NOVA IMS Information Management School

MDSAA

Master's Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Siemens Advanta Consulting: Case study challenge

Diogo, Marquês, number: 20230486

Diogo, Pimenta, number: 20230498

João, Lopes, number: 20230748

João, Maia, number: 20230746

Pedro, Ventura, number: 20230728

Group K

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

April, 2024

INDEX

1. EXECUTIVE SUMMARY

The sales forecasting project for Siemens involved thorough analysis and modeling efforts aimed at accurately predicting monthly sales for selected product groups within the Smart Infrastructure Division. The project encompassed various stages, including data understanding, preparation, modeling, and evaluation, following the CRISP-DM framework.

During the data understanding phase, descriptive statistics and data visualization techniques were used to comprehend the characteristics of the datasets, which included market data and sales data. Insights were derived regarding sales patterns, seasonality, and correlations among variables.

In the data preparation phase, extensive feature engineering was conducted to integrate macroeconomic indicators, seasonal factors, and lag features into the dataset. Missing values were addressed through predictive imputation.

Feature selection techniques were employed to identify the most relevant variables for modeling. Correlation analysis and feature importance assessments were conducted to streamline the dataset and eliminate redundancies.

Various machine learning models were applied, including KNN, Linear Regression, Random Forest, XGBoost, Ensemble, and SARIMAX, to forecast sales for different product groups. The models were evaluated based on their Root Mean Square Error (RMSE) values, with lower RMSE indicating better predictive performance.

The results showed promising outcomes, with different models performing optimally for different product groups. For instance, Linear Regression yielded the lowest RMSE for GCK 4 and GCK 5, while Random Forest performed well for GCK 6 and GCK 8.

Finally, plans for deployment and maintenance were outlined to ensure the continued relevance and effectiveness of the sales forecasting model. Strategies for model improvement were also discussed, including implementing correlation of product sales, utilizing grid search for model optimization, and exploring alternative feature selection metrics.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1 BACKGROUND

Siemens is a multinational conglomerate company that specializes in various fields such as energy, healthcare, infrastructure and transport. The organization wants to forecast monthly sales of selected product groups from one business unit of its Smart Infrastructure Division.

Siemens Smart Infrastructure Division offers a comprehensive portfolio of products and services designed to provide solutions for building automation, energy management, and infrastructure technologies. Their portfolio consists of products and services for smart buildings, power distribution, electrical installation, and grid solutions, helping customers optimize resource consumption, improve safety and enhance the overall quality of life.

2.2 BUSINESS OBJECTIVES

To accurately determine the monthly sales forecast for the ten-month period between May 2022 and February 2023, a thorough analysis of the available data is imperative. The macroeconomic indices of each country must always be taken into consideration, as the data pertains to multiple countries. Finding the most significant variables and creating the model that reduces the forecast error the most is the aim.

2.3 BUSINESS SUCCESS CRITERIA

When considering business success criteria related to sales forecasting and strategic decision-making, it's important to establish measurable benchmarks that align with overarching business goals.

Forecast accuracy is a critical success criterion that directly impacts the effectiveness of sales forecasting models and the overall business performance. By setting specific targets for forecast accuracy, such as achieving a low RMSE (Root Mean Square Error) value between actual and predicted sales figures, organizations can demonstrate the reliability and predictive power of their forecasting methodologies. It plays a pivotal role in guiding strategic decision-making, resource allocation, and operational planning within businesses. A highly accurate forecasting model provides stakeholders with confidence in future sales projections, enabling proactive management of inventory, production, and customer demand. RMSE is a statistical measure that quantifies the average magnitude of errors between predicted values and observed values. A lower RMSE indicates better accuracy, suggesting that the forecasting model effectively captures underlying trends and patterns in sales data.

Ultimately, we aim for the model to prove its efficiency in predicting sales over a ten-month period, demonstrating superior performance compared to actual sales results obtained during that same timeframe. This validation underscores the model's real-world applicability and its value in supporting informed business decisions and strategic initiatives.

2.4 SITUATION ASSESSMENT

For the development of this project, we used the file "Case3_Market data," a dataset comprising 221 rows and 48 columns, containing information about the company's macroeconomic indices

for the most significant countries where the company operates and some world raw materials prices indicators, and the file "Case3_Sales data," with 9802 rows and 3 columns, providing information on monthly sales of each product category in monetary terms. Both files were provided by Siemens.

To prepare the data for analysis, we employed the Python programming language and several of its libraries. Furthermore, for data visualization, which is essential for communicating insights and presenting analytical results, we employed libraries such as Seaborn, Matplotlib and Plotly, and some other software's like Microsoft excel.

In an initial analysis, several challenges arose that required addressing:

- Data Format Issues: We encountered the existence of two files, demanding changes to merge them into a single file to perform the necessary steps for obtaining the final sales prediction model.
- Data Quality Issues: In the "Case3_Market data" file, the dataset contains an initial row providing additional information to the second row. This information can be merged with the second row to make the data clearer and more easily studied in the future. Additionally, addressing the presence of missing values is necessary. Concerning the "Case3_Sales data" file, we encountered some issues that hindered the usability of the data for future modeling, necessitating modifications.
- Data Type Issues: In both files, the variable types are incorrect, requiring adjustments.

The unpredictability of sales is heightened by the fact that the available data covers years affected by a pandemic. This external factor may significantly impact the accuracy of predictions. Additionally, other external factors such as inflation can further complicate predictions. Inflation may affect both production costs and changes in product selling prices. The presence or absence of effective marketing may also translate into an increase/decrease in the number of sales, which the group is unable to predict and relies solely on the available data.

2.5 Determine Machine Learning goals

To evaluate our sales forecasting predictions effectively, it is imperative to establish clear machine learning goals that align with our business objectives and available information. Considering the complexities of forecasting sales trends, we have defined the following technical objectives:

- **Explore diverse models**: Leveraging ensemble learning techniques such as Random Forests or Gradient Boosting enables us to combine multiple models for improved accuracy and resilience against variations in sales data patterns.

