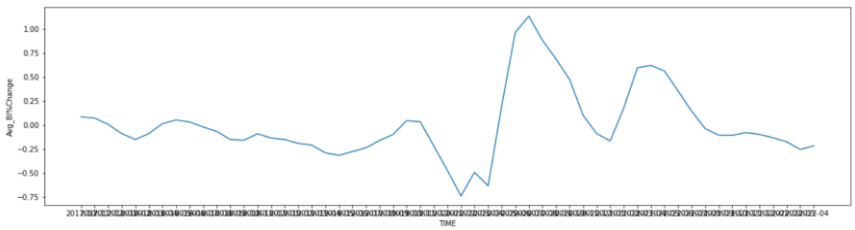

Figure 6 - Average Business Confidence change (%)

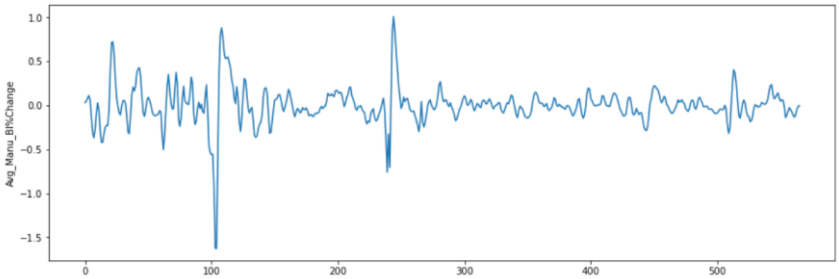

Figure 7 - Average manufacturing BI change (%)

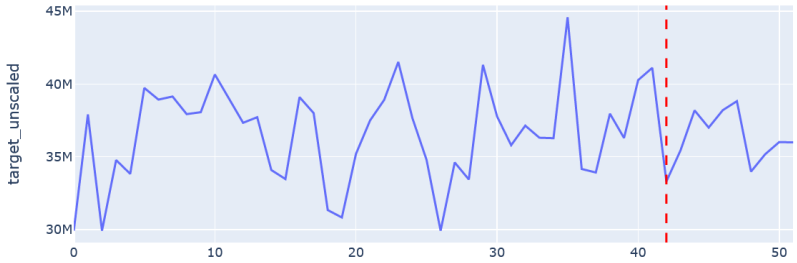

Figure 8 – Forecast for P1

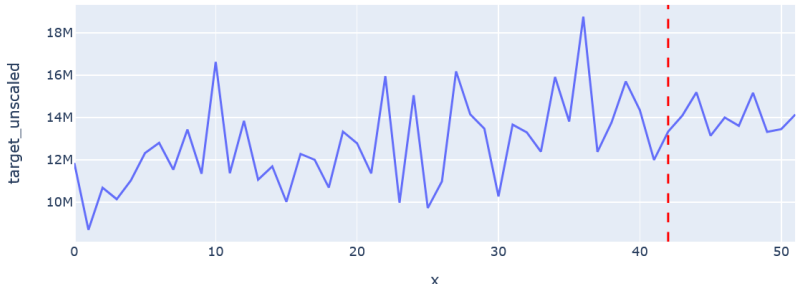

Figure 9 - Forecast for P3

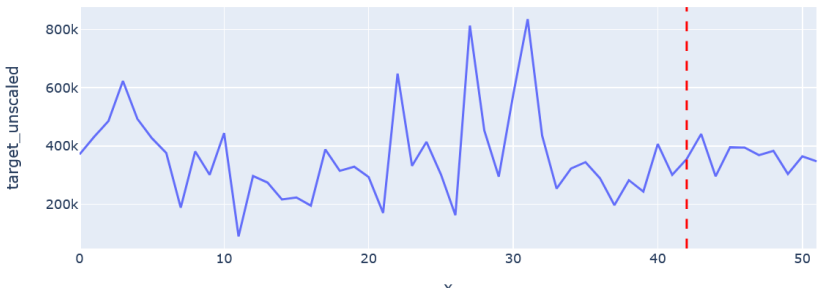

Figure 10 - Forecast for P4

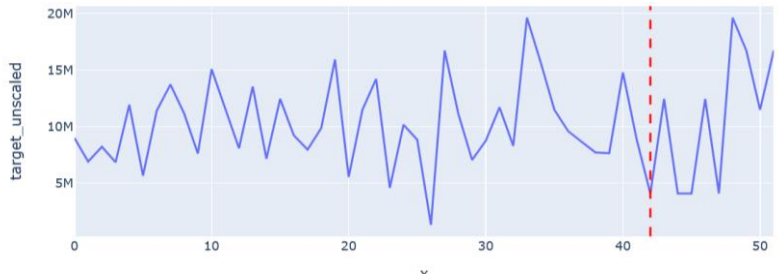

Figure 11 - Predictions for P5.

