# MDSAA

Master Degree Program in

**Data Science and Advanced Analytics**

## Business Cases with Data Science

Case 2: Sales Forecast

David Psuik, number: 20230818
Peter Falterbaum, number: 20230956
Luis Penteado, number: 20230441
Noah Campana, number: 20230996

Group A

**NOVA Information Management School**

**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa

April, 2024

# INDEX

# 1. Executive Summary

Siemens’ Division in Germany is leveraging Artificial Intelligence (AI) to revolutionize its sales forecasting and leveraging it for pricing strategies. This ambitious project targets the development of monthly sales forecasts across 14 product groups for a period of 10 months. This initiative aims to refine pricing strategies, mitigate the costs associated with forecasting inaccuracies, such as over- or understocking, and consequently, boost sales margins.

The project employs a nuanced approach to model selection and evaluation, with a customized strategy for product group 1, reflecting its sales significance. Here, a range of tailored models were tested to find the best fit. For the other groups, a streamlined pipeline compared various models to select the most effective one for each category. This method ensures precise forecasts and sets a scalable blueprint for future efforts.

This project uses the Root Mean Square Error (RMSE) as the primary metric of forecast accuracy. Beyond numerical precision, the impact of AI-driven forecasting is manifested in operational enhancements, including improved sales offerings and customer satisfaction, alongside optimized inventory management. These outcomes affirm the role of qualitative forecasts in driving business process optimization.

Extending the scope of digital transformation within Siemens, this initiative represents an important step towards embedding advanced technological solutions in key business processes. The deployment and ongoing management of this predictive model requires a collaborative effort, ensuring the integration of forecasting insights into strategic decisions across various departments.

In conclusion, this project´s focus on developing AI-driven sales forecasts for Siemen Advanta´s product groups illustrates the impactful possibilities of efficient forecasting and pricing strategies. By successfully implementing these forecasts, substantial enhancements in operational strategies will be enabled.

# 2. Business Needs and Required Outcome

## 2.1. Business Objectives

Siemens follows the strategy to leverage AI in its business processes, gaining efficiency and competitive advantage. Within this context, the Smart Infrastructure Division in Germany aims to develop AI-driven sales forecasts to enhance strategic management decisions.

Throughout this project Siemens Advanta thrives to automate price recommendations based on machine learning to develop more precise pricing strategies, reduce costs associated with inaccurate forecasting, such as over- or understocking, and consequently increase sales margins. The goal is to develop a monthly sales forecast for single product groups. The time span of the prediction captures 10 months and the solely available input parameter for the prediction is the date, making the model in scope a univariate time series forecast.

Employing an effective data-driven forecasting approach underlines thus the effort Siemens spends on enabling digital transformation across all company sectors, leading efficient and effective operations.

## 2.2. Business Success Criteria

The sales forecast will be evaluated on quantitative and qualitative measures, whereas the overall preciseness of the prediction is the strongest indicator and will be assessed using the RMSE (Root Mean Square Error) metric. On a broader perspective, success will be visible on operational levels such as improvements in sales offerings and enhanced customer satisfaction. An adequate management of product stocks can additionally confirm a working forecast model. These criteria showcase the AI produced enhancement in business processes and highlight the direct impact between data insights and business.

## 2.3. Situation assessment

Siemens Advanta is aware of the challenges arising with manual forecasting, specifically with the labor intensity, possible bias and untransparent information. AI-driven solutions on the other hand are replicable, set to certain rules and within this setting open to data-gained insights. Having more data available for analysis in the future increases the need to leverage technical capabilities even more. Training personnel in understanding and overseeing trends and noteworthy details will become an impossibility.

Siemens' strategy in incorporating AI innovations in business processes and practices underlines this informed investigation. The realization of high failure rates among past data science projects drives Siemens to stick thoroughly with their implementation of the CRISP-DM framework, to control known risks and enhance the success rate of upcoming AI implementations. CRISP-DM consists of specific steps to fixate the steps taken to execute a data science project. The framework explicitly allows iterations to drive the agile mindset of fast failure and ongoing adaptions. This report is structured in the steps of CRISP-DM to realize the described benefits.

### 2.4. Determine Data Mining Goals

In this project, our main objective is to leverage patterns identified in past data to forecast sales outcomes effectively. By searching for relationships between certain trends and changes along our target variable, we aim to find meaningful patterns that could influence the sales. This involves an analysis of correlation metrics and the examination of lag effects over different time spans to understand how the sales changes and follows patterns over time. In this report we employee this investigation across 14 different product groups, each treated and forecasted individually.