- **Scalability and Performance**: Developing a robust and scalable sales forecasting system is crucial for efficiently handling large datasets and real-time forecasting demands. By using techniques such as Min-Max scaling, we ensure that our model remains effective and adaptable as data volumes increase and business requirements evolve over time. Scalability considerations are essential to maintain optimal performance and

responsiveness, enabling our forecasting system to meet growing demands and deliver reliable insights to support strategic decision-making. This approach not only enhances efficiency but also future proofs our forecasting capabilities against evolving market dynamics and data complexities.

By addressing these technical objectives, our sales forecasting project aims to deliver actionable insights and reliable predictions that empower strategic decision-making and drive business growth to Siemens.

3. METHODOLOGY

Our approach to tackling the sales forecasting challenge follows the CRISP-DM, which is structured into four main phases: Business Understanding, Data Preparation, Modeling, and Evaluation.

In the Business Understanding phase, we conducted workshops and interviews with domain experts to identify essential sales drivers and develop hypotheses guiding our analysis. Assessing data quality and completeness was critical in this phase to ensure reliable insights.

During Data Preparation, extensive feature engineering was performed, integrating macroeconomic indicators and seasonal factors into our dataset. We addressed missing data through predictive imputation, tailoring features to capture sales trends across various time horizons.

The Modeling phase involved iterative development of machine learning models, starting with benchmark models and refining our approach through feature and model selection. Our goal was to create standardized models optimized for accurate sales forecasting.

In the Evaluation phase, a validation set was strategically defined to assess model performance. We emphasized interpretability of model results, including feature importance analysis, to validate hypotheses about sales drivers. We prioritized meaningful features over complex datasets for effective forecasting.

This structured methodology ensures a systematic and comprehensive approach to sales forecasting that aligns with business objectives. By leveraging the CRISP-DM framework, we aim to deliver actionable insights that contribute to informed decision-making and business success.

3.1 DATA UNDERSTANDING

To delve deeper into comprehending the dataset, we employed descriptive statistics alongside data visualization tools.

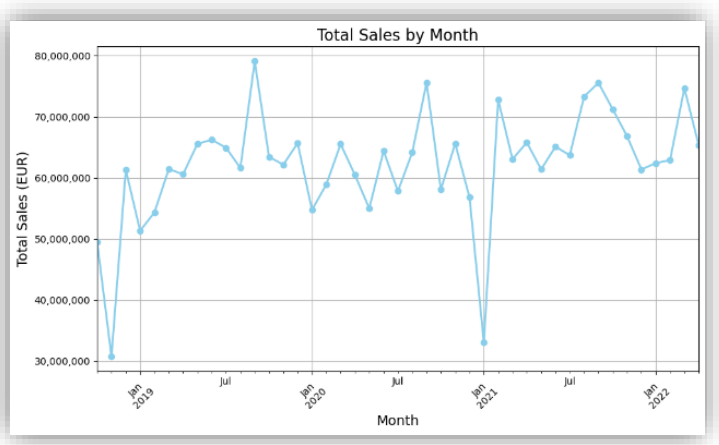
Initially, it was necessary to analyze the datasets separately, both the market dataset and the sales dataset.

The market data comprises 48 variables and 221 rows, including the "*date*", the "*Product Index Machinery and Electricals*" and "*Shipments Index Machinery & Electricals*" for each country, the price of all metals worldwide, and other relevant variables for analysis.

The countries featured in the dataset are Germany, France, Japan, Italy, United Kingdom, Switzerland, China, United States, Europe, and world.

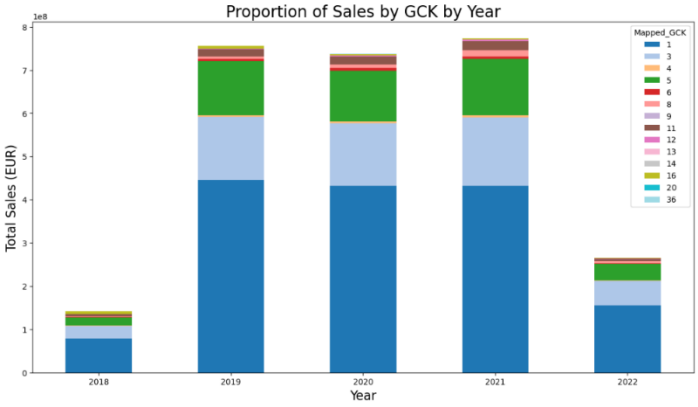
When analyzing the distribution plots of the variables, we can observe graphical similarities between the plots of the "Production Index Machinery and Electricals" and "Shipments Index Machinery and Electricals" for the same country, in most cases as it can be seen in Fig. 1.

The Sales Dataset comprises for 9802 rows but only three variables: "DATE," "Mapped_GCK," and "Sales_EUR." The "DATE" variable indicates the day of the purchase transaction, "Mapped_GCK" denotes the product sold, and the last variable signifies the sale value in euros.



1- Sales by month

We can conclude from the "Sales_EUR" graph that sales fluctuate considerably throughout the year, yet we can identify similar patterns occurring at the same time of the year across different years. We could also observe further into the sales to identify the products that represent the biggest bulk of sales, as depicted in Fig. C.



From this we can clearly see that product 1 consistently holds the largest share. Categories 3 and 5 also contribute significantly to overall sales, maintain sizable shares throughout the period. We also delved into seeing which products most are sold together, these are the pairs (4, 5), (3, 8), (4, 8), (5, 8), as it can be observed in fig e.

After this we checked for seasonality within each product sales with additive decomposition. Additive decomposition is a statistical method widely employed in time series analysis that serves as a crucial tool for dissecting sales data into distinct components. This process aids in discerning underlying patterns pivotal for informed decision-making within businesses.

In essence, additive decomposition dissects the original time series into three fundamental components:

Trend: This component unveils the overarching direction of sales data over time, elucidating whether sales are ascending, descending, or maintaining stability.

Seasonality: By identifying recurring patterns at regular intervals—be it daily, weekly, monthly, or yearly—seasonality highlights the cyclic nature of sales. It offers insights into seasonal peaks or troughs, allowing businesses to anticipate and strategize for fluctuating demand.

Residuals (Noise): Representing the residual variance after trend and seasonal components are extracted, this aspect captures random fluctuations and irregularities inherent in the data. It encapsulates unexplained variations that may influence sales performance unpredictably.

