ARCADA

Car sales analysis in the Nordic Countries

Martinez, Luis

Master's Thesis

Master of Engineering - Big Data Analytics

Spring 2023

MASTER'S THESIS	
Arcada University of Applied Sciences	
Degree Programme:	Master of Engineering - Big Data Analytics
Identification number:	9202
Author:	Martinez, Luis
Title:	Car sales analysis in the Nordic Countries
Supervisor (Arcada):	Magnus Westerlund
Commissioned by:	Arcada University of Applied Sciences

Abstract:

Sales forecasting is an essential component of business intelligence and, artificial intelligence and predictive analytics are now essential tools for companies to predict market trends and forecast sales volumes.

In the automotive industry, where production processes are extremely complex and the logistic operations until the products are retailed and handed over to end customers are very long, the ability to predict future demand and accommodate such prediction to production planning and supply chain processes is even more crucial than in any other industry.

In this work, some classical time-series forecasting techniques are compared with more advanced machine learning methods analyzing their performance in predicting the number of sales of a carmaker in the Nordic countries and studying the correlation between socio-economic indicators and the volume of sales for this specific brand.

Multiple experiments are conducted and evaluated following a quantitative approach and the effectiveness of the models is scrutinized using different performance metrics leading to the conclusion that the introduction of exogenous variables as inputs for ML algorithms can improve the forecasting results overcoming traditional models in the 4 Nordic countries analyzed.

Keywords:	Regression analysis · Machine Learning · Ensemble · Boosting · Bagging · Time Series · Forecasting · Statsmodels · Triple Exponential Smoothing · SARIMA · SARIMAX · Random Forest · eXtreme Gradient Boosting · Light Gradient Boosting Machine · Adaptative Boosting · Extra Tree Regressor · Automotive · Car sales · Feature Engineering · Feature Selection
Number of pages:	**62**
Language:	English
Date of acceptance:	30.05.2023

TABLE OF CONTENTS

FIGURES

TABLES

ABBREVIATIONS

ML	Machine Learning
MTO	Make to Order
MTS	Make to Stock
AI	Artificial Intelligence
JIT	Just in Time
CPI	Consumer Price Index
UR	Unemployment Rate
TIV	Total Industry Volume
MS	Market Share
PC	Passenger Cars
LCV	Light Commercial Vehicles
FS	Feature Selection
FE	Feature Engineering
ADF	Augmented Dickey-Fuller
MA	Moving Average

FOREWORD

This thesis was written for my master's degree in Big Data Analytics at Arcada University of Applied Sciences in Helsinki (Finland) and its subject is related to the analysis of vehicle sales in the Nordic countries – Finland, Sweden, Norway, and Denmark - based on historical data from a particular car manufacturer in combination with the total number of car registrations plus some socio-economic indicators to find a correlation between the selected features and gain some prediction ability based on the models analyzed.

This is a very fascinating research topic as it is a blend of the technical knowledge acquired during the studies and the professional experience in the field that I have been professionally connected to for most of my professional career: the automotive industry.

After thanking my entire family and, especially my wife and my parents for their continuous support and my dearest son for his endless patience, I would like also to thank my supervisor (*Dr. Magnus Westerlund*) for his guidance, encouragement, and recommendations, and the rest of lecturers for their knowledge sharing and their contribution to this work.

Luis Martinez
April 2023

1 INTRODUCTION

1.1 Background

The capacity to forecast the volume of sales has historically been a key asset for every company in any kind of industry but, especially, in the automotive sector, where the production processes are extremely complex, the ability to anticipate consumer demand becomes very valuable information that can be especially useful for car manufacturers to accommodate their production plans and the associated logistic operations and, consequently, to fit consumers' needs and achieve the optimal stock levels.

Compared to other industries with lower-value products and shorter *production-to-customer* lead times, car manufacturers must be highly responsive to customers' demands, which is a significant challenge in today's turbulent business environment (Kobayashi, Tomino, Shintaku, & YoungWon, 2017). An effective demand-and-supply chain, utilizing both make-to-stock (MTS) and make-to-order (MTO) approaches, and keeping the right balance between the two is essential for automakers to cope with the demand fluctuation and, artificial intelligence (AI) and, more concretely, machine learning (ML) algorithms can contribute to extract the necessary business knowledge supporting data-driven decisions to keep such balance in the right spot.

An MTO approach marks a shift in how raw materials are sourced. Rather than maintaining a large inventory of parts in storage, a make-to-order production strategy requires smooth management of just-in-time (JIT) delivery, where the facility receives goods as close as possible to when they are needed, improving efficiency, and reducing costs. Moreover, this kind of production, offers a more personalized shopping experience, making the product more customizable.

On the other side, an MTS production methodology contributes to a steady production workflow minimizing customer waiting times but, on the downside, generates inappropriate inventory levels (sometimes too high and sometimes too

low) due to the unpredictable nature of consumer demand which may, impact the company's cash flow.

1.2 Objective

This thesis aims to evaluate the performance of traditional statistical models - such as Triple Exponential Smoothing (TEST), SARIMA, or SARIMAX – and compare them to more complex ML algorithms based on ensemble techniques with the final goal of getting a better understanding of the automotive market in the four Nordic countries.

1.3 Research questions

The main research questions identified as part of this study are:

1. Is there any correlation between the analyzed external socio-economic factors and the volume of car sales in the Nordic countries?
2. Is there a model good enough to predict the volume of sales and the market trend in the Nordic countries?
3. Are ML algorithms performing better than traditional algorithms for car sales prediction?
4. Are the Nordic markets affecting each other in terms of volume of car sales?

1.4 Limitations

The automotive market is very changing and volatile, and the volume of car sales can be affected by an uncountable number of external factors, some of them may be macroeconomic, such as gross domestic product (GDP), inflation, interest rates or unemployment rates (Islam, Bashawir Abdul Ghani, Kusuma, & Teh Yew Ho, 2016) but, probably, also by marketing campaigns, consumer demand changes, technological aspects or even by unpredictable elements such as pandemics or military conflicts making the forecasting of car sales a very complex and challenging task.

This paper is focused on the analysis of the volume of car sales in the Nordic markets of a specific automaker but, by no means, is the aim of this research to provide a methodology on how to predict car sales volumes in a wider term.

2 LITERATURE REVIEW

Sales forecasting is a deeply studied field and researchers have made a lot of efforts to find accurate ways to predict sales volumes, not only in the automotive industry, but in all business fields, and, many of them, have already tried to predict the number of car sales based on socio-economic indicators but, in most of those studies, the source of data came from *country-specific* numbers rather than *brand-specific* and, the only ones using *brand-specific* sales' volumes are primary focused on Asian, African or US markets but very limited works have been focused on European markets (Homolka, Vu Minh, Drahomíra, Bach, & Dehning, 2020) and there are almost non-existent publications with the spotlight in the Nordic region of Europe.

For instance, in the research for Chery, a Chinese domestic brand, an econometric model is proposed to analyze the dynamic connections among Chinese automobile sales, the number of brand-specific sales, and some economic variables such as CCI, CPI, steel production, and gas prices (Junjie, Yanan, Xiaomin, Han, & Feng, 2018). In that study, they concluded that the fluctuation trends of Chinese automobile sales and Chery's sales are intricately linked, and they also found out that, among the models analyzed, VAR, ARMA, and VECM, the last one offered the best prediction accuracy in terms of RMSE and MAPE in long-term forecasting.

On the other side, other *"non-brand specific"* studies also focused on predicting automotive sales (Sa-ngasoongsong, Bukkapatnam, Kim, S. Iyer, & Suresh, 2012) or (Islam, Bashawir Abdul Ghani, Kusuma, & Teh Yew Ho, 2016), identified a meaningful structural relationship between sales and some economic indicators while some other researchers concluded that the macroeconomic variables do not always influence the sales for all the countries (Sanjog & Shoaib, 2022) and, therefore, there is still room for experimentation in this field.

Moreover, most of the papers published related to sales forecasting are exclusively comparing a few models: either statistical (Makatjane & Ntebogang, 2016) or more advanced ML algorithms (Baržić, Munitić, Bronić, Jelić, & Lešić, 2022) but very few papers combine multiple models of each class in one single publication.

3 DATA

In this paper, a dataset from a car manufacturer containing the number of cars retailed (sold and delivered to end-customer) in the Nordic countries between January 2006 and December 2022 is used for the analysis containing 204 data points whose distribution is shown in the picture below:

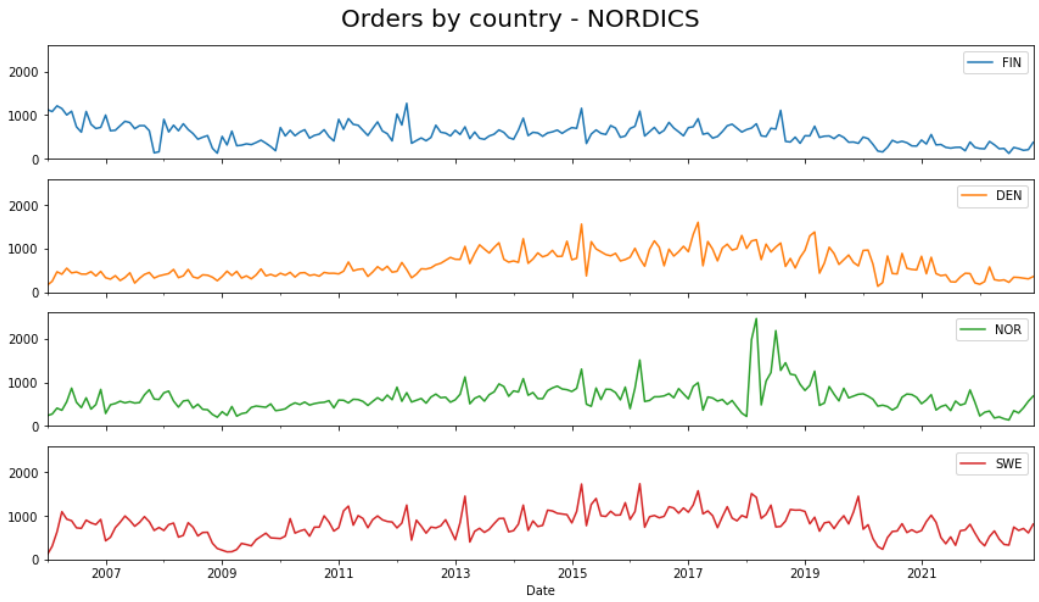


Figure 1: Retails by country between 2006 to 2022

Furthermore, this paper includes an additional dataset containing all the passenger cars (PC) and light commercial vehicles (LCV) registered in the Nordic countries during the same period. This set of data, also known as total industry volume or TIV, is often utilized to provide one of the most extended key performance metrics (KPI) in the automotive industry: the market share (MS). This measure gives an insight into the automaker's performance versus the competitors. The source of this dataset came from the same car manufacturer whose identity will remain hidden to preserve the confidentiality of the data used.

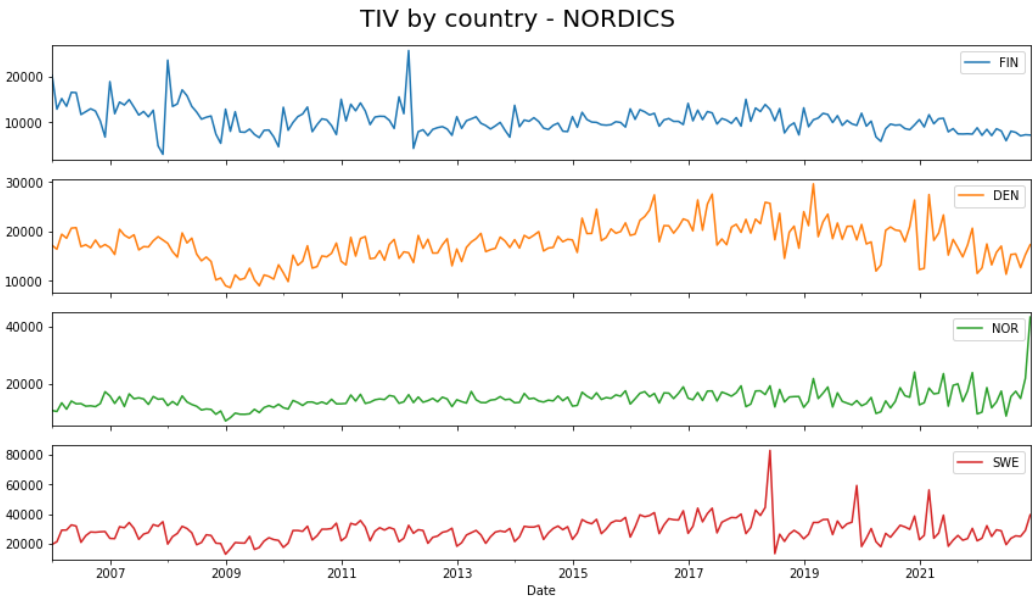


Figure 2: Registrations by country between 2006 to 2022

On the other hand, some exogenous features are incorporated as socio-economic external indicators that could, eventually, have an impact on the volume of automobile sales. These indicators are the consumer price index (CPI), the unemployment rate, and the long-term interest rate in each of the Nordic countries between 2006 and 2022.

