MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 3: Sales Forecast

Ana Miguel Monteiro, number: 20221645

Ana Rita Viseu, number: 20220703

Miguel Cruz, number: 20221391

Rodrigo Brigham, number: 20221607

Sara Galguinho, number: 20220682

Group G

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

May, 2023

Index

1. EXECUTIVE SUMMARY 1
2. BUSINESS NEEDS AND REQUIRED OUTCOME 2
2.1. Background 2
2.2. Business Objectives 2
2.3. Business Success Criteria 3
2.4. Situation Assessment 3
2.5. Determine Data Mining Goals 5
3. METHODOLOGY 6
3.1. Data Understanding 6
3.2. Data Preparation 8
3.2.1. Data Imputation 8
3.2.2. Feature Creation 9
3.2.3. Sales Analysis 9
3.2.4. Feature Selection 10
3.2.5. Outlier Detection 10
3.3. Modeling 10
3.4. Evaluation 11
4. RESULTS EVALUATION 12
5. DEPLOYMENT AND MAINTENANCE PLANS 12
6. CONCLUSION 13
6.1. Considerations for Model Improvement 13
7. REFERENCES 14
8. APPENDIX 16

1. EXECUTIVE SUMMARY

This project aims to assist Siemens Advanta in developing accurate and reliable forecasts for Siemens’ Smart Infrastructure Division. By accurately predicting future sales, the company can make informed business decisions and optimize its production, inventory management, and marketing strategies, which can ultimately lead to increased revenue and profitability. With access to historical sales data from October 2018 to April 2022, we utilized time series models and machine learning algorithms to accurately forecast sales for fourteen different product groups for the following ten months.

During this project, we followed the process model CRISP-DM, starting by gaining a broader understanding of the business. To do this, we explored Siemens’s background, clarified business objectives, and defined success criteria for the problem, which will allow us to measure the performance of our findings and predictions. We also assessed the kind of resources that were available, as well as risks/contingencies and costs/benefits that this project entailed. As a final step, before starting to work on the data provided by Siemens Advanta, we determined what were our data mining goals - objectives in technical terms - for this forecasting project. Afterwards, we proceeded with data preparation and exploration. This involved cleaning and organizing the data, as well as identifying and addressing any missing values or outliers. We also conducted an exploratory data analysis to gain insights into the distribution of the variables, alongside pertinent correlations between them, which helped us identify relevant features for our models. Finally, by applying an ensemble model composed by Means, Medians, XG Boost, Auto-ARIMA and Prophet, we developed a forecasting model that effectively predicts sales based on historical data.

Based on our evaluation, we found that XG Boost and Auto-ARIMA were usually the most accurate models for each product, outperforming the other models in terms of accuracy, with lower values of Root Mean Squared Errors. We used our ensemble model to forecast Siemens' sales for the next ten months.

In summary, this sales forecast model project can provide valuable insights to Siemens by predicting future sales for each of its 14 product groups. By following best practices in data analysis and modeling, we developed a robust and reliable forecast that can inform Siemens' strategic decision-making processes and drive business growth.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. BACKGROUND

Founded in 1847 by Werner Von Siemens, Siemens AG is the largest manufacturing company in Europe, being focused on industry, infrastructure, transport, and healthcare. Among its divisions, there is Siemens Advanta, which is a global consulting and professional service group with a strong focus on digitalization that delivers end-to-end consulting and solutions. Siemens Advanta has significantly shaped the successful story of Siemens, being the consulting partner of choice for its own digital transformation. With headquarters in Munich, Germany, and offices in America, Asia, Europe, and the Middle East, the group develops more than 700 projects globally every year and has optimized more than 100 factories in the last two years. Overall, Siemens Advanta is an important player in the digital transformation landscape, offering a comprehensive suite of solutions to businesses looking to leverage the power of digital technology to drive growth and success.

Siemens's Smart Infrastructure Division is facing challenges related to the increase of sales margins and the lack of consistency and transparency about quotes to customers. By basing itself on the history of sales offers, Siemens Advanta hopes to provide an automated price recommendation via machine learning.

Sales forecasting is an essential aspect of any business strategy since it helps organizations to predict future demand for their products or services. Accurately forecasting sales allows businesses to make informed decisions on everything from production planning to pricing and marketing. By anticipating sales trends, companies can manage their inventory to meet the expected demand.

Thus, by working closely with the Siemens's team/stakeholders, our goal is to leverage historical data provided to us and apply forecasting techniques and machine learning in order to accurately forecast sales. Consequently, Siemens Advanta will be able to make more informed price recommendations to the Smart Infrastructure Division, ultimately improving sales margin, as well as enhance customer transparency.

2.2. BUSINESS OBJECTIVES

Our primary goal for this project is to accurately predict sales by understanding their behavior over time. Through forecasting techniques, we aim to assist Siemens and seek to address the following key business questions:

- Identify the main differences in sales over multiple months and years;
- Identify the main differences between the distinct groups of products;
- Understand how to improve the sales margin and enhance the transparency of Siemens's Smart Infrastructure Division, as well as present recommendations aligned with the company's objectives.

Some of the expected benefits of this project include:

- Improved accuracy in sales forecasting. By using advanced forecasting and machine learning techniques, Siemens can gain a better understanding of how its products are likely to sell in the future;

- Increase of sales margin;
- More effective pricing recommendations;
- Transparency enhancement;
- Better and more informed decision-making: the company can make well informed business decisions related to production, inventory management, marketing strategies and so on;
- Improved customer service: by understanding which products are likely to be in demand at certain times of the year, Siemens can better meet the needs of its customers by ensuring that it has the right products in stock when needed. This can lead to increased customer satisfaction and loyalty.

2.3. Business Success Criteria

In order to evaluate the success of the project, it is necessary to define concrete criteria that enable us to evaluate its progress and outcomes. Therefore, we established the following business success criteria to be assessed by the relevant stakeholders:

- Increase in revenue and profit margins by one percentual point within a year;
- Improvement of operational efficiency by:
    a) increase of inventory turnover ratio by 2%;
    b) reduction of production cycle times by 10%;
    c) reduction of delivery lead times by 3%;
- Improvement of decision making and price recommendations.

2.4. Situation Assessment

To develop this project, we used the file “Case3_Sales data.csv”, a dataset with 3 columns and 9802 rows, containing the date of each transaction and the amount corresponding to it, as well as the category of the product that was sold. The sales range from 2018 to 2022. We also used the file “Case3_Market data.xlsx”, a dataset with 48 columns and 222 rows, which includes various important macroeconomic indices for Siemens, some of them corresponding to specific countries, as well as different dates for each of them. This data goes from 2004 to 2022. Both files were provided by Siemens. The changes made to the initial datasets are explained in chapter 3.1.