To summarize the overall trend of sales, we constructed a table wherein a positive "Trend Change" value signifies an increasing trend in sales over the analyzed period, while a negative value indicates the opposite:

GCK	TREND	TREND CHANGE	SEASONALITY
1	Positive	1870332.1826111898	No
3	Positive	3262564.043488236	Yes
4	Negative	-42991.90045697382	Yes
5	Positive	2727421.258366838	Yes
6	Positive	23349.5437359169	No
8	Positive	1219909.3281764756	Yes
9	Positive	1753.7272625291416	Yes
11	Positive	851240.2255103914	Yes
12	Positive	419127.6296935704	No
13	Negative	-6350.018076923094	No
14	Positive	315.08519862081084	Yes
16	Negative	-1156737.7506468536	No
20	Negative	-2324.8664607614596	Yes
36	Positive	30140.968428030315	Yes

3.2 DATA PREPARATION

In the market dataset, missing values were identified in the following variables: "Switzerland:Production Index Machinery & Electricals", "Switzerland:Shipments Index Machinery & Electricals," "UK:Shipments Index Machinery & Electricals," "US:Shipments Index Machinery & Electricals", "Producer Prices: United Kingdom: Electrical equipment", "Producer Prices: France: Electrical equipment", "Producer Prices: China: Electrical equipment", "production index: Switzerland: Machinery and equipment n.e.c.", "production index: World: Electrical equipment" and "production index: Switzerland: Electrical equipment." To address these missing values, we applied predictions from a linear regression model with a specific parameter that interpolates values in both forward and backward direction.

However, the Sales dataset does not contain variables with missing values, thus requiring no further processing.

Regarding duplicate values, after analyzing both datasets, it was determined that neither dataset contains any duplicate values. Therefore, there is no need to remove any values from the datasets.

3.2.1 Correlations

After examining the correlation among market features, we found numerous variables with high correlations, which did not offer any new insights (Fig d.). To streamline our models and eliminate redundancy, we decided to exclude the following variables from our analysis:

- 'China: Shipments Index Machinery & Electricals'
- 'Production Index: France: Machinery and Equipment n.e.c.'
- 'Production Index: Germany: Machinery and Equipment n.e.c.'
- 'Production Index: Italy: Machinery and Equipment n.e.c.'
- 'Japan: Shipments Index Machinery & Electricals'
- 'Production Index: Japan: Machinery and Equipment n.e.c.'
- 'Switzerland: Shipments Index Machinery & Electricals'
- 'Production Index: United Kingdom: Machinery and Equipment n.e.c.'
- 'Production Index: United States: Machinery and Equipment n.e.c.'
- 'World: Price of Base Metals'
- 'World: Price of Crude Oil, Average'
- 'World: Price of Copper'
- 'Producer Prices: United States: Electrical Equipment'
- 'Producer Prices: Italy: Electrical Equipment'
- 'Producer Prices: United Kingdom: Electrical Equipment'

3.2.2 Feature engineering

In our initial phase of feature engineering, we conducted an analysis using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) on the Sales_EUR data.

The ACF serves to gauge the correlation between a series and its lagged values at different time intervals, shedding light on how each sale correlates with its past sales over varying temporal distances. This analysis assists the model in identifying any underlying seasonality or trend within the data.

Similarly, the PACF examines the correlation between a series and its lagged values while eliminating the influence of intermediate lags, thereby focusing solely on the direct relationship between the series and its lagged values. This enables us to pinpoint the direct impact of past observations on the current observation.

Following a thorough examination of the ACF and PACF results, we opted to incorporate 8 lag features for the sales of each GCK. These represent the historical values of the series at preceding time points. By integrating lag features into our analysis, we aim to capture the temporal dependencies and patterns inherent in the data, thereby enhancing our models' ability to learn from past observations and make more accurate predictions.

We then applied the Moving Average Convergence Divergence (MACD indicator to our dataset.

The MACD is typically utilized in financial analysis to identify changes in the strength, direction, momentum, and duration of a price trend.

The MACD is calculated by computing the difference between two exponential moving averages (EMAs) of an asset's price. Specifically, it involves:

1. Calculating the 12-month EMA (Exponential Moving Average), which gives more weight to recent prices over the last 12 months.
2. Calculating the 26-month EMA, which provides a longer-term average over the last 26 months.
3. Subtracting the 26-month EMA from the 12-month EMA to derive the MACD line.
4. Creating a signal line, typically a 9-month EMA of the MACD line, to identify potential buy or sell signals.
5. The MACD histogram is then generated by plotting the difference between the MACD line and the signal line.

By adapting the MACD calculation to fit monthly data instead of daily data, we tailor the analysis to the temporal scale of our dataset, providing insights into long-term trends and fluctuations in sales performance across different product categories. This allows us to identify potential turning points or shifts in sales patterns, aiding in decision-making processes such as forecasting, strategic planning, and risk management.

we further enriched our dataset by incorporating time-related variables associated with significant periods such as the COVID-19 pandemic, summer vacations in Germany, and the quarters of the year.

Quarter of the Year Feature:

We introduced a new feature called "Quarter," which categorizes each observation based on the quarter of the year it belongs to. This allows us to analyze sales trends and patterns across different quarters, potentially identifying seasonal variations or cyclic behaviors.

COVID-19 Period Indicator:

We defined the start and end dates of the COVID-19 period and created a binary indicator called "during_covid." This variable signifies whether a sale occurred during the COVID-19 period, enabling us to assess the impact of the pandemic on sales performance and behavior.

Summer Vacation Periods in Germany:

We identified the summer vacation periods for Germany between 2018 and 2023 and created a new feature called "during_summer_vacation." This binary indicator determines whether a sale was made during the summer vacation period in Germany. Understanding sales dynamics during vacation periods can provide insights into consumer behavior and demand fluctuations during leisure seasons.

Lastly, we proceeded to enhance our dataset further by creating lag features for market-related variables. For each feature in the list, we generated lagged versions spanning six months, aiming to investigate potential correlations between past market indicators and current sales levels.

The process involved shifting each market-related feature backward in time by up to six months, effectively creating lagged representations of the original variables. By incorporating historical data points, we aimed to capture temporal dependencies and assess whether past market conditions influence present sales performance more significantly than current conditions.