The Consumer Price Index (CPI) (OECD, Inflation (CPI) (indicator), 2023) measures the monthly change in prices paid by urban consumers for a market basket of consumer goods and services. As every consumer has differing tastes and spending habits, the CPI measures prices for a huge assortment of items, not only goods but also services (e.g., hairdressing, taxi fares, insurance, etc.). This collection of items is normally referred to as the basket of goods and services. ("What is the CPI - CSO - Central Statistics Office")

The CPI is one of the most popular measures of inflation which is, a priory, a key factor in car sales because, the higher the inflation is, the less disposable income to spend on a new car when the potential consumers are trying to cover the costs of everyday essentials such as food and housing.

The **unemployment rate** (OECD, Unemployment rate (indicator), 2023) is, according to the Organisation for Economic Co-operation and Development, a measure of

people of working age who are without work, are available for work and have taken specific steps to find work. This indicator is measured in the number of unemployed people as a percentage of the total labor force which is defined as the total number of unemployed people plus those in employment. Everything seems to indicate that the higher the unemployment rate the fewer cars will be sold due to less consumer capacity to afford the cost of the car or to apply for a loan (most cars are purchased using finance).

Long-term interest rates (OECD, Long-term interest rates (indicator), 2023) are determined by the price charged by the lender, the risk from the borrower, and the fall in the capital value. Rising interest rates mean higher loan costs when customers go to buy a car and, according to new statistical data, one-fifth (19.2%) of consumer credit granted by credit institutions to households at the end of April 2021 were vehicle loans (Suomen Pankki, 2021). A priory, it makes sense to anticipate that higher interest rates will mean lower car sales.

All these additional datasets have been collected from the Organisation for Economic Co-operation and Development (https://www.oecd.org/) and their access and use are public.

Figure 3: Orders and exogenous features by country

4 MODELLING

4.1 Times series and forecasting

As it can be seen from Figure 2 and Figure 3, the type of data we are dealing with in this paper corresponds to a succession of data points in chronological order measuring the value change of the number of sales over time, which means, that we are dealing with times series data.

Forecasting in the context of time series is, fundamentally, about predicting the future value of the target (or independent) variable based exclusively on its previous observations (autoregression) or by using other predictors (also named exogenous variables).

Initially, ML algorithms were not designed to deal with times series forecasting problems, but recent studies have shown that these types of models can, not only be used for TS forecasting (Bojer, 2022), but they can perform even better than traditional statistical models (Stoll, 2020). However, to make use of ML regression models with time series data is necessary to transform the input values into a matrix that can be used for supervised learning problems by utilizing lagged values of the input variable as the features to predict the target variable or by including external indicators as exogenous features to predict the output variable.

One of the key aspects to consider when using machine learning algorithms for time series forecasting is the type of input data to deal with, which can be divided into two main types:

- Univariate: one single input variable where the objective is to predict its value in the future using only its past values.

- Multivariate: more than one input variable where the objective is to predict the value in the future of the target using a combination of its past values plus some other additional (also known as exogenous) features.

On the other side, the forecasting strategy and the prediction interval are also aspects particularly important when dealing with time series forecasting problems and, based on that we can differentiate:

- Single-step prediction: the objective is to predict the next value of the time series.

- Multi-step prediction: the objective is to predict the next N values of the time series and can be divided into (Brownlee, 2020):

    - Direct. "A separate model is developed to forecast each forecast lead time".

    - Recursive. "A single model is developed to make one-step forecasts, and the model is used recursively where prior forecasts are used as input to forecast the subsequent lead time."

In this study, a combination of univariate and multivariate data and a direct multi-step prediction with 24 steps (months) forecasting horizon is used to evaluate the performance of ML algorithms.

Also, as part of this work, we will compare some traditional statistical algorithms – such as Triple Exponential Smoothing, SARIMA, or SARIMAX with some more advanced ML techniques such as boosting (AdaBoost, XGBoost, Gradient Boosting, and LGBM) and bagging models (Extra Tree and Random Forest).

4.2 Traditional Time Series Models

4.2.1 Triple Exponential Smoothing

Exponential smoothing is one of the most widely used time series forecasting methods for univariate data and it is remarkably like simple moving averages with the main difference that, where simple moving averages consider past observations

equally, exponential smoothing models put exponentially more weight on recent observations and less on historical observations.

“There are three types of exponential smoothing models: simple exponential smoothing, double exponential smoothing, and triple exponential smoothing.” (Shin, 2022)

Single (or simple) exponential smoothing is useful for time-series data with no seasonality or trend, double exponential smoothing is useful for time-series data with no seasonality – but with a trend, and Triple exponential smoothing, also known as Holt-Winters exponential smoothing, is useful for time-series data with a trend and seasonal pattern. As seen during the EDA in previous chapters, we are dealing with trended and seasonal data and, therefore, TES could be, in principle, considered as a suitable candidate.

4.2.2 SARIMA

SARIMA stands for Seasonal Autoregressive Integrated Moving Average, and it is an evolution of the ARIMA model including the seasonal component which, as we saw in the EDA, is present in the data we are working with and, therefore, it is a viable option as a baseline model in our task.

An autoregressive (AR) model assumes that the value of an observation at a time t is a linear combination of the p past sequence values of itself using these lagged values (t-1, t-2, …, t-p) as input variables for its predictions. The order (represented by “p”) determines how many previous data points (“lags”) will be used as predictors and, the higher the p-value is, the more lagged data points the model will consider.

The integrated (I) part of ARIMA models refers to the ability of this kind of model to make the input data stationary which is a pre-requisite of AR and MA models. The model will use the d differences of the time series to predict the value at time t or, in other words, will predict on stationary data (no trends and no seasonality). The number of times the data needs to be differenced will determine the value of “d.”

Finally, the moving average (MA) model considers model errors at previous time steps to improve the prediction for the current data point. In more technical terms, we model the prediction at time t as a linear combination of q residual errors (t-1, t-2, …, t-q).

4.2.3 SARIMAX

SARIMAX stands for Seasonal Auto-Regressive Integrated Moving Average with Exogenous factors, and it is like SARIMA with the only difference that it accepts exogenous factors as input predictors for the model. This means that, on top of the lagged terms, the model will also use external data for the forecast task which, in principle, may improve the model's performance.

4.3 Machine Learning Models

4.3.1 Ensemble Methods

Ensemble learning methods are following a *divide-and-conquer* approach based on the assumption that multiple models combined can achieve a more powerful, accurate, and outperforming result than a single learner. These methods can be divided in:

- Stacking: This technique uses the predictions from multiple different models to build new features which are used as predictors for a new model that makes the final prediction. This technique aims to improve the prediction accuracy.

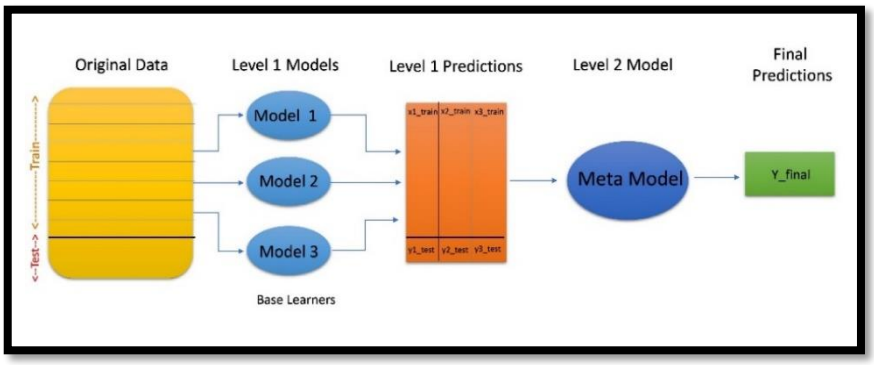


Figure 4: Stacking architecture (Source: OpenGenus)

- Bootstrap aggregating (or bagging): is an ensemble technique where multiple algorithms are trained in parallel using subsamples of the original data and the final prediction comes from the average prediction (or voting mechanism for classification tasks) of all the models. This technique aims to reduce the variance (avoid overfitting).

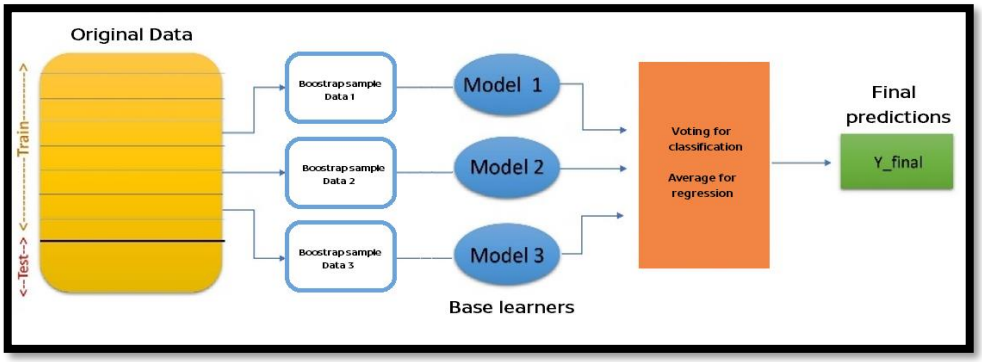


Figure 5: Bagging architecture (Source: Author)

- Boosting: in the boosting technique, models are built sequentially by minimizing the errors from previous models while increasing (or boosting) the influence of the high-performing models. This technique aims to reduce the bias (avoid underfitting).

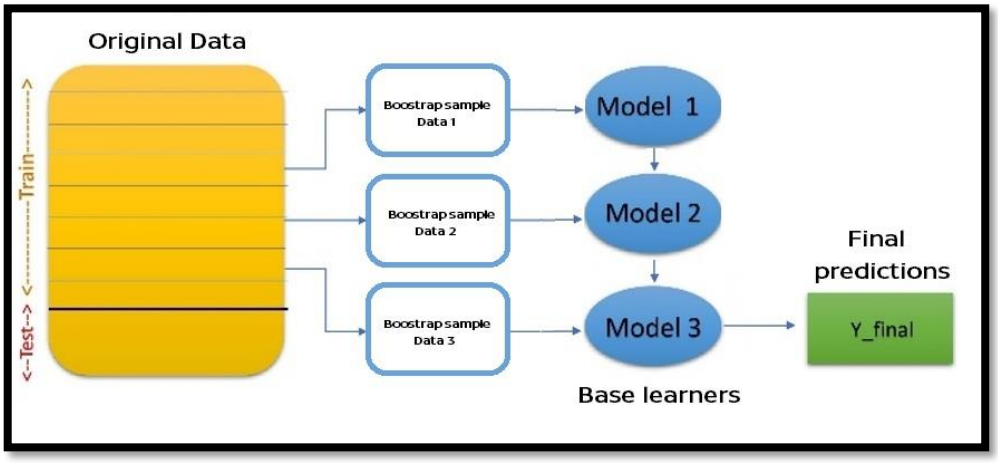


Figure 6: Boosting architecture (Source: Author)

4.3.2 AdaBoost

AdaBoost or Adaptative Boosting (Schapire & Freund, 1997) was the first practical boosting algorithm (Prastiwi & Tahi Ulubalang, 2020) and one of the best-known boosting methods. In this model, each training instance receives a weight that indicates the relative importance which is used to calculate the error. After each iteration, the weights of the instances are adjusted according to the error of the current prediction and the learning focuses on the more difficult cases.

4.3.3 Gradient Boosting Regressor

Another popular boosting ensemble method is Gradient Boosting which is a powerful model able to capture nonlinear relationships between the predictors and the target variable (Masui, 2022). This estimator builds an additive model in a forward stage-wise fashion starting from a naïve prediction of the target variable (i.e., average value) and it improves the prediction by focusing on the prediction errors from previous steps trying to minimize the residuals on each iteration until the model prediction stops improving.

4.3.4 eXtreme Gradient Boosting

XGBoost (Extreme Gradient Boosting) is a scalable end-to-end tree boosting system (Guestrin & Tianqi, 2016) based on the gradient boosting framework that has shown exceptionally reliable, fast, and high-performance results in many ML competitions thanks to the features it offers such as parallelization, cache optimization, auto pruning, and regularization.

4.3.5 Light Gradient Boosting Machine

Light Gradient Boosting Machine or LightGBM (Ke, et al., 2017) is a distributed high-performance framework that uses decision trees for ranking, classification, and regression tasks (Saha, 2023). This framework has shown high accuracy results and it is quite popular when handling large datasets due to the faster training speed, the higher efficiency, and the lower memory usage and because it supports parallel, distributed, and GPU learning.

4.3.6 Random Forest Regressor

Random Forest (Breiman, 2001) is one of the most well-known bagging-based algorithms where a subset of data points and a subset of features is selected (randomly) for constructing each decision tree.

This random selection of features on each decision tree ensures a low correlation among the decision trees that are producing the intermediate predictions that are at the end averaged to produce the final outcome.

4.3.7 Extra tree regressor

Extra tree (Extremely Randomized Trees) (Geurts, Ernst, & Wehenkel, 2006) is an ensemble learning algorithm very similar to Random Forest with the exception that, not only the feature selection is done at random, but also the selection of the split values - instead of looking for the optimal split at each node - which makes it faster in comparison to the Random Forest.