To prepare the data for analysis, we resorted to software tools such as Python and some of its libraries: *Pandas*, *NumPy*, *Scikit-learn*, *ydata_profiling*, *SciPy*, *datetime*, *calendar* and *re*. For data visualization, which is essential to communicate insights and present results from our exploratory analysis, we also used libraries like *Seaborn* and *Matplotlib*. Finally, for the modeling phase of the project, we used various libraries such as *prophet*, *xgboost*, *adfuller* and *pdmarima, statsmodels*.

Besides the given datasets, we added new sources of information as we believed it would be an important resource to get a better understanding of the data and to achieve greater results. We imported two new datasets:

- The file “GDP.csv” is a dataset with information regarding the quarterly Gross Domestic Product of Germany for the years 2004 to 2022 (not including the last trimester of 2022). The dataset initially consisted of two features: *DATE* and *GDP*. The datatype of *DATE* was changed to datetime type and the column was set as the index. Furthermore, the data was resampled to monthly frequency and we forward filled the missing values (meaning we propagated the last valid observation forward). We performed this resampling assuming that the quarterly GDP value corresponds to the sum of the GDP values of the following 3 months (so the monthly GDP value is 1/3 of the quarterly GDP). With the final transformation, we have a dataset with 226 rows, where the index is the *DATE* and the one column is the monthly GDP (which was later added to the Market dataset).

- The file "Cons_price_index.csv" is a dataset regarding information about consumer price index in Germany per month from January 2004 to March 2023, where the value for the year 2020 is set on 100 (meaning the average of the 12 months in 2020 is 100). After some necessary transformations, we ended up with a dataset of 231 rows, which is made of 2 columns: *Date* and *Consumer Price Index (2020=100)*.

Furthermore, as recommended by Siemens Advanta's team, the non-working days (weekends and holidays) could also be relevant information, so we added the number of Saturdays, Sundays and official holidays of Germany that took place in each month. Their integration in the data can be seen in section 3.2.2.

Besides resource assessment, it's important to mention the assumptions that we made about the provided datasets:

- Negative values of sales mean that there were more credits (returns/refunds) than sales;
- Null values of sales indicate products that were not sold;
- The macroeconomic index "Producer Prices Electrical equipment" refers to the cost prices of "generic" electrical equipment and could be used as a proxy for measuring the cost of production.

During the realization of the project, some challenges may arise due to limitations in resources, such as computing power and time. It is also important to consider potential risks, as well as take measures to minimize their impact:

- Data quality issues: The data might contain missing values, incorrect values or inconsistencies that can affect our analysis. To minimize this problem, we resorted to data cleaning techniques, ensuring that the final data was clean, complete, and accurate.
- Unrepresentative samples: The provided sales dataset might not be representative of the entire population, since it only represents a few years of sales. To help this potential problem, it is good to make sure the data is diverse and representative of the sales history.
- Problems with the implementation of some models: To mitigate this risk, we made sure our data was carefully prepared for the desired models to be implemented without problems.
- Lack of domain knowledge: Without a good understanding of the industry and the specific context of the business, we could miss important patterns and make incorrect assumptions when forecasting future sales. To address this issue, we conducted a thorough additional research to build our domain knowledge and gain a better understanding of the market trends and factors that could impact the accuracy of our predictions. We also consulted with industry experts and leveraged external data sources to supplement our analysis and ensure that our models are grounded in a comprehensive understanding of the business.
- The insights and recommendations generated might not be accepted or adopted by the team: To ensure this does not happen, our findings should be communicated with clear and

detailed explanations in a way that is easily understandable and actionable for Siemens's sales representatives and stakeholders.

Finally, this project may involve several direct and indirect costs, such as acquiring and processing the data, computing resources, hiring, or training team members of Siemens, and redefining and implementing strategies. On the other hand, there are also potential benefits that Siemens can expect, like increasing its sales margin, improving its operational efficiency, generating an automated price recommendation, and enhancing transparency. Additionally, the insights gained can help the team to better understand the market trends of the industry it operates on.

2.5. Determine Data Mining Goals

With the aim of evaluating our sales forecast, it is important to determine the data mining results we are expecting, meaning what are the technical objectives of this project. Considering the already mentioned business objectives, our general goal, and the available information provided, we have defined the following data mining goals:

- Low Root Mean Square Error (RMSE): Our goal is to achieve an RMSE score that is as low as possible, indicating that our forecasts are accurate and reliable. This metric shows how far predictions fall from observed true values using the Euclidean distance. RMSE is a widely used metric and it is regarded as one of the best general purpose error metric (Noah, 2022).
- Identify the most important predictors for the different product groups: Our objective is to understand how different macroeconomic indices and other possible exogenous factors affect the existing product groups, as that allows for a more detailed analysis and a better understanding of how different products are affected by external factors.
- Perform forecasting for each product group: We hope to create separate forecasting models for each product group and select the most accurate ones. This will allow us to capture the unique sales patterns and behaviors of the different types of products, resulting in more accurate and tailored forecasts.
- Use multiple forms of performance evaluation: Although the most important metric that we will use is the RMSE, our goal is to use other forms of performance evaluation. There is no one size fits all, so it is better to use a range of metrics to evaluate our models (Howell, 2022).
- Avoid overfitting: The generalization of our models to new data is ultimately what allows us to use them to make predictions (IBM, n.d.), therefore, it is important that the models we train don't fit too closely to the training set, allowing them to perform accurately against unseen data.
- Test various forecasting methods: We wish to evaluate the performance of multiple forecasting methods, such as time-series models and machine learning algorithms, to identify the most effective ones for the case in question. Capturing the variability of the economy and the demand is crucial (Komor, Bony, 2022), so we will take advantage of the several methods available.
- Optimize the parameters of the selected models: Once we have selected the most effective model for each product group, our goal is to fine-tune their parameters for performance optimization. This will involve testing different combinations of parameters and evaluating the resulting performance metrics to find the optimal settings.

- Build a robust and scalable forecasting system: Outside the scope of this project, our goal is to build a forecasting system that can handle large volumes of data and changing market conditions, as well as the integration of new data in the future. Scalability is about handling huge amounts of data and performing a lot of computations in a cost-effective and time-saving way, and it includes benefits like productivity, portability, and cost reduction (Kansal, 2020).