3.2.3 Feature Selection

For our feature selection, we initially partitioned our original data frame into 14 distinct frames, each corresponding to a specific product group (GCK). Subsequently, we applied feature selection techniques to each of these frames.

Feature selection endeavors to pinpoint the most pertinent variables for model evaluation. To accomplish this, we employed various methodologies, applying them across the different product groups. Initially, we examined all lag features created and assessed their correlation with the target variable to determine if any lag feature exhibited stronger correlation than the current one. Following this initial selection process, we evaluated the correlation among all features to identify any redundancies.

Subsequently, we employed two distinct models capable of elucidating feature importance: the XGBoost Regressor, which gauges information gain, and the Random Forest Regressor, which assesses via entropy. Upon executing these models, we compiled a list of features that represented the intersection of both models' selected features, ensuring that these features are indeed the most relevant. We got the following results for each of the product categories:

GCK	Features Selected
1	'12 Month EMA','production index: Germany: Electrical equipment_lag_3','MACD Histogram','production index: United Kingdom: Electrical equipment_lag_1'
3	'Signal Line', 'MACD', 'production index: World: Electrical equipment','Italy:Production Index Machinery & Electricals_lag_3','China:Production Index Machinery & Electricals_lag_3'
4	'12 Month EMA', 'MACD Histogram', '26 Month EMA','production index: Switzerland: Machinery and equipment n.e.c._lag_6'
5	'US:Shipments Index Machinery & Electricals_lag_3','MACD Histogram','France:Production Index Machinery & Electricals_lag_1','production index: Japan: Electrical equipment_lag_1'
6	'Italy:Shipments Index Machinery & Electricals_lag_1','Switzerland:Production Index Machinery & Electricals_lag_6','MACD','MACD Histogram', 'US:Production Index Machinery & Electricals','France:Production Index Machinery & Electricals_lag_5','UK:Shipments Index Machinery & Electricals'
8	'12 Month EMA','MACD Histogram','production index: United Kingdom: Electrical equipment_lag_6', 'World: Price of Natural gas index','Sales_EUR_Lag_3'
9	'12 Month EMA','Sales_EUR_Lag_6','Sales_EUR_Lag_2', 'MACD','MACD Histogram','production index: Japan:Electrical equipment','Sales_EUR_Lag_8'
11	'MACD Histogram', 'MACD'
12	'MACD Histogram','12 Month EMA','MACD','Sales_EUR_Lag_6', 'Italy:Shipments Index Machinery & Electricals','UK:Production Index Machinery & Electricals','World: Price of Energy'
13	'MACD Histogram','production index: Switzerland: Electrical equipment','Sales_EUR_Lag_3','production index: Switzerland: Machinery and equipment n.e.c.','Sales_EUR_Lag_7','China:Production Index Machinery & Electricals_lag_2','MACD','production index: Germany: Electrical equipment_lag_5','production index: World: Electrical equipment_lag_4','Signal Line'
14	'MACD Histogram','MACD','Europe:Production Index Machinery & Electricals_lag_4','UK:Production Index Machinery & Electricals_lag_4','production index: Japan: Electrical equipment', 'production index: Germany: Electrical equipment_lag_4'
16	'Producer Prices: Germany: Electrical equipment_lag_6','MACD', '12 Month EMA'
20	'MACD Histogram','Producer Prices: Germany: Electrical equipment_lag_4','World: Price of Energy_lag_6','12 Month EMA'
36	'MACD Histogram','Europe:Production Index Machinery & Electricals_lag_2','Producer Prices: China: Electrical equipment_lag_5','US:Shipments Index Machinery & Electricals_lag_5', 'Sales_EUR_Lag_2','Italy:Shipments Index Machinery & Electricals_lag_3'

3.3- Explaining the Models

XGBRegressor

We decided to use a regression model because for sales forecasting tasks it is ideal due to the tailored support for regression, making it well-suited for predicting continuous sales values. With regression-specific loss functions and evaluation metrics, *XGBRegressor* offers efficient optimization and accurate predictions. Its performance capabilities, interpretability, and clear intent enhance the modeling process, resulting in reliable sales forecasts.

The *extreme Gradient Boosting* is based on the concept of gradient boosting, which is an ensemble learning technique that builds a series of weak learners (typically decision trees) in a sequential manner. Each new tree is trained to correct the errors made by the previously built trees. These are built recursively by splitting the data into partitions based on feature values to minimize the error in the predicted target values. In this case for regression measures we used (**XGBRegressor** with **objective='reg:squarederror'**) the typical loss function, the mean squared error (MSE).

Random Forest

RandomForestRegressor is well-suited for sales forecasting due to its ability to capture complex relationships and non-linear patterns in sales data. As an ensemble learning method, random forests combine multiple decision trees to produce robust and accurate predictions, handling both numerical and categorical data effectively. They are less prone to overfitting, provide insights into feature importance, and deliver competitive performance in terms of predictive accuracy. With these qualities, RandomForestRegressor is a versatile and reliable choice for modeling and predicting sales based on diverse input factors.

RandomForestRegressor builds multiple decision trees independently using random subsets of the training data and features. Each decision tree is trained on a bootstrapped sample of the data and considers only a subset of features at each node. The trees are diverse and less correlated, which helps in reducing overfitting. During prediction, new data samples are passed through each tree, and the final prediction is made by averaging the predictions of all trees.

Linear Regression

Linear regression is a simple, yet powerful technique used for modeling the relationship between independent variables (features) and a dependent variable (target) by fitting a linear equation to the observed data. It aims to find the best-fitting linear relationship between the features and the target variable. The model estimates coefficients (weights) for each feature, indicating the strength and direction of their influence on the target, then predictions are made by applying the learned linear equation to new data points. This type of regression model fits well having notice that we are predicting sales.

KNN

The *k-Nearest Neighbors (k-NN)* algorithm is a simple and intuitive machine learning method used for regression and classification tasks based on similarity measures between data points.

predicts the target value of a new data point based on the average (or weighted average) of the target values of its k nearest neighbors in the feature space. The choice of n_neighbors (number of neighbors) and *weights* (weighting scheme based on distance) can significantly impact the model's predictive performance.