5 RESEARCH METHODOLOGY

The research methodology of this project follows, as in most machine learning research papers (Mariga & Kamiri, 2021), a quantitative approach with an experimental research design where various techniques and multiples algorithms are tested and compared from a model performance standpoint and the conclusions are based on the obtained predictions and the results of the pre-defined metrics.

This study is, therefore, focused on a supervised learning regression task where a set of statistical processes are used for estimating the relationships between the dependent variable (also known as “target”) and one or more independent variables (also known as “features” or “predictors”) to predict the number of car sales.

The high-level ML workflow can be graphically summarized in the diagram below:

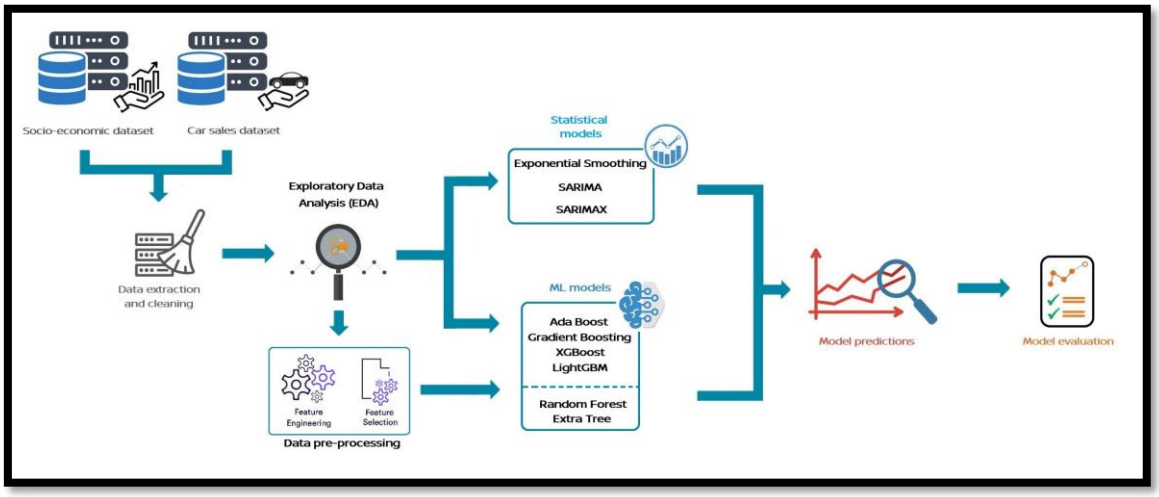


Figure 7: ML workflow

5.1 Data extraction and cleaning

The first step in the presented ML workflow is to load and pre-process the data to make it clean and ready for analysis. In this step, the car sales data is cleansed, sorted, filtered, grouped, and (monthly) indexed to get the number of retails for each Nordic country. Finally, the resulting sales dataset is combined with the external socio-economic dataset to produce the output .XLS file which is used as the input in the rest of the analytics workflow.

For each county under the analysis (FIN, DEN, NOR, and SWE) a specific tab is generated within the Excel file with the below structure. The country objective of the analyses can be selected using the "country" variable as shown in Figure 5.

COLUMN NAME	DESCRIPTION
Date	Monthly index
Orders	Number of retails
CPI	Consumer Price Index
UR	Unemployment Rate
LTIR	Long-Term Interest Rate
TIV	Total Industry Volume

Table 1: Data Structure

As outlined before, we are dealing with a supervised learning regression task which means that the objective is to understand the relationship between the dependent and the independent variables. In this project, the target (or dependent) variable can be set to "Orders" where the objective would be to understand the relationship between the number of cars retails (orders) and the rest of the features, or it can be set to "TIV" where the objective would be to understand the relationship between the total industry volume (TIV) and the rest of the features. For simplicity, this paper will be exclusively focused on the "Orders" as the target variable and the TIV analysis will be left for future work.

Additionally, a Boolean variable (True/False) can be used to decide whether the analysis will include the rest of Nordic countries (orders or TIV depending on the selected target) as exogenous features for the ML algorithms or if those will be kept aside.

5.2 EDA: Exploratory Data Analysis

Exploring and understanding the data you are dealing with is a fundamental part of any data science project and the Exploratory Data Analysis (EDA) provides insights and statistical measures of your data and helps to define and refine key features as well as contributes to better understanding of patterns and relations among the variables.

5.2.1 Feature correlations

The first exploration made to our dataset was to try to identify correlations between the independent variable (number of orders) and the rest of the dependent variables of the dataset using scatter plots. In this case, the rest of the Nordic countries are included as exogenous features.

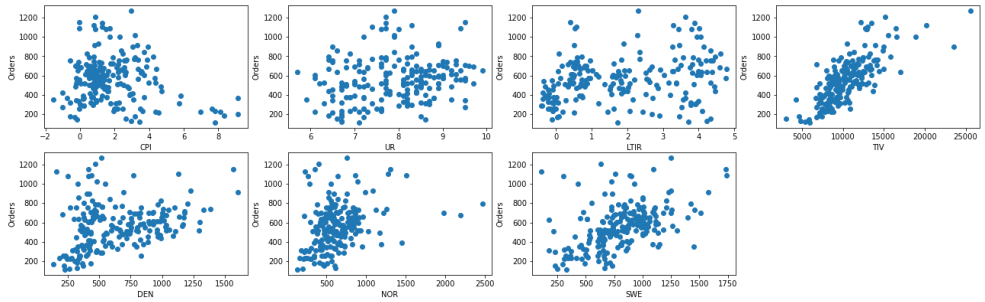


Figure 8: Correlation plot – Finland

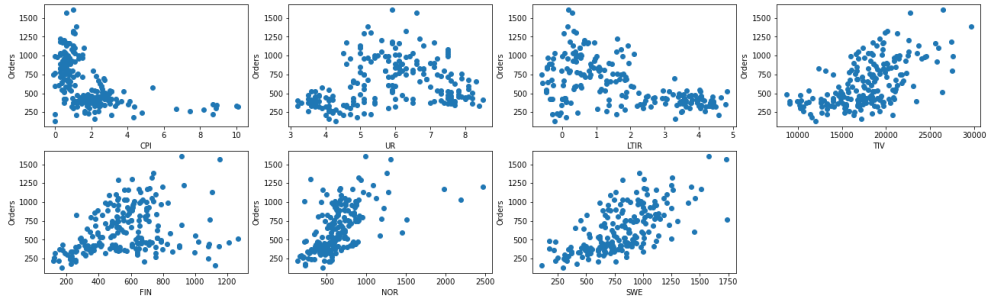


Figure 9: Correlation plot – Target Denmark

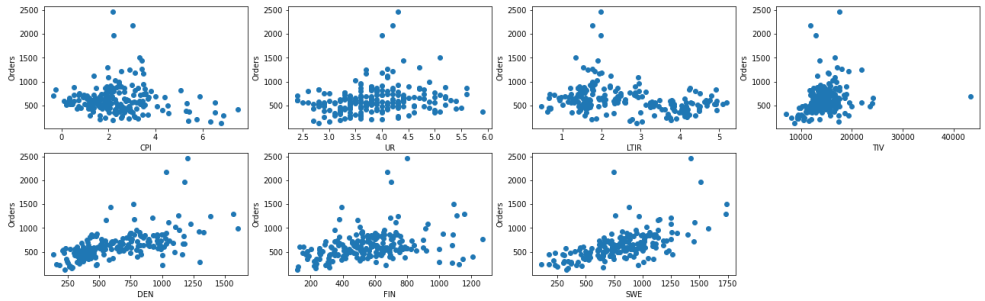


Figure 10: Correlation plot – Target Norway

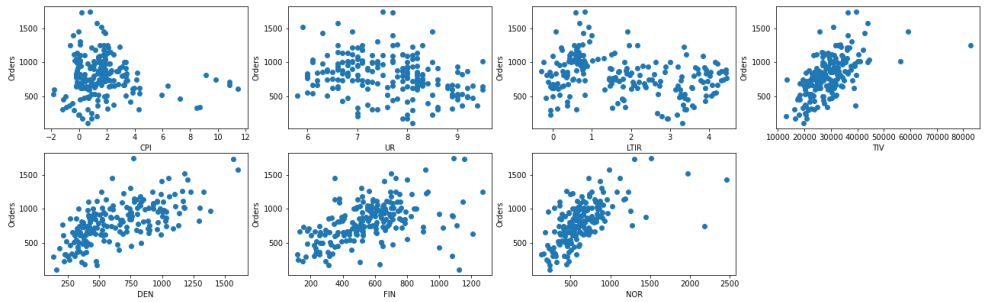


Figure 11: Correlation plot – Target Sweden

Another way of showing this relationship between the features is by computing the Pearson correlation matrix of the dataset. This measure provides the linear correlation between features assigning a value between -1 and 1 to each feature

pair. A positive value indicates a positive correlation, and a negative value indicates a negative correlation.

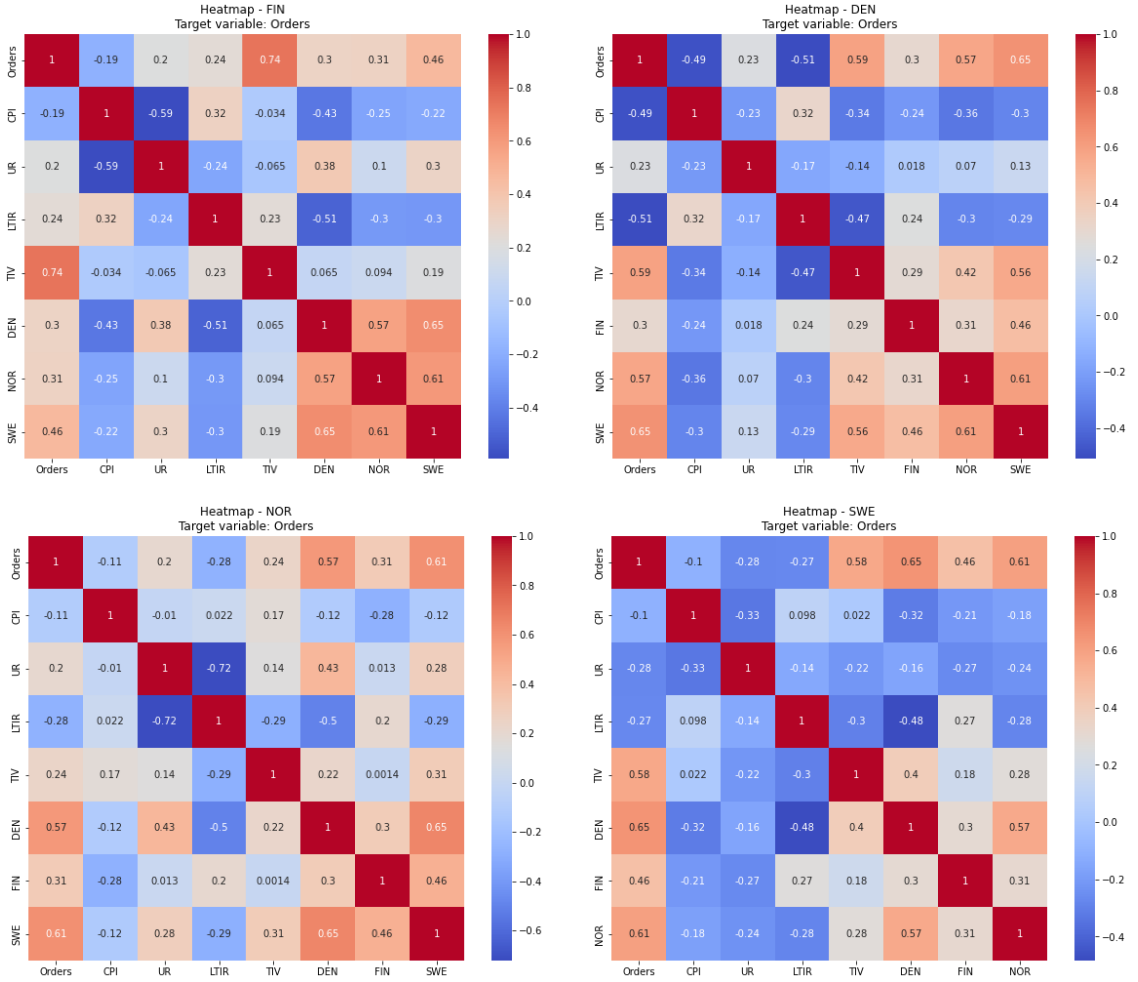


Figure 12: Heatmaps

Some insights can be extracted from the above plots:

- Firstly, the number of Orders which are highly correlated with the TIV in Finland (0.74) and moderately correlated in Denmark and Sweden (0.59 and 0.58 respectively) is almost uncorrelated in Norway (0.24). This supports the theory that the Norwegian market is following its path versus the rest of the Nordic countries.

- On the other side, the number of sales in all countries is not much influenced by the socio-economic indicators. However, while in Finland, Norway, and Sweden this correlation is almost non-existent, in Denmark the number of sales is negatively (weakly) correlated with the consumer price index (0.49) and the long-term interest rates (0.51). This means that the higher those indicators are, the less the sales volume is.

- In the third term, and the most important outcome of this correlation analysis, the number of sales in the countries is, somehow, correlated to each other, but not equally. Finland mostly correlated to Sweden, and Denmark, Sweden, and Norway correlated to each other but not to Finland.