However, our success criteria extend beyond merely achieving an accurate predictive model. Besides pattern recognition, the project's data mining goals can be categorized into 4 target areas, which have the potential to be fueled automated by the outcomes of this analysis:

**Demand Forecasting**: Precise demand predictions are crucial, allowing the company to meet market demands efficiently, maintain optimal inventory levels, and avoid stock shortages or leftovers.

**Price Optimization**: Sales outcomes and forecasts can be leveraged in relation to pricing goals to identify optimal price points that maximize profitability while aligning with consumer expectations and market trends.

**Anomaly Detection**: Using the forecast, it's possible to identify unusual or not expected sales activities. This helps to spot potential consequential issues early on, from supply chain disruptions to unexpected shifts in consumer behavior.

**Seasonality Analysis**: Seasonal trends derived from the developed forecasts enables Siemens Advanta to anticipate periodic fluctuations, optimizing inventory and marketing to leverage peak periods and mitigate downturns.

## 3. METHODOLOGY

### 3.1. Data understanding

This chapter explores the two datasets which build the foundation of this project. The first provides sales data of Siemens Advanta and the second complements these sales by external market data such as resource costs and indices captured for several markets in the world. Although only one was utilized for forecasting, both datasets played a crucial role in the initial data exploration and analysis phases. In the following, both datasets are described separately for source understanding.

#### Sales dataset

The first dataset contains daily sales data for 14 different product groups, spanning from 1st October 2018 to 30th April 2022, with the sales amount stated in Euros. There are some negative values for sales, which were considered as returns. Their total sum is considerably low – especially compared to the total sales per product, thereby having a small impact on a broader perspective. Consequently, the relevance of these returns will be limited to their contribution to the total sales, without investigating on a more granular level.

The data shows significant sparsity and irregularity: 73% of the entries of the data indicate zero sales per day, standing in contrast to days of high sales. On average, the daily sale across all products exceeds 270,000 Euros. This contrast highlights irregular distribution of sales across the time span, suggesting the need for treatment as it presents challenges for forecasting. Days with zero sales insert gaps in the time series, which make it difficult to identify underlying trends and patterns as time series relies on rather continuous data points. Entries with zero sales interrupt potential continuity in the data.

To handle the sparsity of the data and to introduce a monthly perspective, which is needed for the forecast, the sales data has been aggregated monthly by summation. In the future exploration and modelling only monthly aggregated data will be considered relevant.

The frequency of sales varies across the product groups highly, having a range from 70 to 321 products sold in the given time frame. Product Group 16 rises as the most frequent one, while product group 14 ranks lowest. A detailed presentation of counts and total sales per product can be found in appendix 1.

Product group 1,3 and 5 are considered the most important ones as they bring by far the most revenue to the company. Product group 1 accounts for 58% of the total sales, product group 3 for 20% and product group 5 for 16%, cumulating together 94% of the overall sales (for further reference refer appendix 2). Moreover, these products tend to be more expensive as the number of products sold is not significantly higher than the one of other product groups, in contrast to their revenue.

Considering the sales of the products, product group 8 has the overall highest correlations to product group 11 and 12, what might mean they are complementary products which are commonly sold together. A low correlation between product group 11 and 12 indicates that possibly both product groups are solely single complements of product group 8.

The time series for product group 1 has a significant drop in November 2018, which is additionally visible in the overall sales of all product groups, which underlines the big ratio product group 1 holds on the total sales.

The overall monthly counts have 2 significantly outliers in January 2019 and 2020, while the overall sales are rather average as product group 1 has normal sales for both months.

In the data is a spike noticeable every September. That's when the fiscal year ends for Siemens which can explain this artificial peak by invested sales endeavors and budget strategic transactions. Another spike seems to be in March, which might be caused by semestral sales goals.

The amount of sold products correlates highly with the total sales per month and product. Exception from this dynamic is January 2020. In this month the amount of sold products is the lowest of the given dataset, whereas the total sales are on average.

### MARKET DATASET

The second dataset contains 48 columns about different market related measures from February 2004 to April 2022. It provides information about several indices for production and shipment of machinery and electrical equipment in 8 different countries, additionally consolidated for Europe and the world. It also holds prices for raw materials like metals, gas, oil, and energy.

Merging both datasets, sales and market, we can reach conclusions about the indexes' influence on sales. In analyzing the relationship between sales data and various market indexes, including German, European, and global indexes, significant correlations and influences have been identified, underscoring the indexes' impact on sales across different sectors.