Ensemble (Voting)

VotingRegressor ensemble model aggregates predictions from multiple base regression models (*XGBRegressor, RandomForestRegressor, LinearRegression, KNeighborsRegressor*) to make collective predictions on unseen data. By combining diverse modeling approaches, the ensemble aims to achieve better predictive performance compared to individual models alone.

Each base estimator (**model1**, **model1_RF**, **model1_LR**, **model1_KNN**) contributes to the ensemble's prediction based on its individual performance. The ensemble leverages the diversity of different models to potentially improve overall prediction accuracy and robustness. The choice of voting strategy (**voting='hard'** or **voting='soft'**) can impact how predictions are combined **'hard'**: Predictions are based on majority voting.

Sarimax

SARIMAX (Seasonal Autoregressive Integrated Moving Average with exogenous regressors) is a popular and powerful time series forecasting model that extends the capabilities of the *ARIMA* (Autoregressive Integrated Moving Average) model to include seasonal components and external regressors. *SARIMAX* includes an autoregressive (AR) component, which captures the relationship between the target variable and its lagged values. This component reflects how past observations in the time series influence the current value.

3.4 MODELING

To obtain the RMSEs associated with each of the GCKs, we applied *XGBoost*, *Random Forest*, *Linear Regression*, and *KNN*, all of which were discussed in section 3.2. Additionally, we applied an ensemble of these models to all the GCKs. Furthermore, we also applied the *SARIMAX* model exclusively to GCK numbers 4 and 5. To facilitate the prediction process, we used SARIMAX to forecast the values of indicators in the temporal space for which we are making predictions, as these values were not readily available. Consequently, we obtained five different RMSEs for each GCK, and six for GCK 4 and 5. Subsequently, we proceeded to select the model with the lowest RMSE value for each of them.

Model	Mapped_GCK
KNN	1
Linear Regression	3
Linear Regression	4
Linear Regression	5
Random Forest	6
Random Forest	8
Linear Regression	9
Random Forest	11
Linear Regression	12
Linear Regression	13
Linear Regression	14
XGBoost	16
Ensemble	20
Random Forest	36

4. RESULTS EVALUATION

After implementing the best models for each GCK as mentioned in section 3.3, we obtained the following results for each of them:

Model	RMSE	Mapped_GCK
KNN	2574299,846	1
Linear Regression	1071342,997	3
Linear Regression	111325,6246	4
Linear Regression	2718332,028	5
Random Forest	113122,5668	6
Random Forest	303534,0265	8
Linear Regression	7575,82085	9
Random Forest	1094146,122	11
Linear Regression	82851,78045	12
Linear Regression	11282,3649	13
Linear Regression	14084,54736	14
XGBoost	65993,26114	16
Ensemble	1104,632418	20
Random Forest	10337,89908	36

The sales forecast can also be consulted in figure c of the index, which illustrates all the charts for each of its GCKs regarding sales. In these representations, the blue line represents the sales data provided, while the dashed orange line represents the respective forecasts.

Considering Siemens' sales volume, these results appear to be promising regarding a stable future trend line, with no occurrence of abnormal factors that could compromise the company's sales capacity.

5. DEPLOYMENT AND MAINTENANCE PLANS

A strategic approach is required for the sales forecasting model to be deployed to guarantee th at stakeholders are interacting with and applying it effectively. Firstly, the model will be presented to key stakeholders, to ensure that it is readily accessible for decision-making purposes, to provide a comprehensive understanding of its functionality and to present its benefits. There will be training sessions to familiarize stakeholders with the model's interface, features, and interpretation of results. To assure that no doubts arise, in the future, among stakeholders it is essential to have a reference guide. This guide should outline the model's methodology, including data sources, algorithms employed, model's outputs, interpretation guidelines and possible assumptions that were made.

To ensure its ongoing relevance, accuracy, and effectiveness in supporting business objectives is essential doing maintenance. Updates and enhancements will be implemented in response to changes in market dynamics, customer behaviour, or internal business processes. Quality assurance protocols will be established to validate data integrity and consistency, mitigating the risk of errors or discrepancies in forecasting outputs. Iterative improvements to the model will be possible through the establishment of constant feedback loops that will collect end-user insights and suggestions for improvement.

6. CONCLUSIONS

In conclusion, the development and implementation of a sales forecasting model for Siemens represents a great advantage in planning, management and enhancing decision-making capabilities by having information about potential sales values.

By leveraging historical data, advanced algorithms, and market insights through predictive modelling techniques, Siemens gains valuable knowledge into future demand trends and with that it will be possible to allocate resources in a strategic way, inventory management, and generate an optimization of revenue.

Also, through comprehensive documentation and stakeholder engagement initiatives, Siemens ensures the smooth deployment and effective utilization of the forecasting model across the organization.

Ultimately, the implementation of the sales forecasting model is a key instrument that enables Siemens to successfully through market dynamics, developing operational excellence and solidifying its position as a leader in the intelligent infrastructure solutions sector.

6.2 CONSIDERATIONS FOR MODEL IMPROVEMENT

There are some key considerations and strategies for enhancing the sales forecasting model, comprehensive methods such as implementing correlation of product sales, using grid search for model optimization, and exploring alternative feature selection metrics.

Implementing Correlation of Product Sales: Exploring the correlation between the sales of different products could provide valuable insights for model improvement. The forecasting model can more accurately identify interdependencies and adjust predictions by examining the effects of one product's sales on another.

Utilizing Grid Search for Model Optimization: Grid search involves searching through a specified parameter space to identify the optimal combination of parameters that yield the best performance. The use of grid search method in each tested model provides an approach to optimize model parameters, improve prediction accuracy, and achieve better forecasting results.