This exploration, even being useful as an initial analysis, does not show through the "lagged" correlation between features, it is, how an observation n times before (t-n) can explain its value (or the value of the other features) at a time t. In the next chapters, we will explore this to get more granularity of the correlation between features.

5.2.2 Time series decomposition

The next step in our EDA process is to carry out a time series decomposition, it is, separating the time series into its components: trend, seasonality, and residuals.

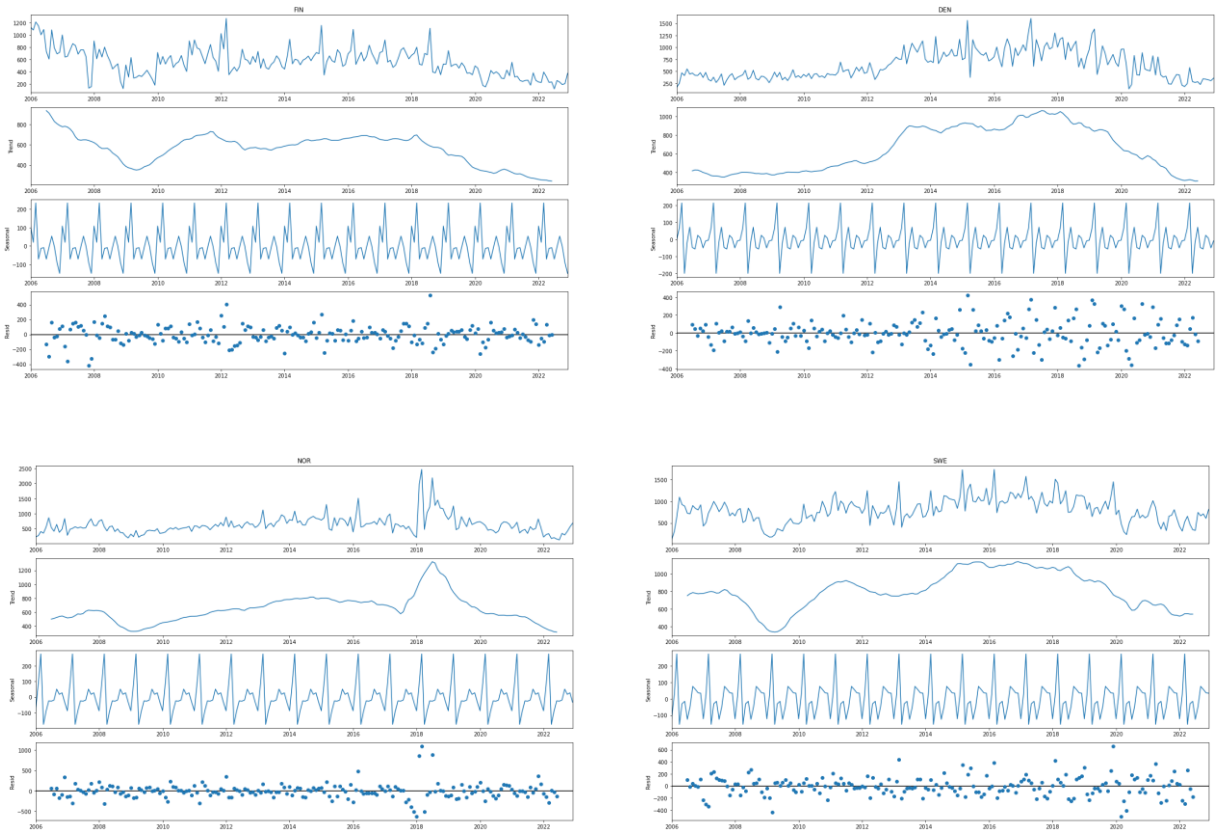


Figure 13: Seasonal Decomposition

5.2.3 Trend

The trend is a pattern in data that shows the movement of a series to higher or lower values over a prolonged period, and it can be classified as an uptrend, downtrend, or horizontal (or stationary).

So stationary series are time series that have constant statistical properties (mean, variance, etc.) over time. Or in other words, the observations in such time series are not dependent on time. And this concept is relevant because, in some cases, machine learning models make more accurate predictions when working with stationary data.

With the time series decomposition plot does not seem easy to conclude whether the time series has any sort of trend or not and, therefore, a further analysis is necessary.

Another method to determine if time-series data is stationary or not is the Augmented Dickey-Fuller Test which has been used as part of this project and will be explained in the data pre-processing chapter.

5.2.4 Seasonality

Another common characteristic of times-series is its seasonal component which is repeated cycles over the time-series observations. Understanding (and processing) the seasonal component can, in some cases, improve the model performance and, especially in machine learning algorithms, can support the learning process by providing clearer signals resulting in a clearer relationship between input and output variables or adding additional features about the seasonal component which can provide added information to improve model performance.

As we can observe below, there are complex patterns of seasonality in the data and not all the countries are showing the same pattern. All of them, however, have a significant peak in March and a decrease in the month of April coinciding with the fiscal year closing of the car brand in this study.

On the other hand, in some countries, the holiday periods may help to explain the down trends during July and December.

With the outcome of the time series decomposition and the above analysis, we can conclude that there is a yearly seasonality in the data and that adding dummy variables with information on the month and quarters could, potentially, contribute to the learning of the machine learning models.

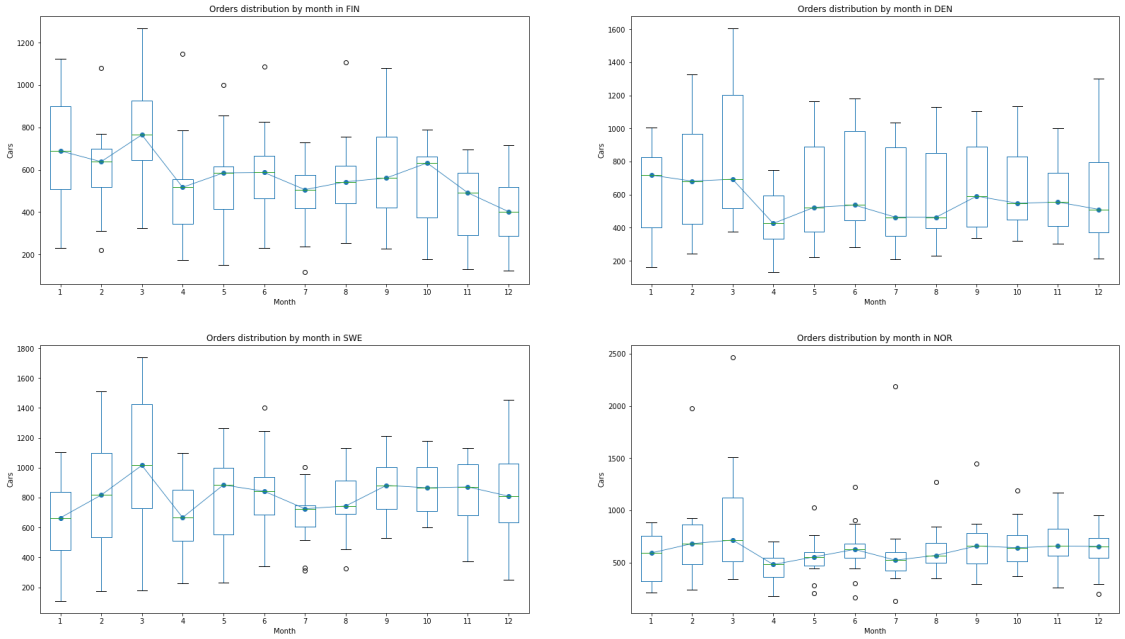


Figure 14: Seasonality plots

5.2.5 Residuals

The last component of the time series decomposition belongs to the residuals which can be defined as the irregular component consisting of the fluctuations in the time series after removing the previous components. These residuals are sometimes also known as noise, and we can appreciate a significant amount of noise in our data that could negatively affect the prediction capacity of the models.

5.2.6 Autocorrelation and partial autocorrelation

The autocorrelation function (ACF) is particularly useful to understand the properties of time series data and, especially in regression problems, they provide valuable information. In contrast, the partial autocorrelation function (PACF) is more useful during the specification process for an autoregressive model, and it can support the decision of the AR (autocorrelation terms) to use in the model fitting.

Autocorrelation is the correlation between two observations at different points in a time series or, in other words, the influence that past values of a time series have on its current value (Frost, s.f.). The distance between observations is known as "lag." From the plots below we can see that there is some sort of autocorrelation in

the number of sales, and this is a good indicator that we should incorporate lagged values of the time series into our regression analysis to model the data appropriately. It also tells us that autoregressive models could be a viable choice in the model selection.

We can also extract from the plots that we are dealing with non-stationary series and that there is some sort of trend because the autocorrelation function drops slowly with significant terms for the first lags of the variable.

Seasonality patterns are observed as well in the ACF plot with peaks of significance after 12 lags which indicates a yearly seasonality in the data, something we already discovered in previous chapters.

On the other side, from the partial autocorrelation plots, we can see that in Finland lags 1 and 2 are statistically significant, in Denmark lags 1,2, and 3 are statistically significant and in Sweden and Norway only lag 1 is significant. Consequently, this PACF suggests fitting different order autoregressive models depending on the country. In this study, however, we made use of the *auto-arima* package which automatically found the appropriate parameters to fit the autoregressive models but will be good to keep in mind our findings to contrast the selection made.

Figure 15: ACF and PACF Finland

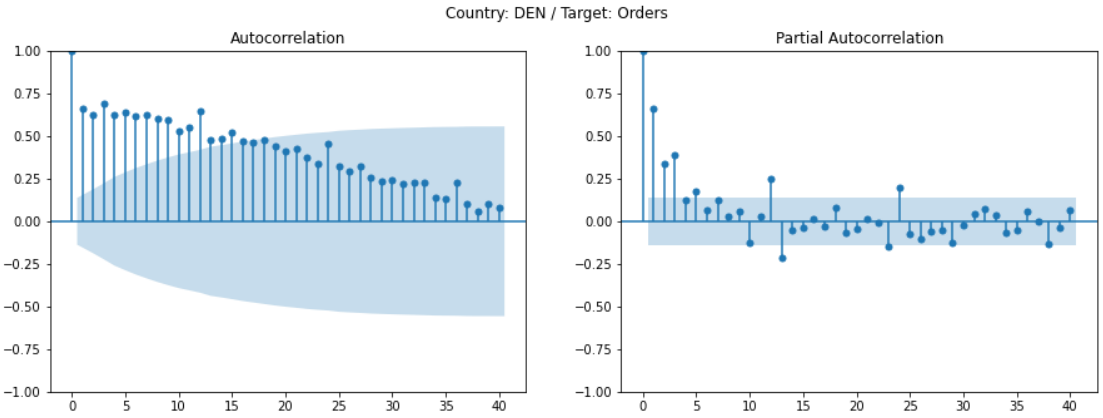


Figure 16: ACF and PACF Denmark

Figure 17: ACF and PACF Norway

Figure 18: ACF and PACF Sweden

5.3 Data pre-processing

Data pre-processing is a fundamental step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of the models to learn. (Kumar, 2018)

In this section, we will cover the different techniques applied to the dataset.

5.3.1 Feature Engineering

As we saw in previous chapters, the input time series we are working with are not stationary but, to statistically confirm this hypothesis, the Augmented Dickey-Fuller Test (ADF) is made to all of them and based on the results, the difference of the series is conducted until each variable becomes stationary.

The ADF test makes conclusions on the hypothesis based on the resulting p-value.

- Null Hypothesis: The data is not stationary.
- Alternative Hypothesis: The data is stationary.

For the data to be stationary (reject the null hypothesis), the ADF test should have:

- p-value <= significance level (0.01, 0.05, 0.10, etc.)
- If the p-value is greater than the significance level, then we can say that it is likely that the data is not stationary.

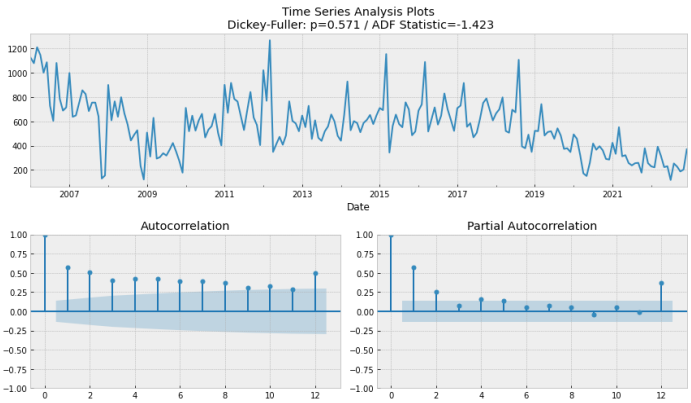


Figure 19: ADF test - original series

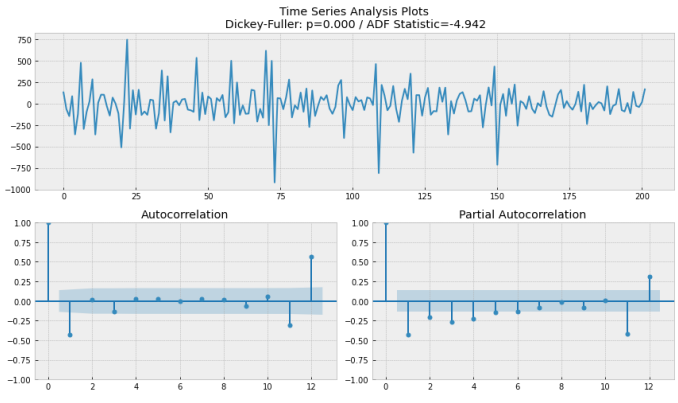


Figure 20: ADF test - 1st difference

The table below shows, by country, the number of differencing terms made to each feature to turn it into a stationary time series.