Electrical equipment and product groups 8, 12, and 3, alongside the Machinery and Equipment sector, show strong correlations with specific market indexes, highlighting their potential as indicators of economic trends and sector-specific demands. This suggests that sales predictions could be more accurate when tailored to each product group due to their distinct market influences. Moreover, the immediate response of these groups to global index changes, especially with metals, underscores the dynamic nature of market impacts on sales.

## 3.2. Data preparation

To overcome rapid fluctuation and high number of missing values (73% zero sales entries) a monthly data aggregation was conducted. This approach not only effectively streamlines the dataset, but additionally algins better with the characteristics of RMSE to prioritize substantial prediction errors – those that could significantly impact decision-making and strategic planning – over smaller discrepancies. The updated dataset spans across 43 months, which constitutes in a 30 times smaller dataset.

In our data exploration, we identified the potential impact of specific time features – like the number of Mondays and weekends, and the fiscal year-end – on sales trends for various product groups. These insights led us to engineer features such as number of Mondays, number of weekends, whether it’s the fiscal year end, and the fiscal quarter. With that, an improvement in forecast performance by capturing these temporal influences is anticipated. Notably, these features are directly calculable for future periods, making them practical for inclusion in our models without requiring external variables forecasting. This approach is particularly valuable in regression models, where capturing the date component can significantly inform the analysis and outcomes.

The second dataset contains a variety of market and resource indices, as outlined in the previous chapter. In time series forecasting, incorporating external variables – factors influencing the time series without being part of it – can significantly enhance the model’s performance. These variables, which do not directly relate to the sales data itself, provide beneficial contrast and depth to identify underlying trends and patterns. External variables can introduce relevant and supportive information to the forecast. Although they present a lot of challenges – namely the need to include them as parameters for predictions – their potential benefits cannot be understated. A common strategy for this scenario is to predict these external variables first, using them as features to predict the target variable secondly. This technique increases the model complexity by far, elevating in turn the potential of overfitting and noise interference.

Because of the increasing complexity, the introduction of noise and higher chances of overfitting to artificial data, we chose to not include external variables and solely focus on the sales data itself such as the generated time features which don't need forecasting as they can simply be calculated for future dates.

## 3.3. Modeling

In this chapter the design of the employed modelling will be elaborated in detail. In scope is the sales forecast of the 14 product groups for a time frame of 10 months. To implement for each product group an individual forecasting model, the python framework pycaret was used to find the best-suited model in a generic, pre-defined way. Product group 1 was treated individually as it accounts for 58% of the total sales and has therefore an outstanding impact on business side. The design of both approaches is layed out in the following.

### Pycaret Pipeline

The forecasting pipeline employed in this project utilizes the PyCaret library[1]. The package can be leveraged to implement a pipeline in which model training and evaluation is done sequentially, returning a ranking of best performing models. This decision was driven by the need for a systematic approach to identify the most effective model for each of the 14 product groups.

Initially, several time series models are applied to the dataset. PyCaret facilitates a comparative analysis to determine the best-performing model. The model evaluation is an automated process that relies on comparing predictions against actual values using the RMSE as primary metric. In parallel, several regression models are evaluated in the same manner. This two-way approach ensures a comprehensive exploration of both time series and regression to identify the most suitable model for each product group.

The overall best performing model will be selected and finalized by training it on the entire dataset. This step ensures that the model leverages all the available data. The final model incorporates the exogenous time related variables to enable a comprehensive forecast.

This methodology streamlines the forecasting process while ensuring a high-performance forecast by comparing a wide range of models to select the optimal one for each specific case.

[1] PyCaret - pycaret 3.0.4 documentation

The automation-friendly aspect of this approach lay the groundwork for it to be seamlessly adapted for future projects, enhancing scalability and efficiency of forecasting efforts.

### Modelling Product Group 1

For Product Group 1 we aimed to develop a more individualized model, as it accounts for over half of the total sales of the German Smart Infrastructure Division. Models such as ARIMA and Prophet, which are generally considered rather simple, were used to capture the underlying trends and seasonality. We also explored advanced machine learning techniques using LSTM and XGBoost algorithms, integrating several kinds of additional features such as lags and rolling mean calculations to encapsulate temporal dependencies. Additionally, date features like year and month were utilized to enhance the models' perception of cyclical patterns, ensuring a comprehensive time series analysis for more precise forecasting.

## 3.4. Evaluation

This chapter first outlines the setting for evaluating the developed forecasting models and how the metric of evaluation is influencing the model training and assessment. Subsequently, the performance of the developed models will be presented for the division's most crucial product groups 1, 3 and 5. This exemplary showcase aims to offer a transparent view into the underlying, involved technical procedures.