Exploring Alternative Feature Selection Metrics: Effective forecasting models require careful consideration of features. By considering more feature selection methods such as mutual information, or recursive feature elimination it will be possible to select the most relevant features in a way that increases the predictive power of the model and its robustness

7. APPENDIX

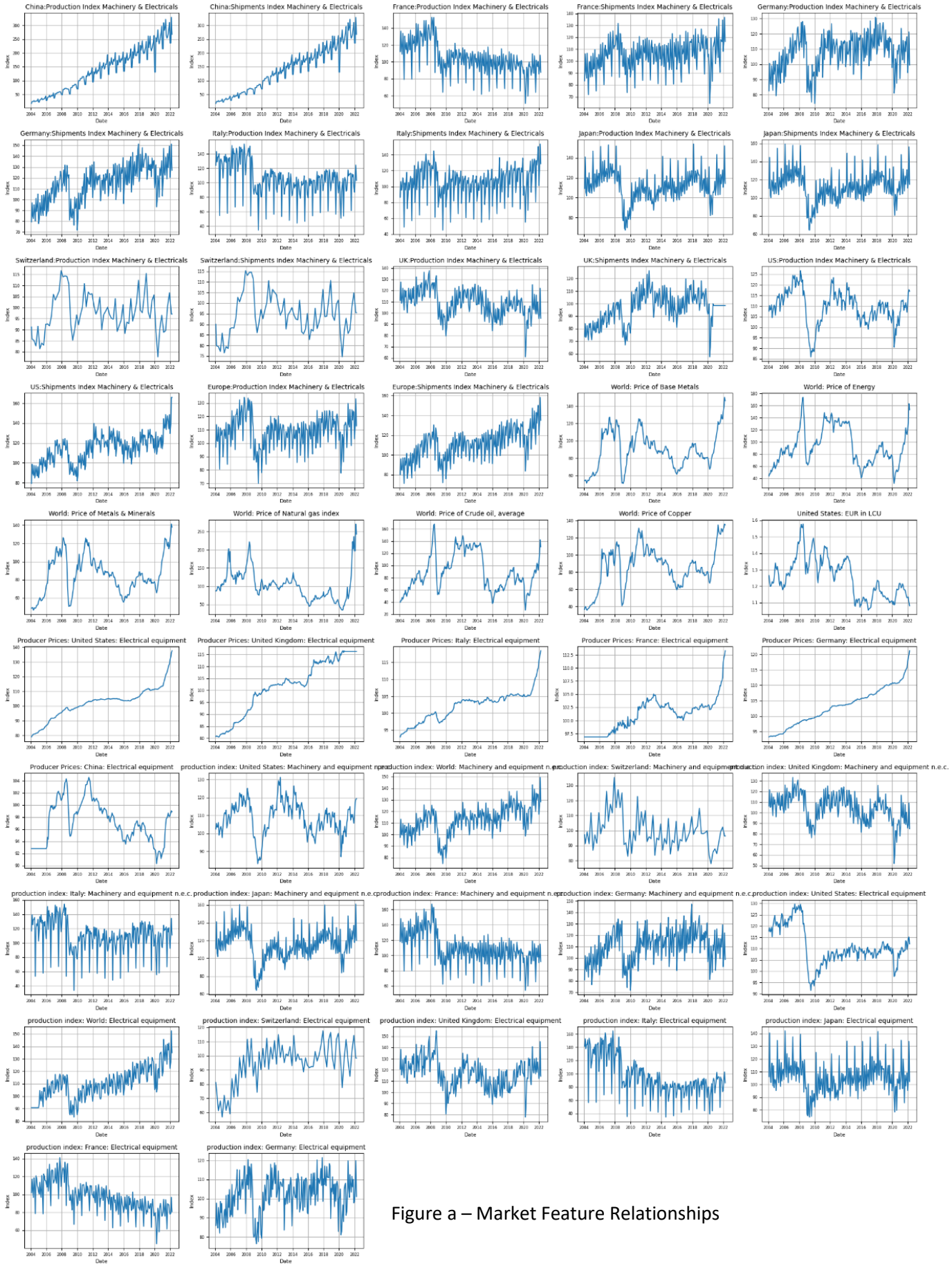


Figure a – Market Feature Relationships

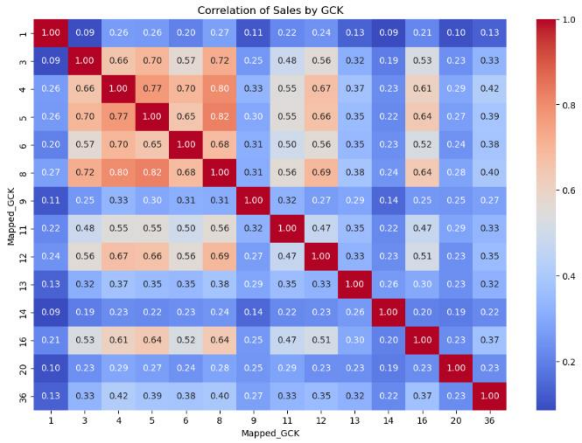
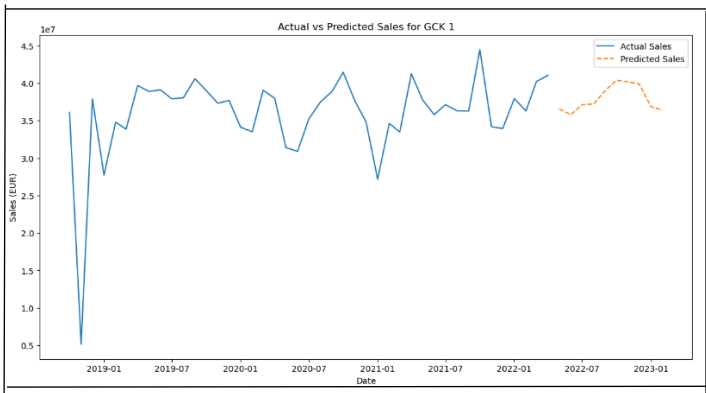
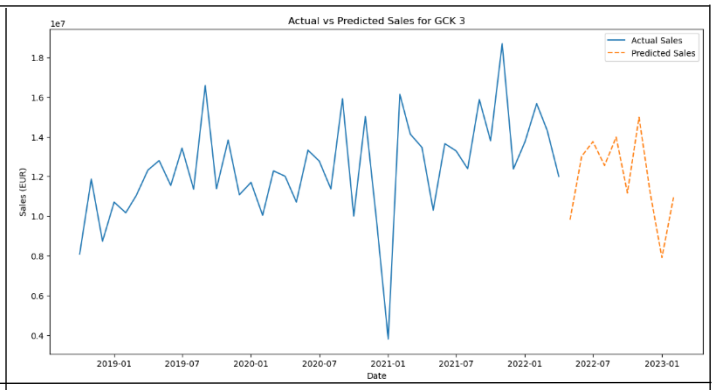
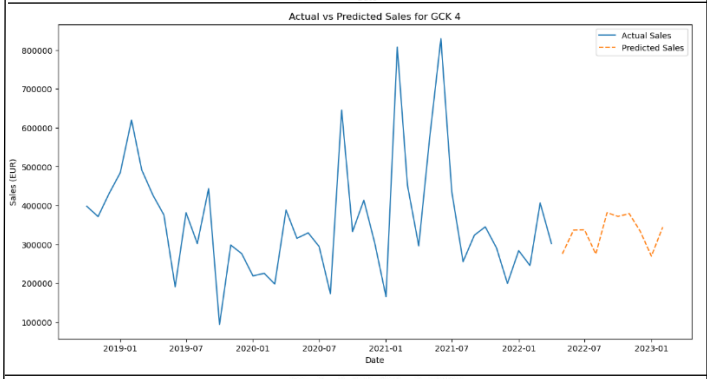
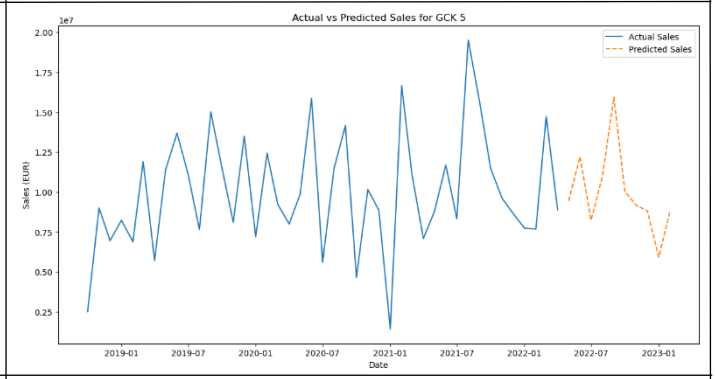