	Finland	*Denmark*	*Norway*	*Sweden*
Orders	1	1	1	1
CPI	1	1	1	1
UR	0	1	1	0
LTIR	1	1	1	1
TIV	1	1	1	1

Table 2: Number of differences to be made to achieve stationarity

Additionally, lagged features from the target variable were created from month m-1 to m-12 to allow the models to capture the yearly seasonality we identified during the EDA.

Also lagged variables of the exogenous features were created, in this case for the previous 1, 3, 6, 9, and 12 months. The objective of this transformation is to capture the impact of the variations in the macroeconomic indexes on the volume of sales as it is assumed that the changes in those indexes do not have an immediate effect on the volume of sales.

Another transformation done to capture seasonal patterns was the creation of a feature with the number of months encoded using the dummy encoding capability of pandas.

Finally, a 12-month window size rolling mean and rolling standard from the target variable was created.

After all the transformations described above, the resultant dataset had the below features:

Original features	12m lag target variable	1,3,6,9,12m lagged	Month "dummy-encoded"	Rolling target
Orders	Orders_lag1	CPI_lag1	Month_2	rolling_avg
UR	Orders_lag2	LTIR_lag1	Month_3	rolling_std
CPI	Orders_lag3	TIV_lag1	Month_4	
LTIR	Orders_lag4	UR_lag1	Month_5	
TIV	Orders_lag5	CPI_lag3	Month_6	
	Orders_lag6	LTIR_lag3	Month_7	
	Orders_lag7	TIV_lag3	Month_8	
	Orders_lag8	UR_lag3	Month_9	
	Orders_lag9	CPI_lag6	Month_10	
	Orders_lag10	LTIR_lag6	Month_11	
	Orders_lag11	TIV_lag6	Month_12	
	Orders_lag12	UR_lag6		
		CPI_lag9		
		LTIR_lag9		
		TIV_lag9		
		UR_lag9		
		CPI_lag12		
		LTIR_lag12		
		TIV_lag12		
		UR_lag12		

Table 3: Feature engineered dataset

5.3.2 Feature Selection

Feature selection (FS) refers to the process of identifying (and selecting) the most significant and relevant features of a given dataset. Different techniques exist for feature selection, divided into filter techniques, wrapper techniques (i.e., adding or removing features iterative), and embedded techniques (i.e., the selection is already part of the forecasting method).

The three main advantages of feature selection are:

- Simplifying the interpretation of the model.
- Reducing the variance of the model to avoid overfitting.
- Reducing the computational cost (and time) for model training.

In this study, we have used a filter-based feature selection using the Pearson correlation between the features and the target variable as the measure to identify the most relevant ones. A 0.4 threshold has been defined where all features below the threshold were kept aside and a Boolean value defines whether FS should be applied before training the ML model or not.

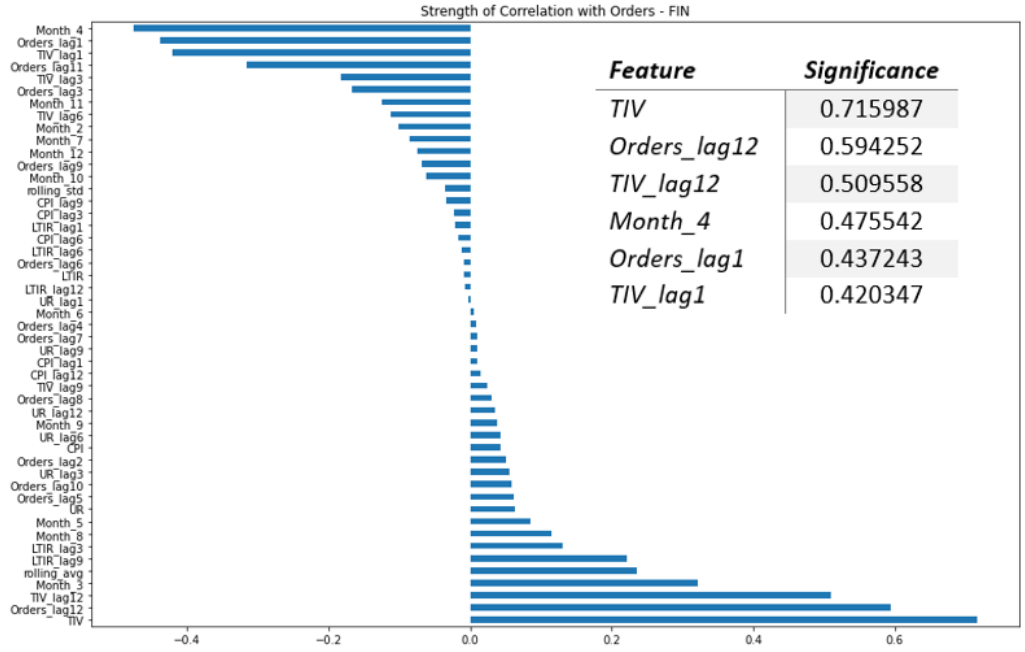


Figure 21: Strength of correlation – Finland

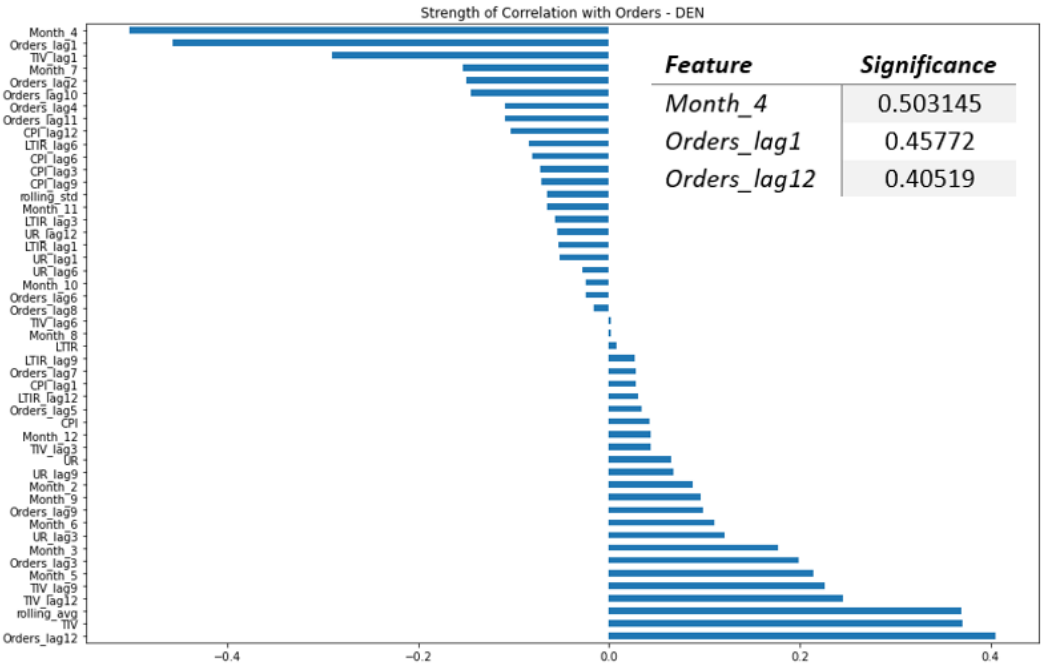


Figure 22: Strength of correlation – Denmark

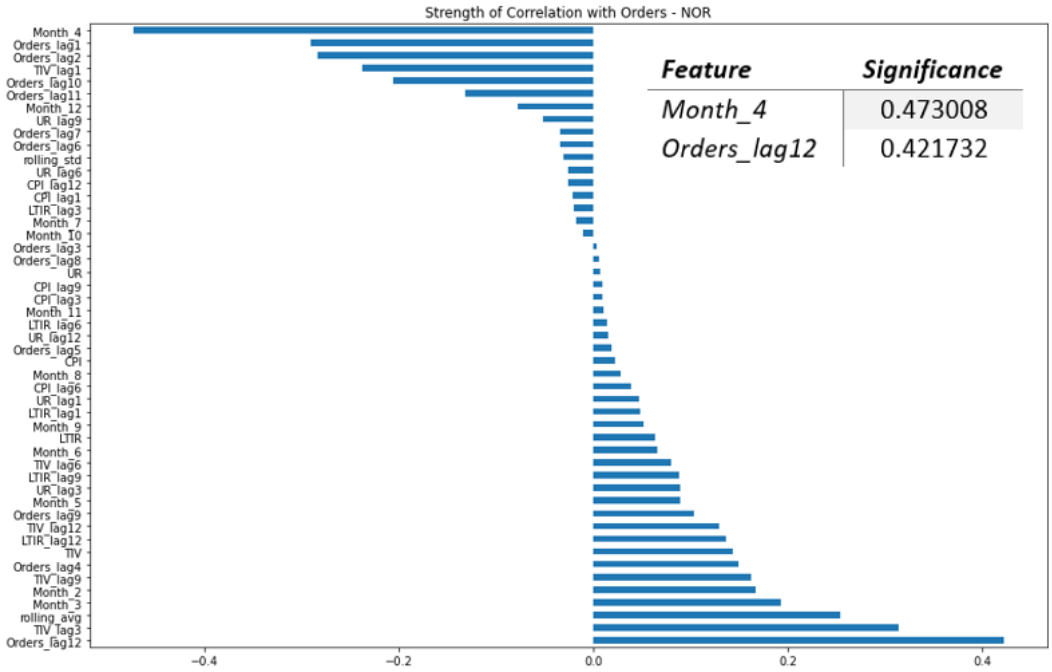


Figure 23: Strength of correlation - Norway

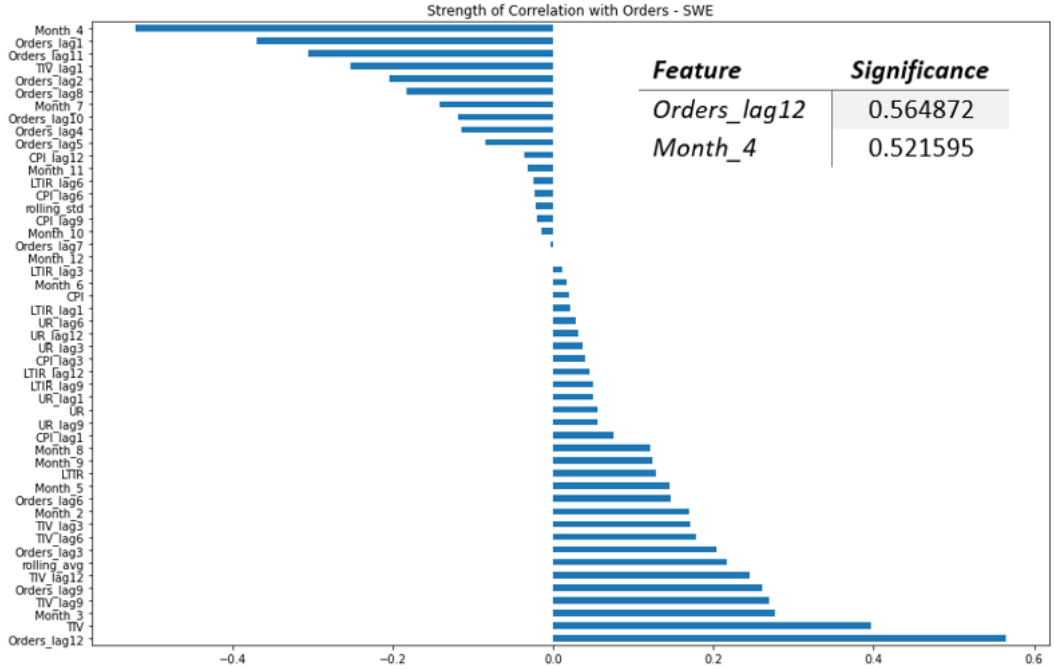


Figure 24: Strength of correlation – Sweden

5.3.3 Train/test split

When working with time series data a simple random split of the data into train and test sets is not a good practice because:

1. It does not maintain the temporal dependence in the data that can be important for modeling
2. It can cause leakage by using future (unknown at the time of training) data during the training process

For this project, we have used the last 24 records (2 years) for testing and the rest of the data available for training.