3. METHODOLOGY

To present a solution for the proposed problem, we followed the CRISP-DM methodology, consisting of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The first phase (developed previously in section 2) involves defining the problem and project objectives. In the second phase, data is collected and analysed to gain a better understanding of the available information. The third phase involves cleaning, transforming, and integrating the data to prepare it for modeling. In the modeling phase, various techniques can be used to build and test models. The fifth phase is evaluation, which involves assessing the quality and effectiveness of the respective models. Finally, in the deployment phase, the chosen model is put into action and monitored for performance. It is important to note that the sequence of these phases is not rigid, and we moved back and forth between steps when necessary.

3.1. DATA UNDERSTANDING

As mentioned before, we uploaded new information to get a better understanding of the business, which culminated in having two extra datasets. After these additions, we began to perform an initial exploration of the available information and to try to identify data patterns, structural problems and draw early conclusions. After uploading the provided data files into *pandas* data frames, we started by becoming familiar with our data and identifying the existing features.

We conducted various transformations on the Sales dataset to improve its usability and accuracy. The first step was to remove the "#" symbol from the *Mapped_GCK* column, followed by replacing "," with "." in the *Sales_EUR* column, enabling us to convert its datatype to float. We also changed the datatype of the *Mapped_GCK* column to integer for consistency. The *DATE* column was converted to a datetime type and transformed to “YYYY-MM-DD” format for easier manipulation. Additionally, we renamed the columns for better readability. Next, we created the feature *DateMY*, corresponding only to the month and year of each date. Moreover, we aggregated the data by month and organized it per product group. After that, we added the column *Consumer Price Index (2020=100)* (from “Cons_price_index.csv”). Finally, a new column, *Adjusted Sales (€)*, was created, which multiplies the sales value by the consumer price index. From now on, the adjusted sales will always be used, and, in the end, they will be readjusted to get the accurate predictions. The features of the final datasets are as follows in Table 1 and Table 2.

FEATURES	DESCRIPTION
DateMY	Month and year of when a sale was made ("YYYY-MM" format)
GCK	GCK stands for "Global Category Key" and represents a unique identifier for the category of products that were sold
Adjusted Sales (€)	Total amount of revenue (in euros) generated by the sale

Table 1 – Features/Columns in the *Sales* dataset

MARKET DATASET	
Date	Index of the dataset indicating the month and year ("YYYY-MM-01" format)
Macroeconomic Indices	Each column is a macroeconomic index worldwide or in a specific country (there is a total of 47 indices)
GDP Germany	Monthly GDP of Germany

Table 2 – Description of the *Market* dataset

Afterwards, we used descriptive statistics and data visualization tools (like pandas profiling) to generate an overall report of the data and discover inconsistencies and potential insights:

- The variable *GCK* is a unique identifier for each product group. In total, there are 14 different groups;
- The Sales data is from October 2018 to April 2022;
- There are negative values in sales (more returns/refunds than sales);
- From the pandas profile, we can see that 72.8% of the column *Sales (€)* are values of 0. As mentioned before, we assumed these are products that were not sold;
- The Market data is from January 2004 to October 2022.

For a more detailed data understanding, we separated our sales data by product group and performed an exploration of the subsequent fourteen time-series. For each, we plotted the time-series, a lag 1 plot (a scatter plot showing the potential relation between y(t) and y(t+1)), the autocorrelation function with a 95% confidence interval, and the partial autocorrelation function with a 95% confidence interval as well. We also performed the Augmented Dickey-Fuller test with a 5% significance level test on each time series in order to gather some statistical evidence on whether the series is stationary or non-stationary.

These were some of the conclusions we were able to extract from our analysis, given the previously set confidence levels. In Figure 8.6 and Figure 8.7, examples of the ACF and PACF plots can be seen:

- Product Group 1: There is no statistical evidence of a lagged effect from either the autocorrelation or the partial autocorrelation function. However, there is statistical evidence for the stationarity of the series.
- Product Group 3: There is no statistical evidence of a lagged effect from the ACF, however, for the PACF plot there is statistical evidence of a 1 lag effect on lag 12. There is statistical evidence that the time-series is stationary.
- Product Group 4: There is no statistical evidence of a lagged effect from ACF, however, for the PACF plot there is statistical evidence for a 3 lags effect on lag 6, 16 and 17. There is statistical evidence that the time-series is stationary.
- Product Group 5: There is statistical evidence for a one lag effect on lag 6 from both ACF and PACF plots. There is statistical evidence that the time-series is stationary.
- Product Group 6: There is no statistical evidence of a lagged effect from the ACF, however, for the PACF plot there is statistical evidence of a 1 lag effect on lag 17. There is statistical evidence that the time-series is stationary.
- Product Group 8: There is statistical evidence for a 3 lags effect on lag 1, 2 and 3 from ACF. In the PACF plot there is statistical evidence for a 2 lags effect on lag 1 and 3. There is no statistical evidence that the time-series is stationary.

- Product Group 9: There is statistical evidence for a 1 lag effect on lag 12 from ACF. In the PACF plot there is statistical evidence for a 3 lags effect on lag 9, 10 and 12. There is statistical evidence that the time-series is stationary.
- Product Group 11: There is no statistical evidence of a lagged effect from the ACF, however, for the PACF plot there is statistical evidence of a 2 lags effect on lag 12 and 16. There is statistical evidence that the time-series is stationary.
- Product Group 12: There is statistical evidence for a 3 lags effect on lag 1, 2 and 3 from ACF. In the PACF plot there is statistical evidence for a 6 lags effect on lag 1, 2, 3, 12, 13 and 17. There is statistical evidence that the time-series is stationary.
- Product Group 13: There is no statistical evidence of a lagged effect from the ACF, however, for the PACF plot there is statistical evidence of a 1 lag effect on lag 17. There is statistical evidence that the time-series is stationary.
- Product Group 14: There is no statistical evidence of a lagged effect from the ACF, however, for the PACF plot there is statistical evidence of a 1 lag effect on lag 6. There is statistical evidence that the time-series is stationary.
- Product Group 16: There is statistical evidence for a 1 lag effect on lag 3 from ACF. In the PACF plot there is statistical evidence for a 1 lag effect on lag 3. There is statistical evidence that the time-series is stationary.
- Product Group 20: There is no statistical evidence of a lagged effect from the ACF, however, for the PACF plot there is statistical evidence of a 2 lags effect on lag 10 and 13. There is statistical evidence that the time-series is stationary.
- Product Group 36: There is statistical evidence for a 1 lag effect on lag 10 from ACF. In the PACF plot there is statistical evidence for a 1 lag effect on lag 10. There is statistical evidence that the time-series is stationary.

3.2. Data Preparation

The following actions are done to ensure that the data is as clean and accurate as possible, which is crucial for the subsequent analysis and modeling.