### Setting

As the main evaluation metric for this project the Root Mean Square Error (RMSE) has been chosen, as specified by the business requirements. The RMSE squares discrepancies in predictions by squaring the errors before averaging them. This means that larger errors have a more weighted effect on the overall metric compared to smaller deviations.

### Results

**Group 1:** ARIMA, Prophet, LSTM and XGBoost were individually trained and evaluated on the first product group. After assessing and modifying all these models, the XGBoost algorithm achieved the best performance regarding the RMSE.

The resulting forecast can be seen in the following visualization. The forecast begins in May 2022 and endures 10 months.

**Group 3:** The forecast for product group 3 was calculated by utilizing the created PyCaret Pipeline. The resulting best model is an elastic net with an RMSE score of 2,141,569. This RMSE value indicates the model´s average prediction error across a 10-month forecasting period. A more interpretable metric in this case is the mean absolute percentage error (MAPE) with a value of 0,1943. This value indicates that, on average, the model´s predictions were off by 19.43% from the actual sales values.

As illustrated in Figure 1, it's not intuitively apparent from the data whether there is a definitive trend or seasonal pattern, complicating the forecasting process. The accuracy of forecasts is generally dependent on both the quantity and quality of data fed into the model. Despite these challenges, the model manages to produce a reasonable forecast for the specified timeframe.

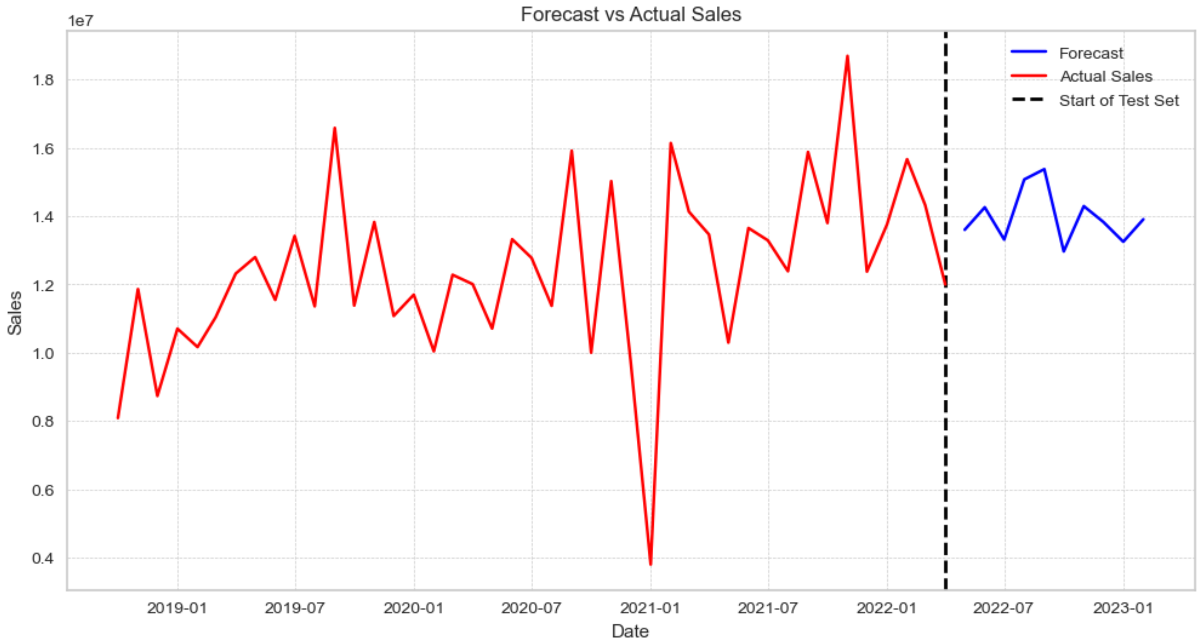


Figure 1 - Product Group 5: Forecast and Actual Sales over time.

**Group 5:** Group 5 ranks as the third most significant product group by overall sales volume. Like the prior approach, the forecast was generated by using the forecasting pipeline. For the test set, the model achieving the best performance was a Ridge Regression, which posted an RMSE of 3,110,096.

The final forecast applied to the actual test data is illustrated in appendix 4. This timeseries also lacks a clear trend or discernible structure, presenting challenges for accurate prediction.