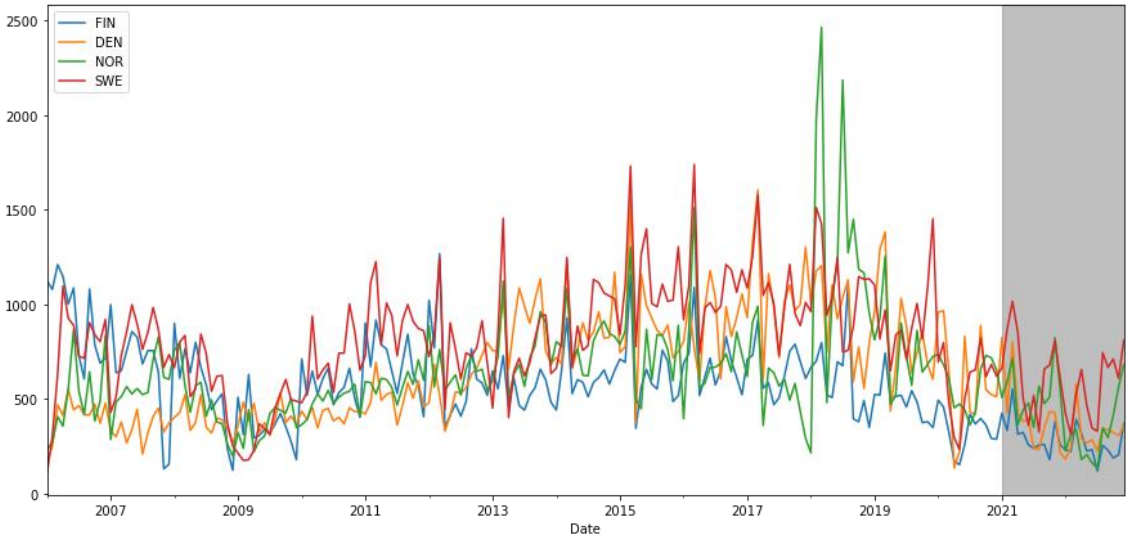


Figure 25: Train/test split

5.4 Performance Metrics

Numerous machine learning (ML)-based predictive modeling techniques are used in this study and, therefore, there is a need to measure the performance of each model and its prediction accuracy.

The metrics used to assess the effectiveness of the model in predicting the outcome are particularly important since they influence the conclusion. In this study, we will use some of the most popular metrics used for regression machine learning models, which are:

- Mean absolute error (MAE): is the average absolute error between actual and predicted values and explains the model performance over the whole dataset.

MAE is a popular metric to use as the error value is easily interpreted because it is on the same scale as the target prediction – in this case the number of retails – and can be seen as the average error that the model's predictions have in comparison with their corresponding actual targets.

"The closer MAE is to 0, the more accurate the model is." (Allwright, 2022)

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

$\mathbf{MAE}$ = mean absolute error
y_i = prediction
x_i = true value
n = total number of data points

Figure 26: MAE

- Root mean square error (RMSE): is the square root of the average squared error between actual and predicted values or, in other words, the square root of the mean squared. The closer RMSE is to 0, the more accurate the model is, and the value is returned on the same scale as the prediction target (Allwright, 2022). For example, calculating RMSE for a car sales prediction model would give the error in terms of car sales, which can help to easily understand model performance.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}}$$

$\mathbf{RMSE}$ = root-mean-square error
i = variable i
N = number of non-missing data points
x_i = actual observations time series
$\hat{x}_i$ = estimated time series

Figure 27: RMSE

- Mean Absolute Percentage Error (MAPE): MAPE is defined as the average absolute percentage difference between predicted values and actual values (Roberts, 2023), and it is a measure of prediction accuracy of forecasting methods in statistics commonly used as a loss function for regression

problems and in model evaluation, because of its very intuitive interpretation in terms of relative error. It usually expresses the accuracy as a percentage ratio defined by the formula below:

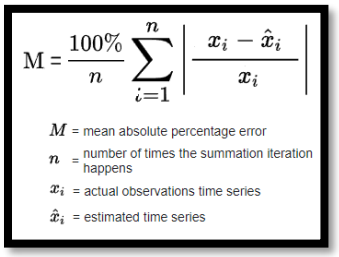


Figure 28: MAPE

6 RESULTS

Now that the reasearch methodolog and the different metrics used in this study have been presented we will turn our attention into the obtained results as part of the multiple experiments carried out. The table below shows a summary of the different possible combinations, where 0 means False, 1 means True and X means not relevant because another selection has priority (i.e. doesn't make sense to apply feature selection when feature engineering is set to False).

With all these possible combinations and, considering the country ('FIN', 'DEN', 'NOR' and 'SWE') and the target selection ('Orders' or 'TIV') up to 64 different experiments could be done with the same piece of code.

Include Nordics	Feature Engineering	Feature Selection	Lagged target
0	0	X	0
0	0	X	1
0	1	0	X
0	1	1	X
1	0	X	0
1	0	X	1
1	1	0	X
1	1	1	X

Table 4: Experiment grid

For the sake of clarity, this project will only be focused on the Orders as the target variable.

In total, as part of this thesis, 32 different experiments have been conducted with the selected machine learning algorithms and 12 additional experiments with the statistical algorithms, giving a total of 44 experiments.

A summary of the obtained results is presented next and compared against the baseline accuracy obtained with the traditional time series models.

6.1 Baseline: traditional algorithms results

Traditional algorithms yielded a mix of results with SARIMAX being the model with the best results in Finland, Denmark, and Sweden (for that one in terms of RMSE) while, in Norway, the Triple Exponential Smoothing algorithm showed the best - particularly poor - results.

COUNTRY	METRIC	TES	SARIMA	SARIMAX
FIN	MAE	90.25	80.15	69.45
	RMSE	102.52	102.07	81.97
	MAPE	37.24	29.95	27.26
DEN	MAE	193.92	240.85	172.51
	RMSE	221.30	272.03	214.60
	MAPE	66.50	82.58	54.36
NOR	MAE	271.61	374.85	281.70
	RMSE	328.46	429.60	343.43
	MAPE	96.63	127.36	102.53
SWE	MAE	172.78	380.45	157.42
	RMSE	214.25	435.23	188.48
	MAPE	29.61	61.65	31.24

Table 5: Traditional models baseline results

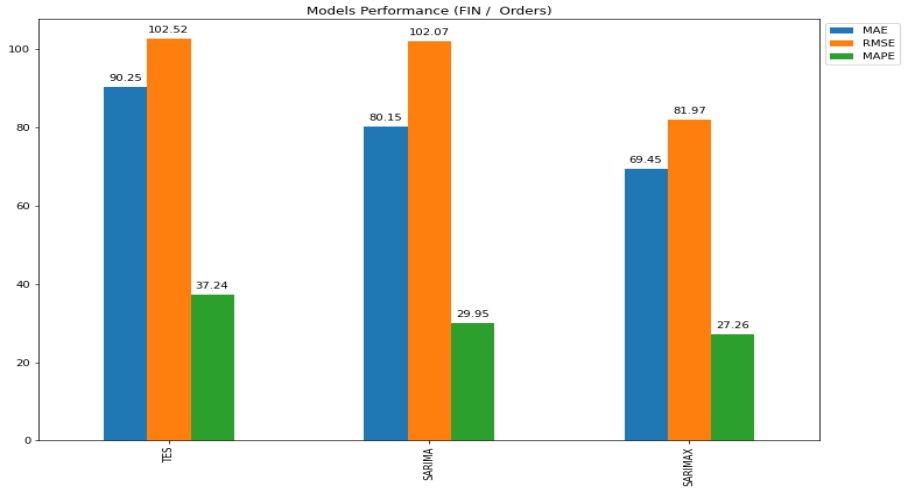
Nevertheless, the accuracy in Denmark and Norway – especially in Norway – was much lower than in Finland and Sweden with an extremely inferior performance as seen in the Table 5, which means that the volume of sales in these countries is not easily explainable with autoregressive algorithms, something that we already anticipated during the correlation analysis done in the EDA.

In the following pages, we deep dive into the results of the traditional models by country.

6.1.1 Finland

SARIMAX with a mean absolute percentage error of 27.26% is the best-performing model and, more importantly, the one with the best fit to the actual values with a RMSE of 82 units. In general, we got quite good-performing results in this scenario, and we can conclude that traditional AR models can predict the sales volume reasonably well. In any case, we can extract from this experiment that the use of external features contributes to the learning rate by improving the overall results.

COUNTRY	METRIC	TES	SARIMA	SARIMAX
FIN	MAE	90.25	80.15	69.45
	RMSE	102.52	102.07	81.97
	MAPE	37.24	29.95	27.26



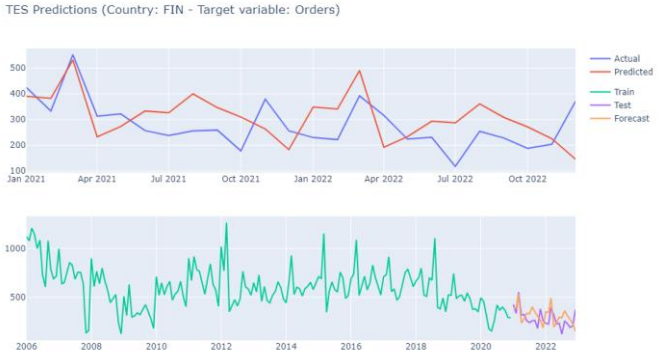
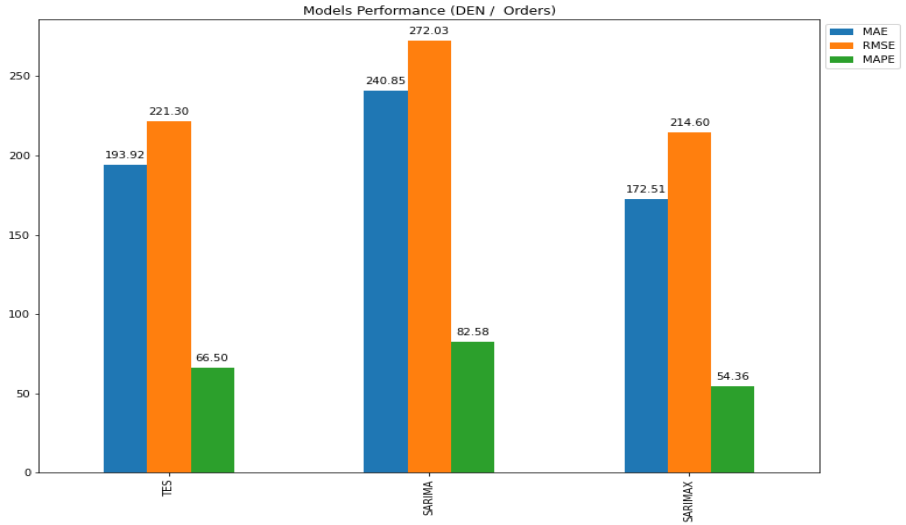


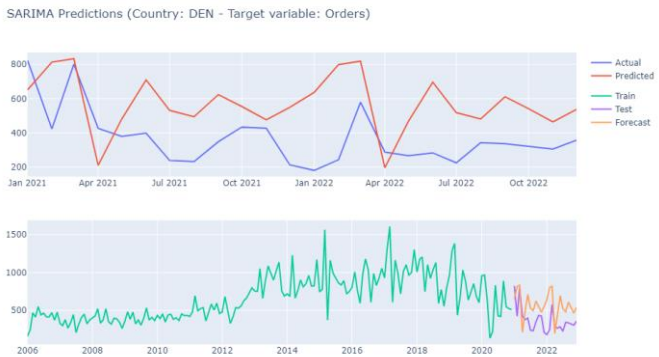
Table 6: Baseline Results Finland

6.1.2 Denmark

Again, in this case, SARIMAX with a MAPE of 54.36% is the best-performing model among the 3 but, opposite to the previous case, in Denmark the accuracy of these models is quite poor with all models above the 50% mean absolute error meaning that sales volumes are not explainable with previous order values but, we can observe that, one more time, the use of external features contributes to the learning rate by improving the overall results.