3.2.1. Data Imputation

There were missing values in the following features:

- *Production Index Electrical equipment Switzerland* (0.46%);
- *Production Index Electrical equipment World* (5.02%);
- *Production Index Machinery and equipment n.e.c. Switzerland* (0.46%);
- *Producer Prices Electrical equipment* China (10.5%);
- *Producer Prices Electrical equipment France* (15.98%);
- *Producer Prices Electrical equipment United Kingdom* (8.22%);
- *Shipments Index Machinery & Electricals United States* (0.46%);
- *Shipments Index Machinery & Electricals United Kingdom*(8.22%);
- *Shipments Index Machinery & Electricals Switzerland* (0.46%);
- *Production Index Machinery & Electricals Switzerland* (0.46%);

For data imputation, we started by checking what years these missing values were on. For the ones that corresponded to earlier years (considering the data goes from 2004 to 2022), we left them as missing values since we would not use this data for modeling.

Next, for each of these features, we analyzed their behavior over time by plotting their respective time-series. If they seemed to behave in a stable way considering the mean, variance, and autocorrelation, then we analyzed their distribution by plotting their respective histograms and density plots.

For the features that appeared to have a normal distribution, we performed the Kolmogorov-Smirnov test with a confidence interval of 95%. This test is used to decide if a sample comes from a population with a specific distribution (National Institute of Standards and Technology, n.d.). Consequently, if the p-value was greater than 0.05, meaning there wasn't statistical evidence that the variable did not follow a normal distribution, we predicted the missing values based on the fitted distribution.

For the features that didn't follow a normal distribution, we either used the Prophet model to predict the missing values or calculated the mean of a certain time period, depending on the case.

3.2.2. Feature Creation

As stated before, we added Germany's GDP and adjusted the sales considering the inflation. Additionally, we added the feature *Non-working Days*, which is the sum of the number of weekend days and official holidays. This could be an important feature for the sales forecasting.

3.2.3. Sales Analysis

In order to get a better understanding of the past behavior of the sales of Siemens' Smart Infrastructure Division, we conducted a brief analysis of the historical data provided.

When observing the daily sales (Figure 8.1), we can see that there are very prominent spikes in certain days, which could indicate transactions of significant value, or simply days where many sales were made.

Furthermore, in Figure 8.2, we can analyze the monthly sales throughout the years, from October 2018 to April 2022. As a result, it is possible to infer that the last trimester of 2018 was quite unstable, the month of September tends to be the one with the highest value in sales, the lowest sales value was in November 2018 and January 2021, and the sales seem to be more stable in 2020 and 2021 (with the exception of January for the year 2021). In Figure 8.3, we can confirm that September is clearly the month with the highest average sales and, on the contrary, January has the lowest average sales, followed by the month of November.

Finally, the bar plot in Figure 8.4 indicates that the third trimester of the year has, on average, the highest sales value and the last trimester has the lowest, despite the differences not being very significant. Moreover, Figure 8.5 shows that Friday is the day of the week with the highest average sales value, closely followed by Monday.

3.2.4. Feature Selection

With the aim of not having irrelevant or redundant data, as it could negatively impact the performance of our models, we performed a thorough feature selection of the Market data. This is a crucial step in data preparation since some macroeconomic indices could be irrelevant for the prediction of the sales. Furthermore, each index could possibly affect certain product groups, while not influencing others. Besides, the external factors could have different time lags of influence on sales, so it is also important to select the best time lags.

Thus, we started by finding the time lag of each macroeconomic index that had the highest correlation with sales, for each product group. Then, we removed the indices that had a Spearman correlation with sales lower than 1%. Afterwards, we checked which indexes had a correlation higher than 99% between each other and removed the ones that had the lowest correlation with the target (sales). Finally, we applied XGBoost to each product group to analyse the importance of the indices and kept the ones that had a value of feature importance higher than 0.05.

3.2.5. Outlier Detection

Once we selected the most important features for each product group, we performed outlier detection in the Sales dataset. Identifying outliers in time series forecasting is essential. Even a small number of outliers, particularly towards the beginning or end of a time series, can reduce the accuracy and reliability of the forecasts (esri, n.d.).

For the Product group 8, which was non-stationary, we used an Isolation Forest model to detect outliers. For the remaining product groups, which were stationary, we used the Z-Score method, considering as outliers all the observations that were more than 3 standard deviations apart from the mean (0).

Once the outliers were identified, we replaced the values of these observations with NaNs and performed linear interpolation to fill them in. It is important to note that, during the modeling, we tested some models with and without the original values of these outliers.

The existence of these outliers could represent interesting events or trends that affected Siemens' sales. For example, for the Product Group 14, the identified outliers seen in Figure 8.8 were very high values in June of 2019 and July of 2021, which could possibly indicate seasonality, even though we only have data on these months from 2019 to 2021. The lack of high sales during those months in 2020 could be explained the Covid-19 Pandemic. Another event that could have negatively impacted the sales was the blockage of Egypt's Suez Canal that took place in March of 2021, however, we did not find any evidence of this.

3.3. MODELING

Due to the diverse nature and behavior of the sales of each product group, as we saw in the previous sections, it was of the utmost importance that we defined a scalable and efficient way to train, test and hyper tune the parameters of each of our potential models.

To achieve this goal, we built certain functions to assist us:

- For each model that we used, we created a function that finds the most efficient split of the data (considering the RMSE) given the default parameters of the models.
- A function that outputs performance metrics: Mean Average Error (MAE), Root Mean Squared Error (RMSE), R-Squared and the Mean Absolute Percentage Error (MAPE).
- A function that generates a forecast for both the train and the test sets. We also built code that would perform a grid-search on XGBoost, Prophet and the Ensemble Models with the goal of hyper tuning their parameters.

For each product group, we tested the models XGBoost, Prophet, Auto-Arima, Mean, Median, Distribution based Simulation, and an Ensemble Model that averages the results of XGBoost and Auto-Arima. Besides, we also tested these models' using data with and without altering the previously detected outliers, as well as data with different time lags.

Thus, for each product group, we started by determining the best split of the data considering the RMSE. Then, for the models XGBoost and Prophet, we hyper tuned their parameters in order to optimize the model. Afterwards, considering the four different performance metrics mentioned previously, the overfitting, and the visualization of the behavior of the time series, we picked the best model.

When the selected model was XGBoost, Auto-ARIMA or Prophet, we had to deal with the non-existence of values for the exogeneous regressors. To overcome this, we used two different strategies. First, we performed an analysis of the lag effects that each regressor variable had on the adjusted sales. This meant that we were able to extract some real information of the series value and use it as regressors for our models. However, there were usually many values that we could not obtain this way. For these situations, we used the whole historical information of that variables time-series and applied a prophet model to have reasonable estimations for our regressors. Once this process was complete for all necessary regressor, we were able to train our models and generate a forecast for the sales.