## 4. Results Evaluation

The monthly sales forecasts natively support strategic management decisions through AI-driven insights. The use of an univariate approach enables the potential for a simple and smooth integration into different business areas. However, the real effect of the developed forecasting models can only be fully assessed in real-world environments. With that, an evaluation of the model's predictive accuracy, its potential to adapt to market changes, and its effectiveness in optimizing inventory levels to avoid overstocking and understocking can be performed.

To comprehensively analyze and understand the forecasting ability, a proof of concept could be implemented with further monitoring over a certain time frame. While comparing the forecasts with actual sales data, the models can be further refined iteratively, sticking with the most recent trends.

Each of the defined data mining areas can be assessed in real-world enrollment, using key performance indicators (KPIs). This KPIs could be set to evaluate different scenarios, which can be evaluated based on the real and the predicted sales. For example, a baseline for inventory KPIs can be documented before using the forecasting and compared to the KPIs after using the forecasts. This could showcase practical changes related to the forecasts.

## 5. Deployment and Maintenance Plans

The successful deployment and ongoing management of a predictive sales model require a collaborative effort across multiple departments. This means that the success of this model's deployment would be dependent not only on the accuracy of the predictions, but also on how the company is able to assimilate the predictions into its operational, financial and strategic decision-making processes. To reach the real value for the product make clear their diverse responsibilities with the information provided. The following will outline a comprehensive deployment process and the necessary steps to ensure that the model delivers its intended benefits while allowing for agility and responsiveness to unforeseen changes.

Input from different departments: Predicting sales isn't always straightforward when you just look at past trends. It is necessary and useful, but not the whole picture. Since Siemens Advanta is not isolated by other businesses, the sales team needs to be aware of upcoming engineering projects. Knowing what other companies are planning helps to assess the potential market volume and to start building connections to future customers.

Launching new products is another important factor. To expand the company's catalog, addressing different market demands and as a competitive advantage may potentially revolutionize sales outcomes and should be also input for the predictions.

Different scenarios: In addition to the results of the prediction model, the input from different areas should be introduced to the model, after considering and specifying different scenarios. By executing a scenario analysis, the company can predefine some actions to deal faster with the business volatility. Using external information like upcoming projects on the one hand, and internal insights such as details about releasing new products, we can improve the forecast not only with past trends but also with initially unpredictable factors.

Dashboard Development: To help track the results and ensure user acceptance, a single dashboard will be developed with all the metric sets and predictions versus actualized sales. This tool will be useful to all stakeholders as it provides a detailed but simplistic visual comparison between prediction and reality, easily identifying areas of variance while signaling department heads about any anomalies. The tool shall have a drill-down function for detailed analysis purposes and should also be customizable for access based on the department. This enables the company for business-relevant analysis on different detail levels.

Price recommendation: Use the knowledge from the sales forecast generated by our machine learning model as leverage to inform the automated price recommendation system. Via A/B testing it is possible to ensure that these strategies truly benefit the decisions taken using the insights provided by the forecast and if it brings value as expected.

## 6. Conclusions

In conclusion, this research showed ways to enable historical sales data from Siemens Advanta to predict future sales. A ready-to-deploy pipeline was created which trains and evaluates a plethora of different forecasting models and returns the best one, which can be applied for predictions. These forecasts potentially leverage existing processes in Siemens to increased efficiency – e.g. optimizing inventory levels – and enable qualitatively not yet implemented opportunities such as seasonality adapted offers for customer.

This report outlines specific steps to test, refine and enroll the data-driven approach in real world scenarios, offering automation integration in several areas such as price recommenders and production management based on sales forecasts.

Furthermore, the proposed approach of sales forecasting can be integrated into automated forecasting, leading to predictions based on most recent information.

Further evaluation of the developed forecasting models is essential to determine their real-world influence across various business segments and to assess the durability of their forecasting accuracy. Continuous performance monitoring over extended periods will reveal how well the models predict future trends and adapt to evolving market conditions.