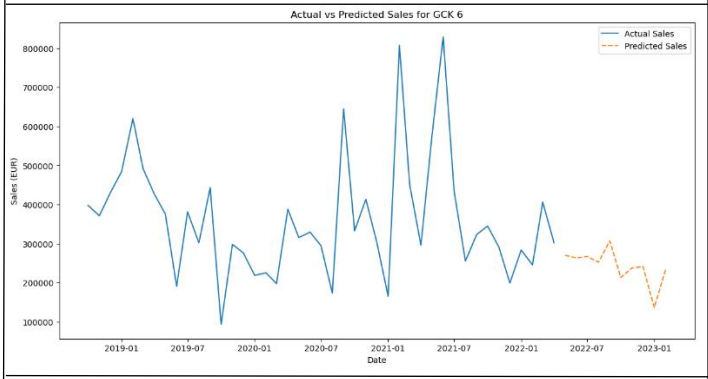
Figure b- Correlated product sales

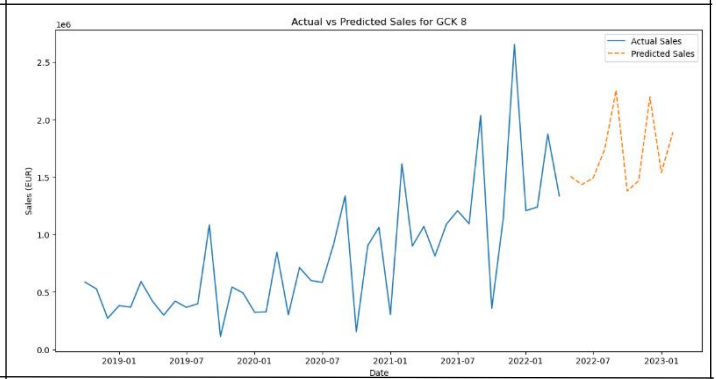

Actual vs Predicted Sales for GCK 1
Actual Sales
Predicted Sales
Sales (EUR)
Date

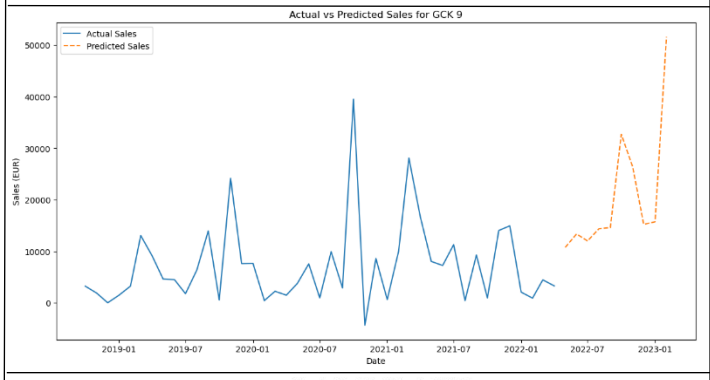

Actual vs Predicted Sales for GCK 3
Actual Sales
Predicted Sales
Sales (EUR)
Date

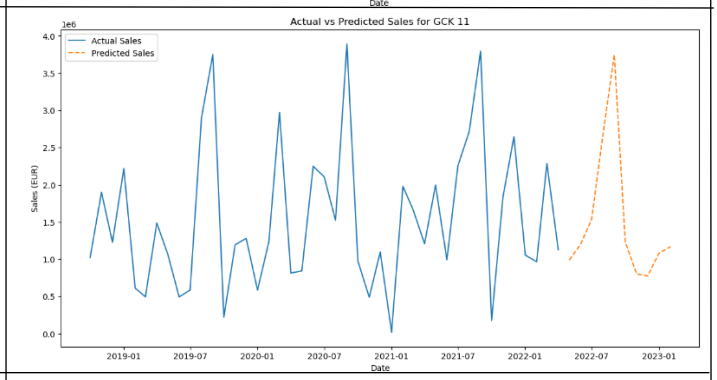

Actual vs Predicted Sales for GCK 4
Actual Sales
Predicted Sales
Sales (EUR)
Date