Finally, after choosing the best model for each product group, we wrapped all fourteen of them into one single ensemble model.

3.4. Evaluation

As previously discussed, we utilized a function that provides performance metrics to evaluate the performance of each model for every group of products. The evaluation of our final forecast model is determined by the sum of the individual Root Mean Squared Errors, as this combination of models minimizes both training and testing errors. Our final model consisted of four XGBoosts, seven Auto-ARIMA, one Prophet, one Mean and one Median. By considering the performance metrics of each model and combining the most accurate ones, we have developed a forecast model that provides reliable and accurate predictions for each group of products. This approach allows us to identify and address any potential issues with the individual models, ensuring that the final forecast model is both robust and accurate.

In the Appendix, we can see the the Train, Test and Forecast graphs of some product groups.

4. RESULTS EVALUATION

Our **FINAL MODEL GENERATED A FORECAST FOR THE FOLLOWING TEN MONTHS, AND ALTHOUGH IT IS NOT POSSIBLE TO CONFIRM THE ACCURACY OF THE PREDICTIONS AT THIS MOMENT, WE ARE CONFIDENT IN THE METHODOLOGY AND APPROACH USED TO CREATE THE MODEL.** As part of our evaluation, we have previously defined concrete business success criteria that we believe the model should meet in order to be considered effective, as well as data mining goals such as identifying the most relevant predictors and avoiding overfitting. While we must wait for time to confirm the accuracy of our predictions, we are confident to have achieved these previously set goals. We will continue to monitor and evaluate the performance of the model and make any necessary adjustments in order to ensure its continued success.

5. DEPLOYMENT AND MAINTENANCE PLANS

Upon completion of the forecasting model, we defined the following deployment and maintenance plan, which is divided into four different phases:

1. Our team will present the model and its results to Siemens' relevant employees (such as executives and tech and financial teams) to ensure we are meeting the needs of the company. If not, we will go back and rework the model until the needs are met and the results are actionable.
2. Next, we will put the model into production, ensuring compatibility with the current technological infrastructure. The model must also undergo integration and system testing to ensure it is compatible with the other existing systems within the data ecosystem and the necessary changes are made.
3. For an accelerated and smooth adoption, we will provide the necessary training and support to ensure the relevant employees can interpret the model's outputs effectively. We will present essential guidelines that enable the teams to make more informed decisions on production, inventory, and sales strategies based on the forecast results. We recognize that the company's success in using the forecast model depends on their ability to maintain and update it. Therefore, regular training sessions to equip the team with the necessary skills to manage and update the forecast model over time will be encouraged and organized.
4. We will continuously improve and monitor the performance of the model, considering the changes in the market and the variability of the economy. This can be done, for instance, every six months:
    a) As it is crucial to maintain the accuracy and reliability of the forecasts, regular updates are necessary to incorporate new and more recent data and adjust the model's parameters. During these updates, we will reassess the model's performance by comparing its predictions to the actual sales results. This process will enable us to identify any potential areas for improvement and make the right adjustments. Moreover, we will provide a detailed report on the model's performance during each update, which will highlight its strengths and limitations. Any potential issues that require further investigation will also be identified and reported.

b) The deployment and maintenance plan will be revaluated to ensure it remains cost-effective. An analysis of the cost of acquiring and processing the necessary data and updating the model will be conducted. Additionally, the potential benefits of the results, such as improved sales, inventory management and reduced costs due to more accurate production planning, will be assessed. Based on this analysis, recommendations will be made on the most cost-effective plan. The objective is to ensure that the benefits outweigh the costs. To achieve this, we will consider various factors such as the frequency of updates, the amount and quality of data required, and the resources needed to maintain and update the model. This will enable us to optimize the deployment and maintenance plan and ensure that the forecast model remains a cost-effective tool for Siemens.

6. CONCLUSION

In conclusion, we were able to achieve most of our objectives and generate valuable insights to help us understand Siemens' sales behavior and accurately predict them. These insights, coupled with the predictions we made to improve the business, will contribute to its growth and success in the long term. Our analysis of Siemens' sales data shows that there are clear seasonal patterns in the sales of certain products. These patterns were modeled and predicted using time series forecasting methods.

Moreover, we have also outlined a comprehensive deployment and maintenance plan that aims to monitor the progress of the developed model and potentially make new adjustments and updates periodically. We are confident that the insights and results obtained, along with the continual development of our model, will help differentiate Siemens from its competitors and increase their sales margin.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

There are several ways to improve our model and the results we obtained:

- Incorporate data from additional years, which would help generate new patterns and better understand tendencies and patterns over time;
- Create and/or improve automated data pipelines, for instance when it comes to pre-processing, which would help to speed-up and improve the current processes;
- Explore more advanced forecasting models, such as LSTM neural networks, to improve the accuracy of our predictions;
- Use an STL decomposition method, which is a flexible method that would make our analysis more robust;
- Collect more data on customer behavior, market trends, macroeconomic indices, and competitors to gain a more comprehensive understanding of the external factors that influence Siemens' sales;
- Refine feature engineering to train the model, which can have a significant impact on its performance and can help to extract the most relevant and useful information from data;

- Incorporate feedback from Siemens' financial or customer service teams, since they may have relevant insights into customer behavior that could be useful in refining our predictions and ultimately improving the accuracy of our model;
- Collaborate with other departments, such as marketing or operations, which can help to gain a more complete understanding of the factors that may affect sales or production cycles;
- Implement a real-time monitoring system to track sales and inventory levels, which could help to identify trends or anomalies in real-time and adjust our model or supply chain accordingly.