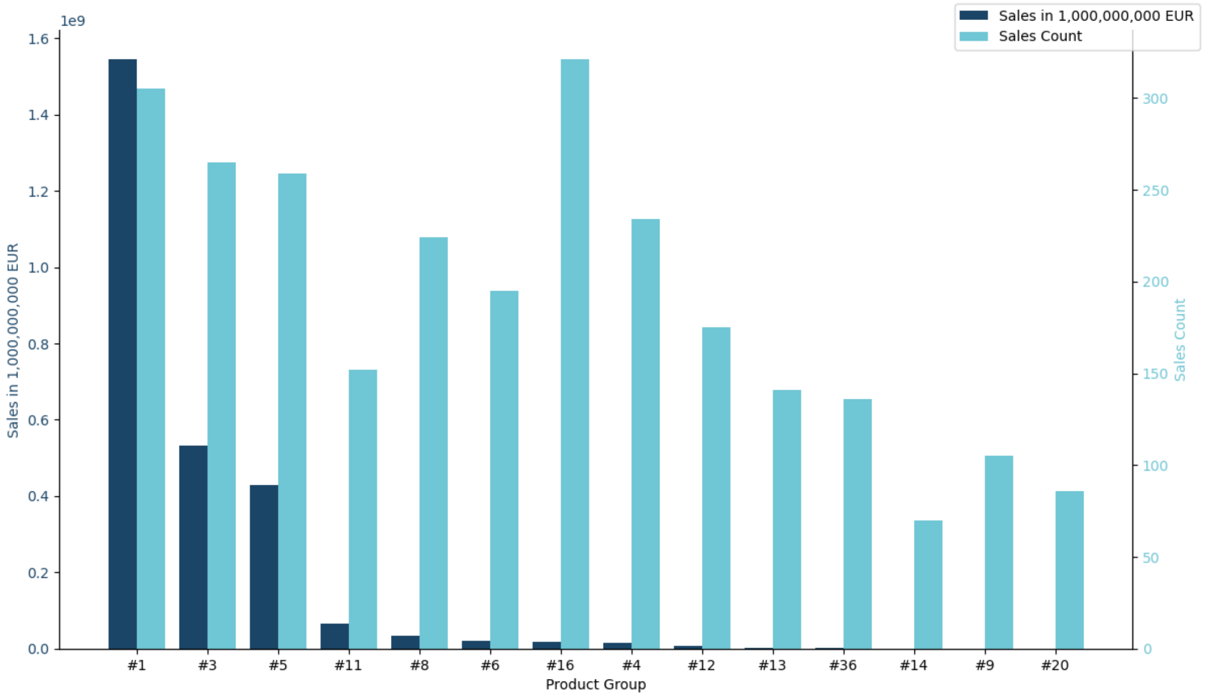
The project has limitations that can be addressed in further research. Having only limited business-related insights leads to the recommendation for in-depth expert insights, prioritizing certain product groups and point out to already known trends. Furthermore, our approach focused on model training based on time features, whereas external data was not used to enhance the model’s complexity. As outlined in the report, increasing the model’s complexity has the potential to enhance the predictive power, which needs to be investigated in a more extensive design.

# REFERENCES

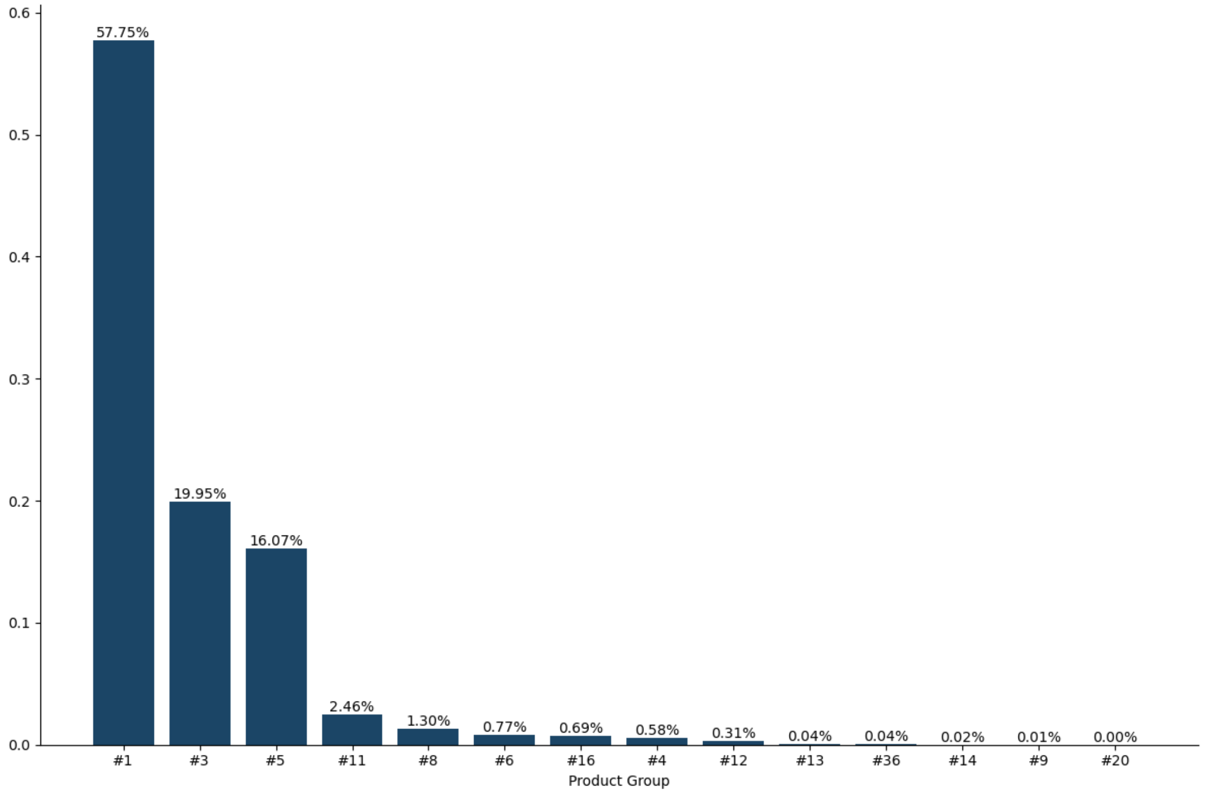
*PyCaret — pycaret 3.0.4 documentation*. (n.d.). https://pycaret.readthedocs.io/en/latest/