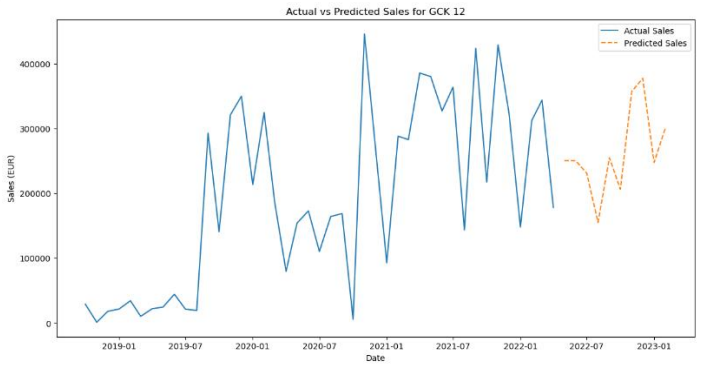

Actual vs Predicted Sales for GCK 5
Actual Sales
Predicted Sales
Sales (EUR)
Date

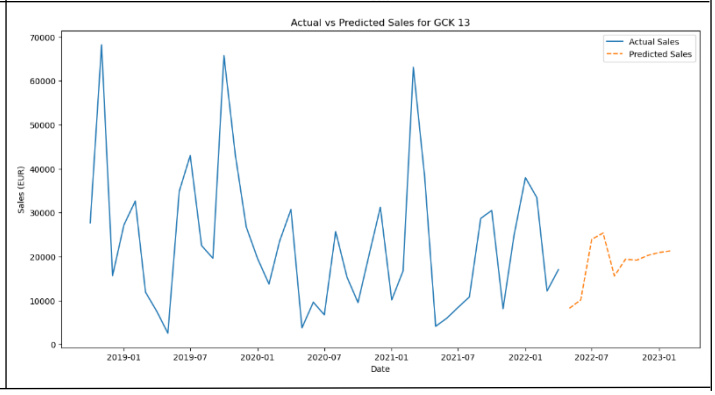

Actual vs Predicted Sales for GCK 6
Actual Sales
Predicted Sales
Sales (EUR)
Date


Actual vs Predicted Sales for GCK 8
Actual Sales
Predicted Sales
Sales (EUR)
Date


Actual vs Predicted Sales for GCK 9
Actual Sales
Predicted Sales
Sales (EUR)
Date


Actual vs Predicted Sales for GCK 11
Actual Sales
Predicted Sales
Sales (EUR)
Date


Actual vs Predicted Sales for GCK 12
Actual Sales
Predicted Sales
Sales (EUR)
Date


Actual vs Predicted Sales for GCK 13
Actual Sales
Predicted Sales
Sales (EUR)
Date

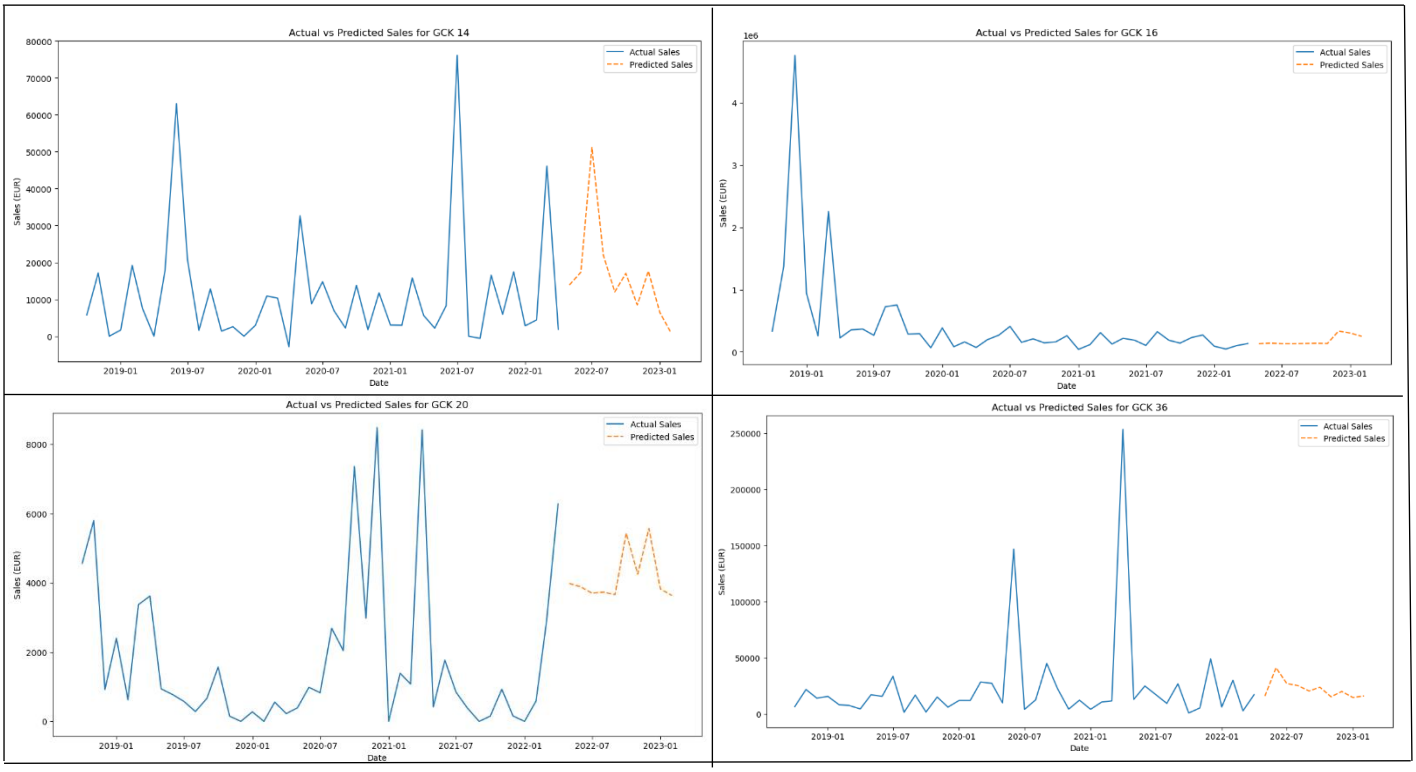


Figure c – Sales forecast for each GCK