7. REFERENCES

Siemens (2023). Siemens Advanta – Your Leader in Digital Transformation (siemens-advanta.com)

Federal Reserve Bank of St. Louis (2023). *Gross Domestic Product for Germany*. https://fred.stlouisfed.org/series/CPMNACNSAB1GQDE

Statistisches Bundesamt (Destatis) (2023). *Consumer price index*. https://www.destatis.de/EN/Themes/Economy/Prices/Consumer-Price-Index/_node.html#sprg481842

timeanddate (n.d.). *Holidays and Observances in Germany in 2018*. https://www.timeanddate.com/holidays/germany/2018?hol=9

Noah (2022). *Forecasting: The RMSE Measure*. Forecasting: The RMSE Measure – modeladvisor.com

Howell, E. (2022). *An Overview Of Forecasting Performance Metrics*. An Overview Of Forecasting Performance Metrics | by Egor Howell | Towards Data Science

Komor, G., Bony, C. (2022). *How to* Choose a Forecasting Model. How to Choose a Forecasting Model | by Gosia Komor | Towards Data Science

Kansal, S. (2020). *Machine Learning: Why Scaling Matters*. Machine Learning: Why Scaling Matters (codementor.io)

IBM (n.d.). *What is overfitting?*. What is Overfitting? | IBM

National Institute of Standards and Technology (n.d.). *1.3.5.16.Kolmogorov-Smirnov Goodness-of-Fit Test*. 1.3.5.16. Kolmogorov-Smirnov Goodness-of-Fit Test (nist.gov)

esri (n.d.). *Understanding outliers in time series analysis*. Understanding outliers in time series analysis—ArcGIS Pro | Documentation

Banerjee, P. (2020). *A Guide on XGBoost hyperparameters tuning*. A Guide on XGBoost hyperparameters tuning | Kaggle

Amy (2022). *Hyperparameter Tuning and Regularization for Time Series Model Using Prophet in Python*. Hyperparameter Tuning and Regularization for Time Series Model Using Prophet in Python | by Amy @GrabNGoInfo | GrabNGoInfo | Medium