# APPENDIX

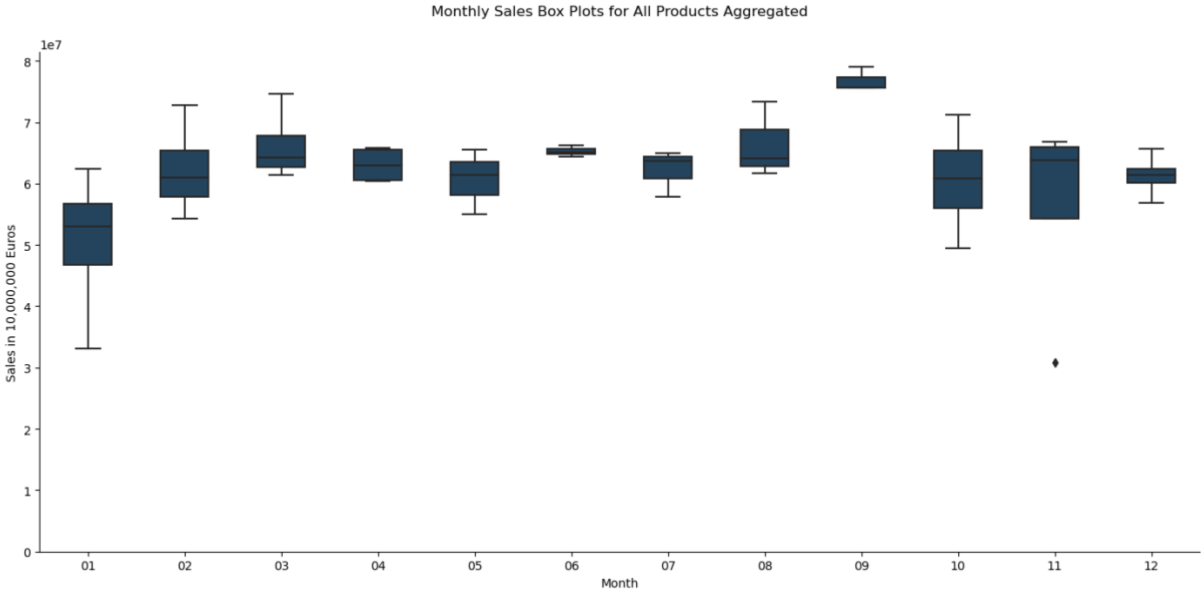
Appendix 1: Comparison of Product Groups by Sales EUR and Count, Sorted by Sales.