COUNTRY	METRIC	TES	SARIMA	SARIMAX
DEN	MAE	193.92	240.85	172.51
	RMSE	221.30	272.03	214.60
	MAPE	66.50	82.58	54.36





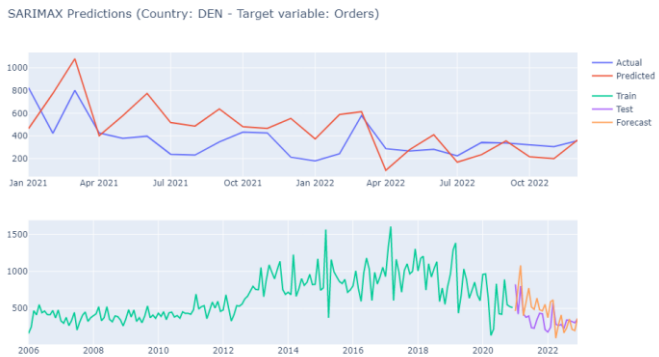
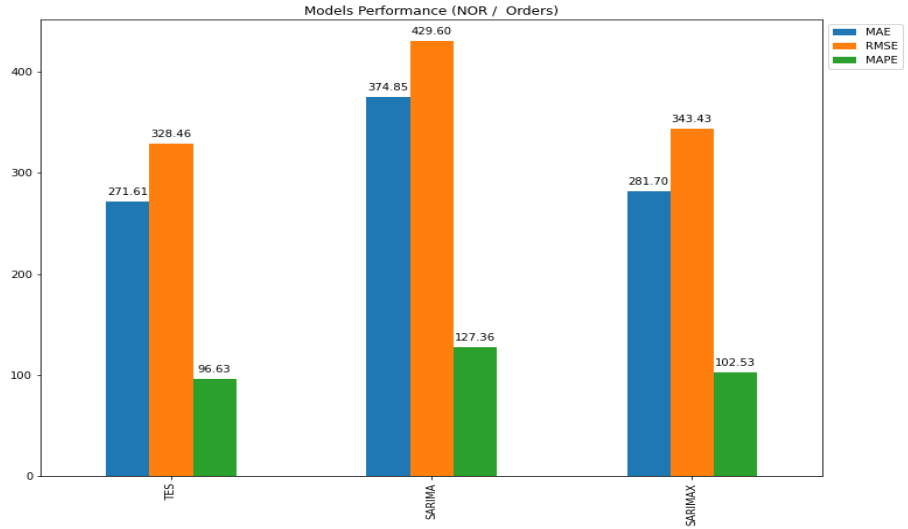


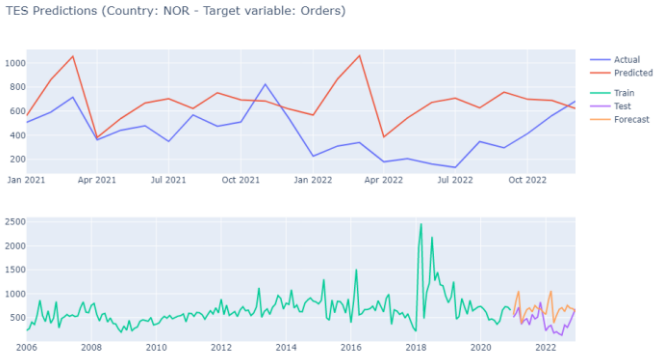
Table 7: Baseline Results Denmark

6.1.3 Norway

As we can see in the charts and graphs, the forecasted values in Norway are random and we can say that the autoregressive models are almost useless to predict the sales volumes. Out of the three models, TES is the one offering the best result but, with a MAPE of 96.63%. The uncorrelation between current and past observations seen during the data exploration explains the poor results of these models for this experiment.

COUNTRY	METRIC	TES	SARIMA	SARIMAX
NOR	MAE	271.61	374.85	281.70
	RMSE	328.46	429.60	343.43
	MAPE	96.63	127.36	102.53





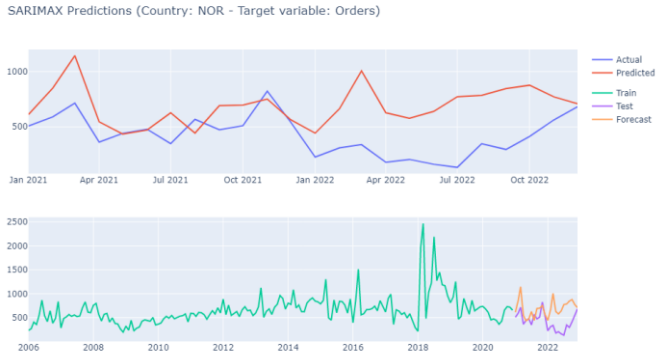
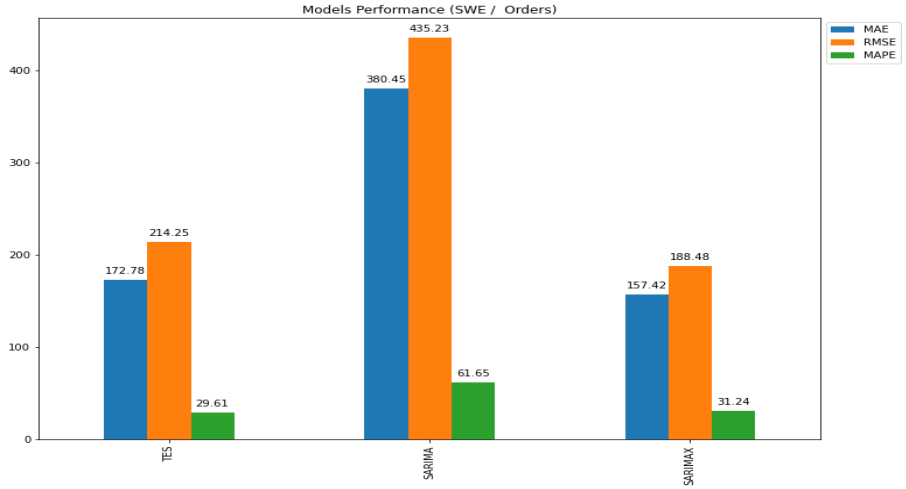


Table 8: Baseline Results Norway

6.1.4 Sweden

Slightly better performance of the TES model in this case if we look at the mean absolute percentage error with 29.61% but not far from the error obtained with SARIMAX with 31.24% and, with a lower RMSE, we can conclude that we got a better fit with the last one. However, considering that TES is only using lagged terms, the results are quite promising. Again, in this experiment, the use of external features contributes to the learning of the model by improving the results of the SARIMA model.

COUNTRY	METRIC	TES	SARIMA	SARIMAX
SWE	MAE	172.78	380.45	157.42
	RMSE	214.25	435.23	188.48
	MAPE	29.61	61.65	31.24



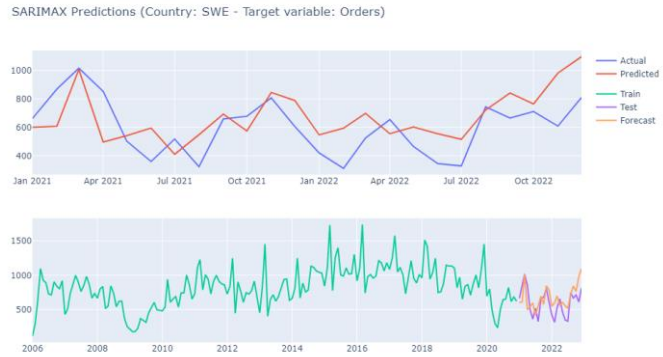


Table 9: Baseline Results Sweden

6.2 ML algorithms results

			FE = 0 / FS = X / LT = 0							FE = 0 / FS = X / LT = 1							FE = 1 / FS = 0 / LT = X							FE = 1 / FS = 1 / LT = X						
	Country	Metric	AB	GB	XGB	LGBM	RF	ET	AVG	AB	GB	XGB	LGBM	RF	ET	AVG	AB	GB	XGB	LGBM	RF	ET	AVG	AB	GB	XGB	LGBM	RF	ET	AVG
INCLUDE NORDICS: 0	FIN	MAE	193.96	130.72	136.96	122.35	147.55	159.81	148.56	115.69	63.18	75.17	78.64	96.88	106.25	89.30	216.69	296.13	273.97	202.29	79.16	67.53	189.29	96.49	270.59	325.49	100.50	181.36	65.42	173.31
		RMSE	209.68	154.00	157.73	149.49	165.21	178.91	169.17	130.16	79.04	91.27	92.98	107.58	117.39	103.07	241.04	324.11	310.73	245.80	94.41	78.66	215.79	110.72	298.64	364.27	125.26	213.50	79.72	198.68
		MAPE	82.83	56.86	58.57	52.21	62.46	67.62	63.43	49.50	21.93	27.13	31.04	39.73	46.15	35.92	95.31	130.26	119.85	83.52	33.49	28.20	81.77	42.19	118.44	143.46	43.16	81.58	24.96	75.63
	DEN	MAE	256.11	300.56	228.16	313.66	314.22	309.21	286.99	162.96	185.27	177.54	213.15	175.37	182.41	182.78	155.64	91.29	199.68	850.07	229.75	665.20	365.27	447.31	893.84	1557.17	573.55	1151.91	1032.98	942.79
		RMSE	297.92	355.20	284.44	348.22	367.78	357.06	335.10	214.92	208.86	212.40	237.64	194.95	200.55	211.55	178.88	112.19	231.07	1062.42	281.17	707.37	428.85	502.74	1027.08	1771.65	664.75	1300.69	1160.67	1071.27
		MAPE	81.79	98.22	72.82	102.65	101.87	100.09	92.91	53.12	61.66	58.86	70.19	57.57	61.55	60.49	43.88	29.34	63.16	274.37	68.03	213.02	115.30	150.42	298.91	523.11	197.01	388.79	346.99	317.54
	NOR	MAE	414.25	473.93	559.67	364.32	575.75	387.79	462.62	182.97	312.97	276.94	252.80	317.89	295.26	273.14	125.99	348.17	292.90	760.55	355.73	244.22	354.59	328.67	531.10	365.08	407.65	526.20	497.71	442.73
		RMSE	453.98	521.13	636.47	414.34	632.49	423.25	513.61	212.52	439.45	347.95	284.81	426.38	359.20	345.05	150.72	493.97	383.20	921.41	393.15	318.21	443.45	407.18	609.76	468.91	464.43	602.69	591.63	524.10
		MAPE	145.18	164.14	178.51	128.82	202.30	128.52	157.91	65.80	125.91	105.53	88.12	129.34	112.48	104.53	42.05	108.72	92.44	231.24	123.02	92.97	115.07	112.77	162.29	97.51	143.07	166.22	144.90	137.79
	SWE	MAE	262.54	251.98	231.82	265.55	263.39	228.97	250.71	196.17	180.13	190.64	182.39	171.25	155.53	179.35	204.17	234.65	226.08	731.91	214.85	271.07	313.79	674.13	262.24	498.28	410.55	341.39	546.52	455.52
		RMSE	324.49	303.89	288.59	338.91	326.32	295.02	312.87	233.99	216.19	233.27	225.09	205.34	193.87	217.96	242.81	276.01	258.47	832.11	278.60	310.10	366.35	711.94	310.01	633.74	473.58	404.81	657.70	531.96
		MAPE	58.81	54.50	50.69	59.15	58.42	51.02	55.43	41.05	34.98	35.40	33.49	34.53	32.30	35.29	35.38	38.42	36.48	129.16	32.11	57.62	54.86	113.54	50.78	105.36	65.17	65.92	112.65	85.57
	Country	Metric	AB	GB	XGB	LGBM	RF	ET	AVG	AB	GB	XGB	LGBM	RF	ET	AVG	AB	GB	XGB	LGBM	RF	ET	AVG	AB	GB	XGB	LGBM	RF	ET	AVG
INCLUDE NORDICS: 1	FIN	MAE	116.27	78.86	91.81	81.77	93.73	104.89	94.55	98.49	63.61	69.44	60.96	80.79	100.33	78.94	87.69	60.54	267.31	329.03	97.12	75.95	152.94	176.86	301.02	141.69	200.19	70.35	168.59	176.45
		RMSE	131.38	95.83	115.49	105.53	110.44	121.63	113.38	115.89	81.35	80.38	71.97	100.49	115.04	94.18	101.65	76.86	320.27	419.04	125.84	91.81	189.25	203.15	311.96	158.88	240.78	85.42	203.66	200.64
		MAPE	47.36	31.23	34.03	31.81	39.19	43.85	37.91	42.77	26.57	26.34	24.14	34.44	42.89	32.86	39.00	25.93	119.44	142.35	39.88	27.33	65.65	78.03	125.10	56.83	85.27	28.48	70.02	73.95
	DEN	MAE	163.90	151.11	131.83	172.89	136.14	210.22	161.02	120.82	144.69	127.59	159.24	139.68	159.02	141.84	147.17	371.18	478.78	297.39	153.76	646.76	349.17	536.22	550.36	634.17	685.21	426.90	596.17	571.50
		RMSE	189.76	183.28	179.40	191.89	163.83	230.96	189.85	139.54	155.19	155.06	188.77	155.83	173.80	161.36	178.27	427.75	512.50	356.63	178.83	679.78	388.96	610.97	629.07	708.82	792.51	479.74	661.80	647.15
		MAPE	51.87	48.26	39.78	56.06	42.70	67.95	51.10	40.14	46.93	37.55	53.44	45.59	53.71	46.23	42.20	115.10	161.25	99.73	44.78	205.44	111.42	180.65	186.87	216.65	231.38	146.04	204.62	194.37
	NOR	MAE	250.80	223.25	142.19	180.50	158.54	178.79	189.01	147.50	141.84	149.80	161.30	154.20	141.88	149.42	251.08	615.99	206.62	568.71	370.11	221.82	372.39	437.04	366.97	282.85	486.69	336.13	292.47	367.02
		RMSE	294.47	260.37	176.00	217.22	204.20	215.79	228.01	170.28	165.23	174.72	186.29	185.85	166.29	174.78	305.15	734.20	256.41	692.75	440.09	266.66	449.21	496.40	460.19	325.05	587.97	361.30	333.55	427.41
		MAPE	94.42	77.58	47.54	65.65	55.80	66.02	67.84	51.26	46.14	46.71	57.04	57.19	51.50	51.64	97.31	188.81	46.21	175.59	136.41	58.65	117.16	156.86	101.93	82.96	174.22	105.29	101.22	120.41
	SWE	MAE	133.78	139.95	170.41	144.02	133.54	133.39	142.52	142.46	142.53	141.49	142.63	144.22	139.50	142.14	476.45	147.57	454.58	450.83	414.06	204.19	357.95	480.76	520.20	490.37	783.19	886.29	163.81	554.10
		RMSE	167.44	165.49	201.64	172.76	161.11	166.11	172.43	168.61	178.29	172.18	174.22	172.61	165.81	171.95	501.75	188.23	502.75	495.72	442.16	239.72	395.06	531.30	579.02	599.09	834.94	919.31	211.44	612.52
		MAPE	27.98	28.50	29.99	28.32	26.78	26.07	27.94	27.75	27.29	24.52	25.23	27.49	25.79	26.35	82.16	31.59	79.95	79.98	69.74	44.31	64.62	80.82	90.77	83.06	137.41	157.49	28.76	96.38

Table 10: ML Results

In the Table 10 the results from the different ML algorithms for each experiment combination are shown and the performance of the ML models is calculated to determine the best (on average) performing combination among the analyzed algorithms.