8. APPENDIX

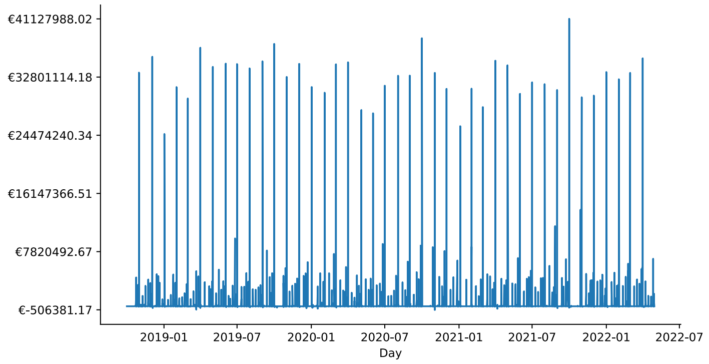

Figure 8.1 – Total Sales per Day

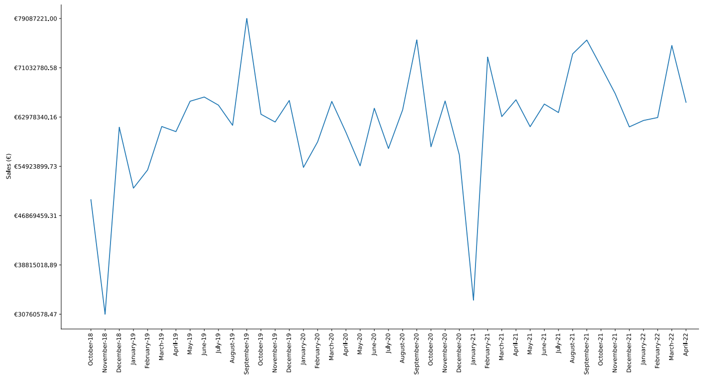

Figure 8.2 – Total Sales per Month of Each Year

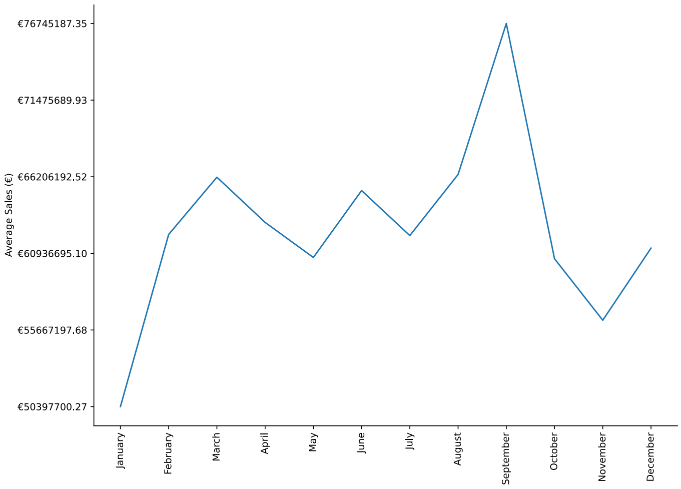

Figure 8.3 – Average Sales per Month

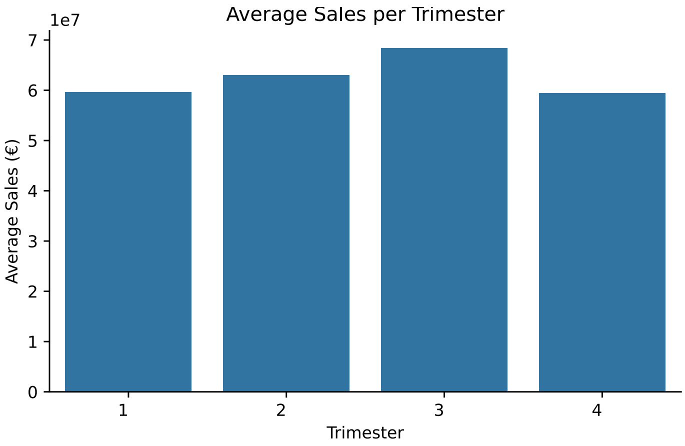

Figure 8.4 – Average Sales per Trimester

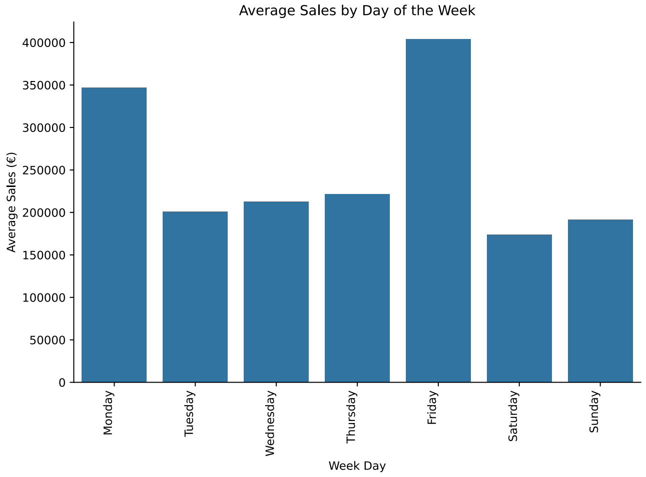

Figure 8.5 – Average Sales by Day of the Week

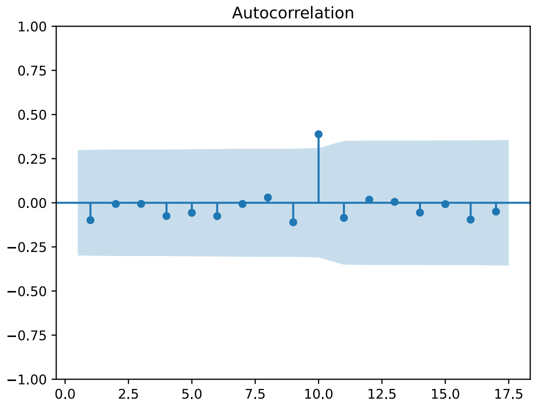

Figure 8.6 – ACF of Product Group 36 Time Series

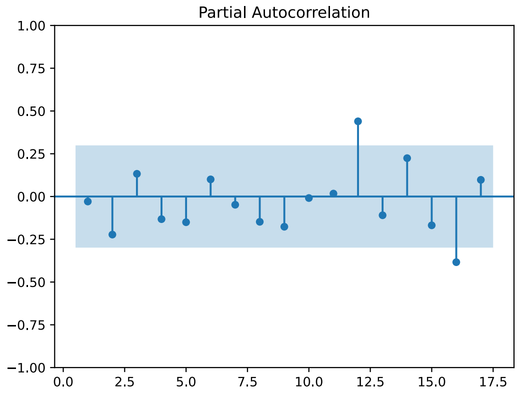

Figure 8.7 – PACF of Product 11 Time Series

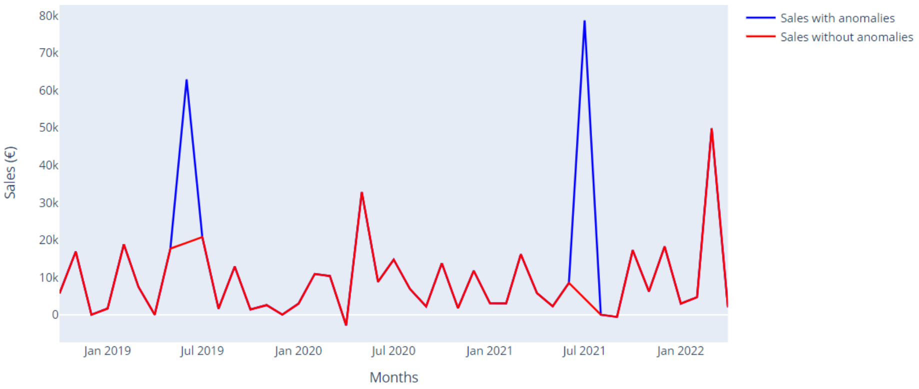

Figure 8.8 – Sales of Product Group 14 With and Without Anomalies

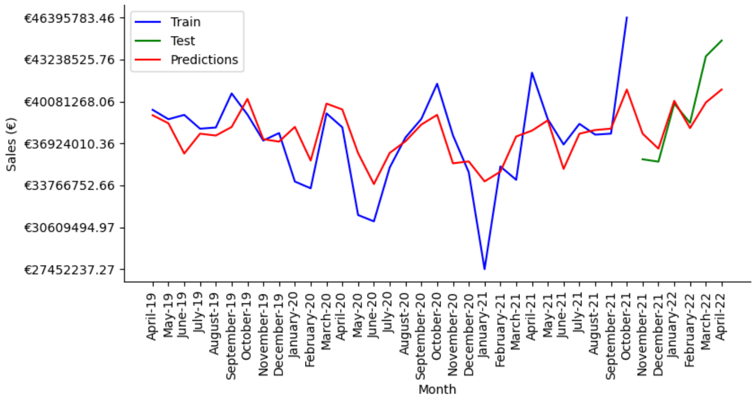

Figure 8.9 – Train, Test and Forecasts of Product Group 1

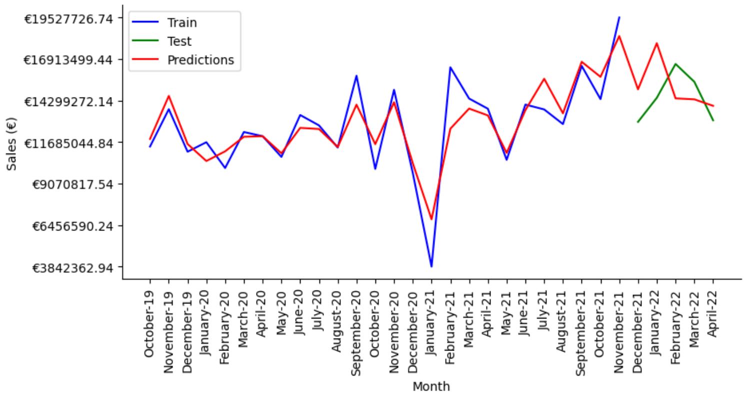