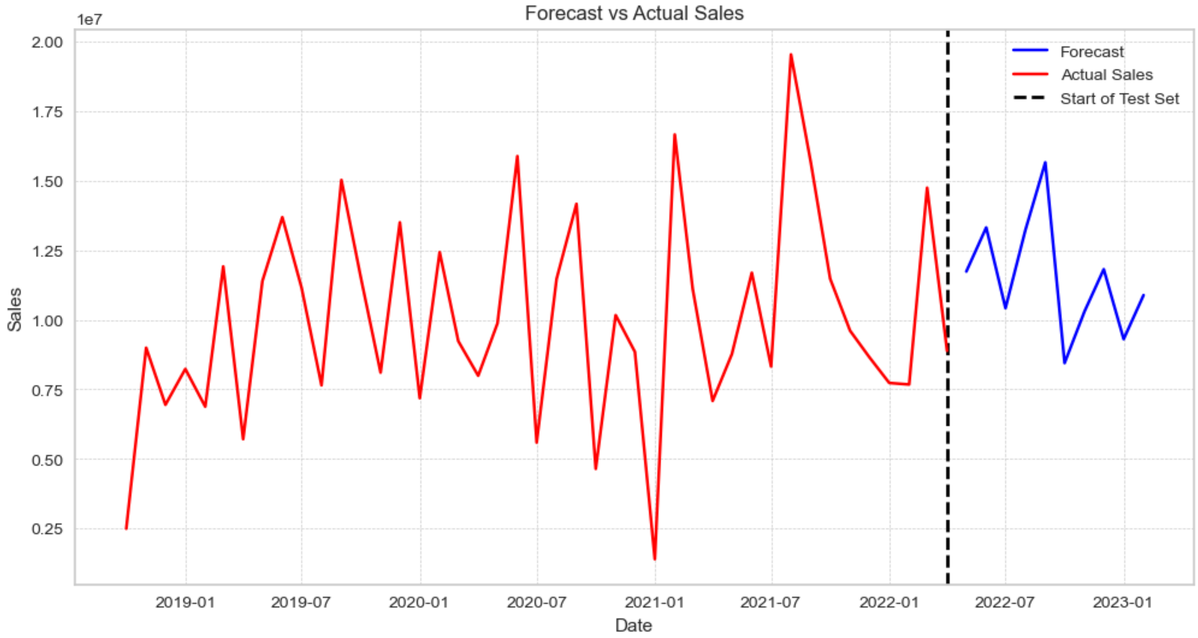
Appendix 2: Proportion of Each Product Group's Sales to Overall Total Sales.



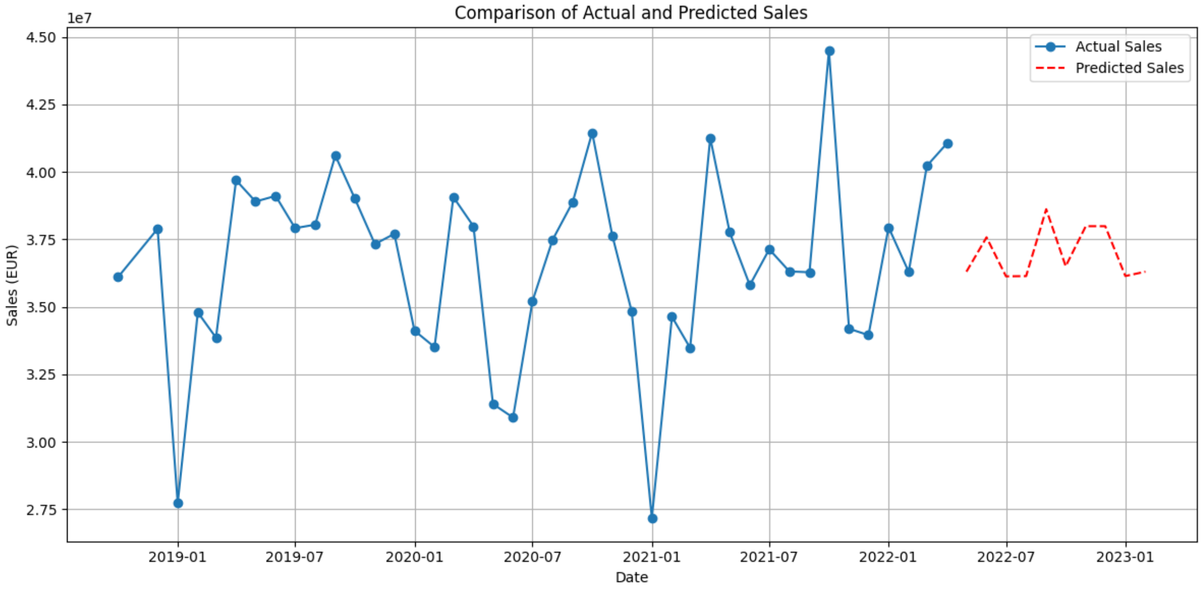
Appendix 3: Monthly Sales Box Plots for All Products Aggregated.



Appendix 4: Product Group 5 – Forecast and Actual Sales over time.



Appendix 5: Product Group 1 – Forecast and Actual Sales over time


Comparison of Actual and Predicted Sales
1e7
Sales (EUR)
Date
Actual Sales
Predicted Sales