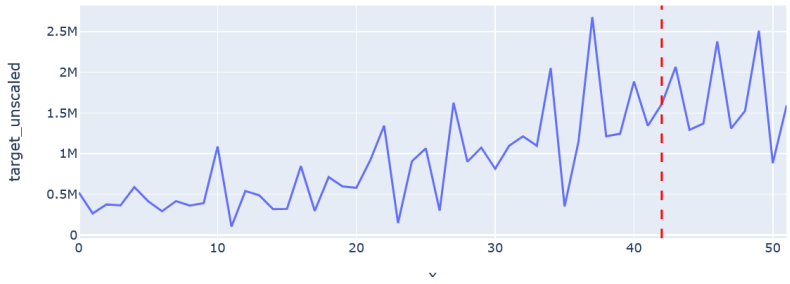

Figure 12 - Forecast for P6.

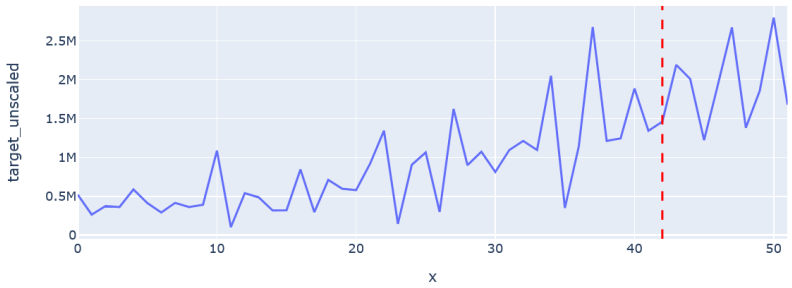

Figure 13 - Forecast for P8.

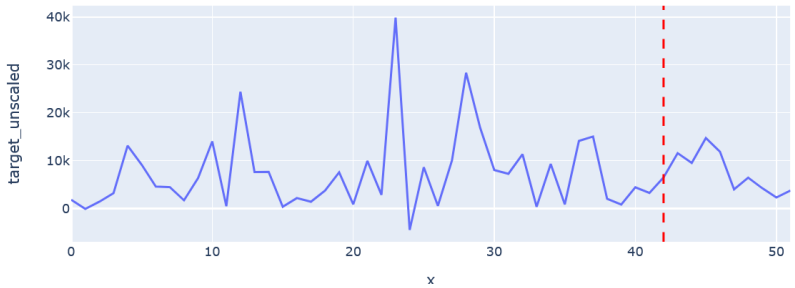

Figure 14 - Forecast for P9.

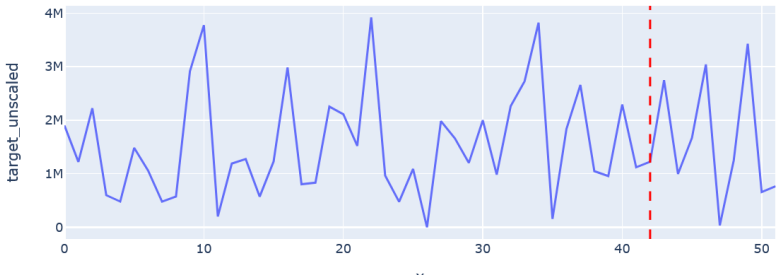

Figure 15 - Forecast for P11.