Figure 8.10 – Train, Test and Forecasts of Product Group 3

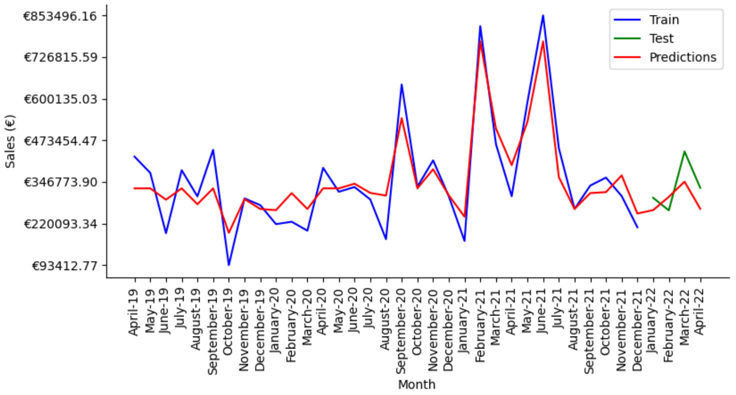

Figure 8.11 – Train, Test and Forecasts of Product Group 4

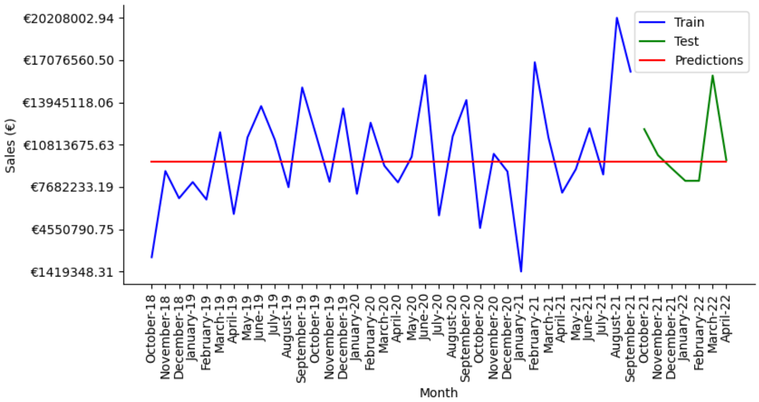

Figure 8.12 – Train, Test and Forecasts of Product Group 5

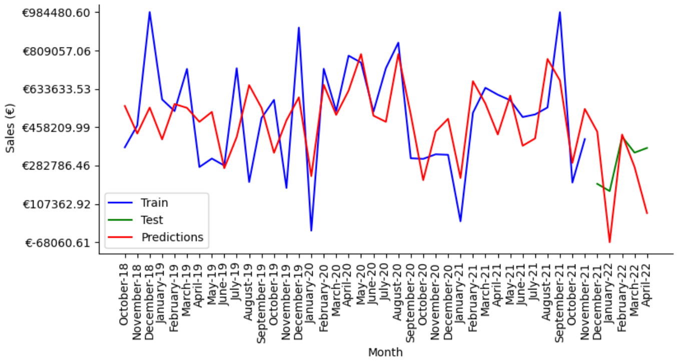

Figure 8.13 – Train, Test and Forecasts of Product Group 6

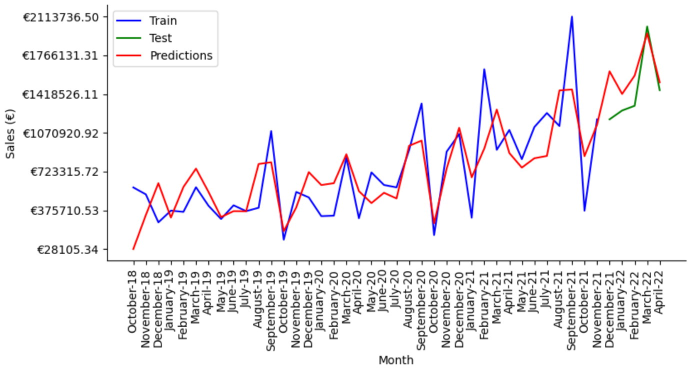

Figure 8.14 – Train, Test and Forecasts of Product Group 8

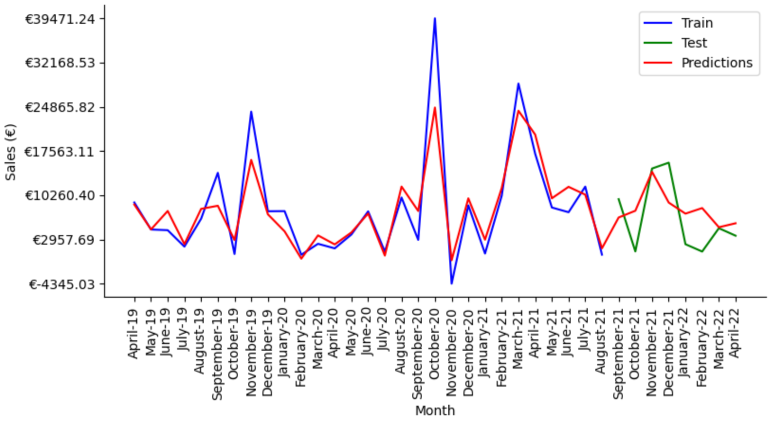

Figure 8.15 – Train, Test and Forecasts of Product Group 9

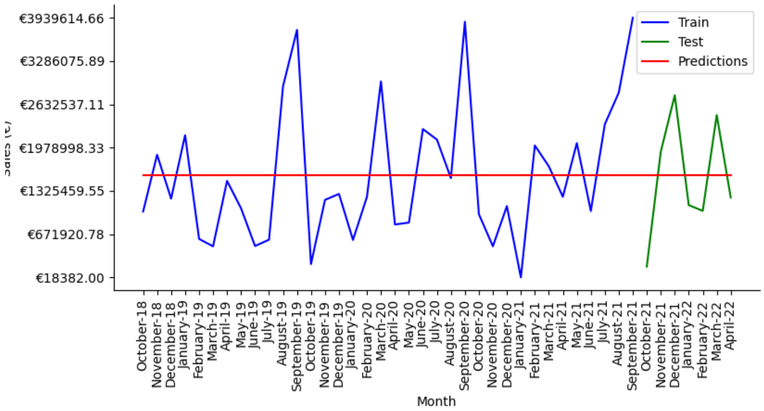

Figure 8.16 – Train, Test and Forecasts of Product Group 11

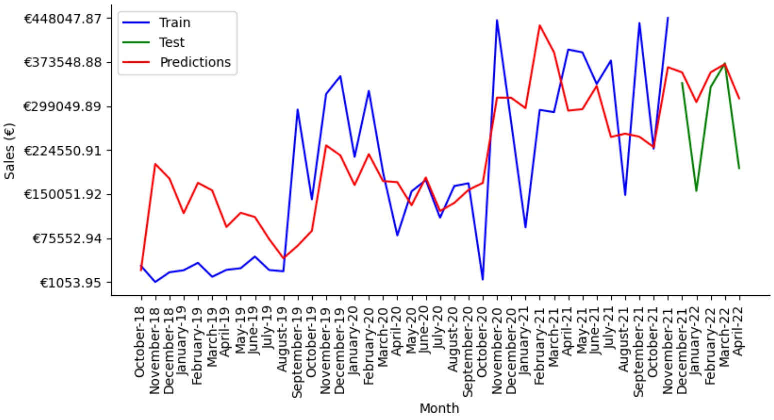

Figure 8.17 – Train, Test and Forecasts of Product Group 12

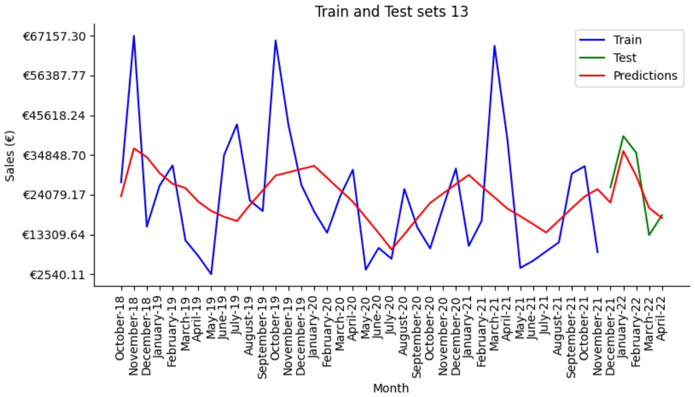

Figure 8.18 – Train, Test and Forecasts of Product Group 13