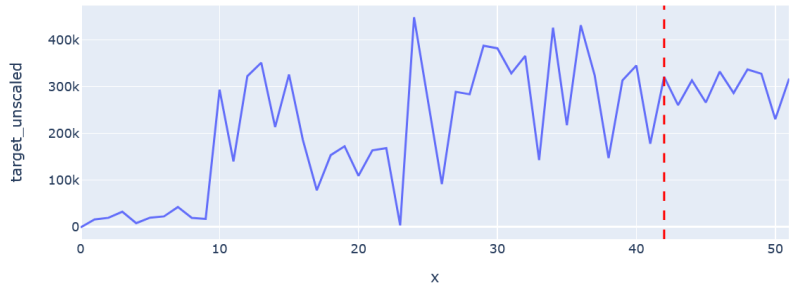

Figure 16 - Forecast for P12

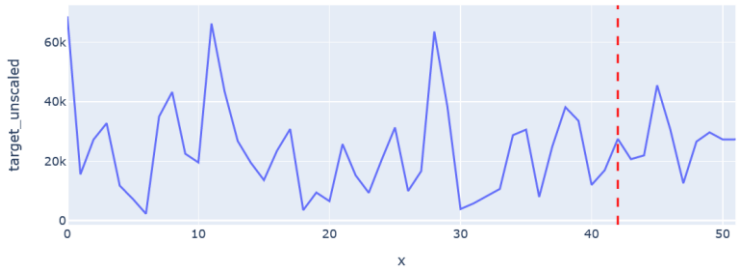

Figure 17 - Forecast for P13.

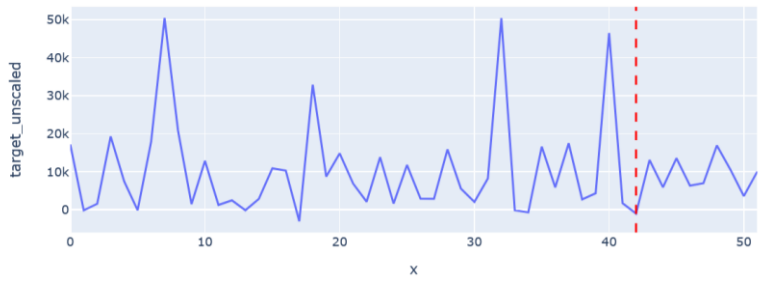

Figure 18 - Forecast for P14.

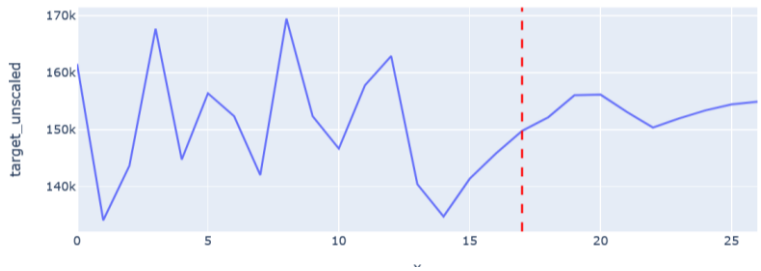

Figure 19 - Forecast for P16.

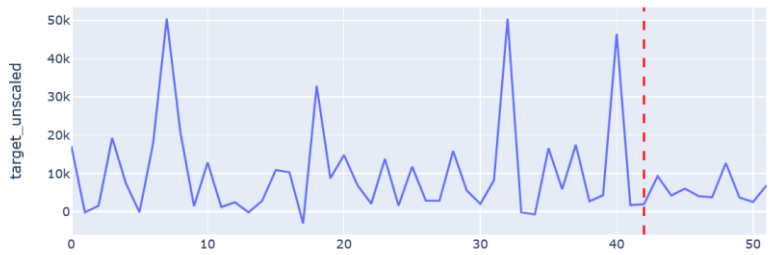

Figure 20 - Forecast for P20.

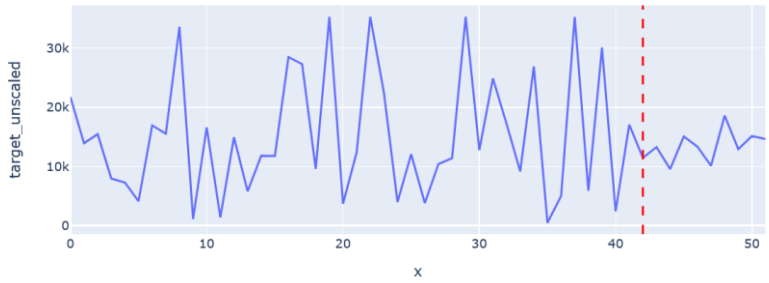

Figure 21 - Forecast for P36.