The rationale to calculate the average performance among the ML algorithms part of this study comes from the objective of this research, which is to determine if those state-of-the-art algorithms can provide better results than traditional times series forecasting strategies. In that sense, rather than finding the best-performing model, the target is to find which one of the different pre-processing strategies offers the best results.

As we can see, on average and, in all countries, the best-performing combination is the one including the rest of the Nordic countries and the lagged target values as exogenous features without feature engineering/selection which leads us to some preliminary conclusions:

- Increasing the complexity of ML models with additional features does not necessarily improve their accuracy.
- Using the volume of sales in the rest of the Nordic countries as input features contributes to the learning capacity of ML models.

6.2.1 Finland

In Finland, the best-performing model among all the models analyzed is the Light Gradient Boosting Machine with an MAPE of 24.14% followed by eXtreme Gradient Boosting with a 26.34%.

In general, we can see that boosting algorithms perform better (except for the Adaptative Boosting) than bagging algorithms and we can also conclude that Machine Learning algorithms can produce better results than the traditional statistical models.

Country	Metric	ADA	GB	XGB	LGBM	RF	ET	TES	SARIMA	SARIMAX
FIN	MAE	112.21	63.34	69.44	60.96	81.29	85.50	90.25	80.15	69.45
	RMSE	129.62	81.24	80.38	71.97	98.60	97.26	102.52	102.07	81.97
	MAPE	49.30	26.54	26.34	24.14	34.43	36.28	37.24	29.95	27.26

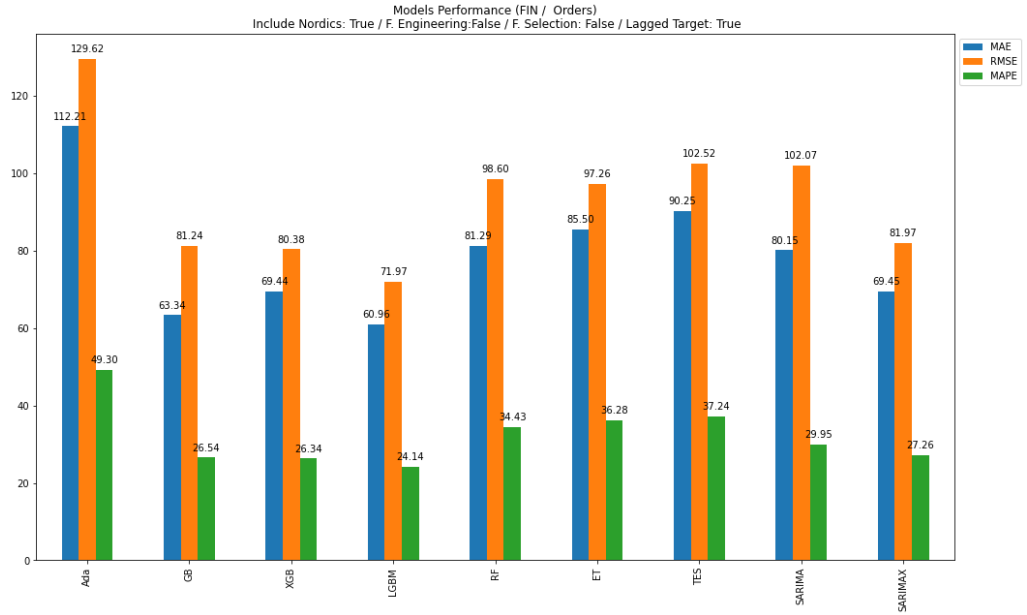


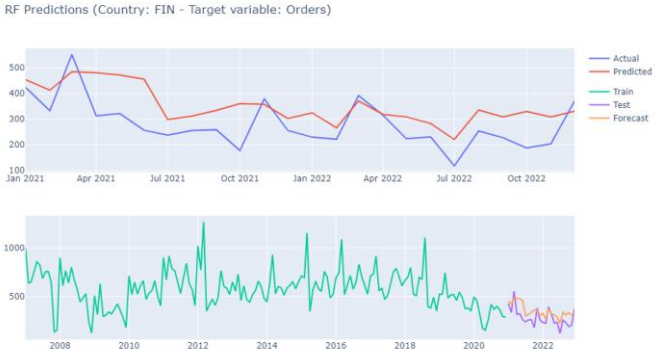
Table 11: Finland Results


Ada Predictions (Country: FIN - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
2008
2010
2012
2014
2016
2018
2020
2022


GB Predictions (Country: FIN - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
2008
2010
2012
2014
2016
2018
2020
2022


XGB Predictions (Country: FIN - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
2008
2010
2012
2014
2016
2018
2020
2022


LGBM Predictions (Country: FIN - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
2008
2010
2012
2014
2016
2018
2020
2022


RF Predictions (Country: FIN - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
2008
2010
2012
2014
2016
2018
2020
2022


ET Predictions (Country: FIN - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
2008
2010
2012
2014
2016
2018
2020
2022

6.2.2 Denmark

In Denmark, the best-performing models belong (again) to the boosting strategy with a mean absolute percentage error of 37.51% yielded by the Adaptative Boosting Model and a 37.55% obtained with the eXtreme Gradient Boosting Model showing these two models a slightly better performance than the rest of machine learning models and much better than traditional autoregressive models.

Country	Metric	ADA	GB	XGB	LGBM	RF	ET	TES	SARIMA	SARIMAX
DEN	MAE	115.27	148.88	127.59	159.24	146.57	154.75	193.92	240.85	172.51
	RMSE	137.21	160.84	155.06	188.77	157.92	166.60	221.30	272.03	214.60
	MAPE	37.51	48.03	37.55	53.44	48.17	51.46	66.50	82.58	54.36

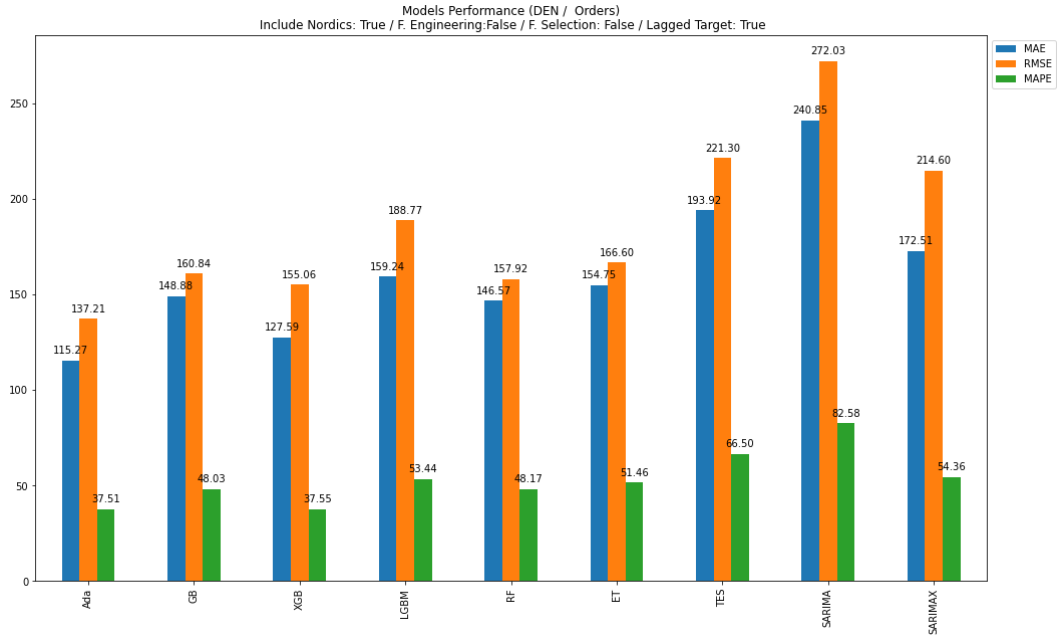


Table 12: Denmark Results


Ada Predictions (Country: DEN - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast
800
600
400
200
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
1500
1000
500
2008
2010
2012
2014
2016
2018
2020
2022


GB Predictions (Country: DEN - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast
800
600
400
200
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
1500
1000
500
2008
2010
2012
2014
2016
2018
2020
2022


XGB Predictions (Country: DEN - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast
800
600
400
200
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
1500
1000
500
2008
2010
2012
2014
2016
2018
2020
2022


LGBM Predictions (Country: DEN - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast
800
600
400
200
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
1500
1000
500
2008
2010
2012
2014
2016
2018
2020
2022


RF Predictions (Country: DEN - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast
800
600
400
200
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
1500
1000
500
2008
2010
2012
2014
2016
2018
2020
2022


ET Predictions (Country: DEN - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast
800
600
400
200
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Apr 2022
Jul 2022
Oct 2022
1500
1000
500
2008
2010
2012
2014
2016
2018
2020
2022

6.2.3 Norway

Gradient Boosting and eXtreme Gradient Boosting are the most accurate models in the case of Norway with a MAPE of 46.54% and a 46.71% respectively showing again that boosting techniques are offering the best performance in terms of car sales forecasting.

In this experiment looks quite evident the learning improvement of machine learning algorithms in respect of traditional statistical algorithms, mostly because lagged terms of the target variable are not self-explanatory in the case of Norway and, therefore, more complex machine learning techniques can capture patterns that remain hidden for the other models.

Country	Metric	ADA	GB	XGB	LGBM	RF	ET	TES	SARIMA	SARIMAX
NOR	MAE	150.15	142.75	149.80	161.30	156.36	138.70	271.61	374.85	281.70
	RMSE	175.08	167.57	174.72	186.29	191.57	163.35	328.46	429.60	343.43
	MAPE	55.02	46.54	46.71	57.04	58.51	49.77	96.63	127.36	102.53

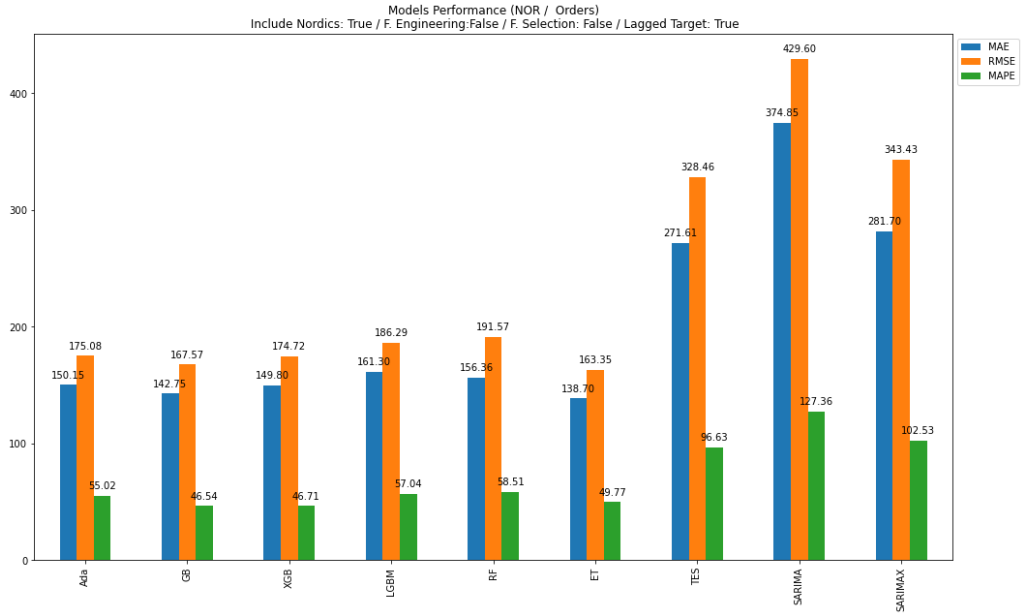
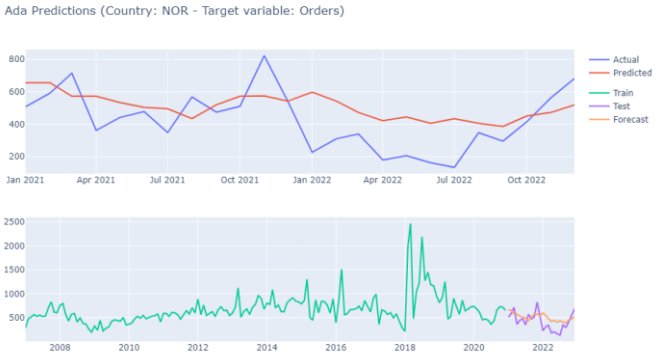
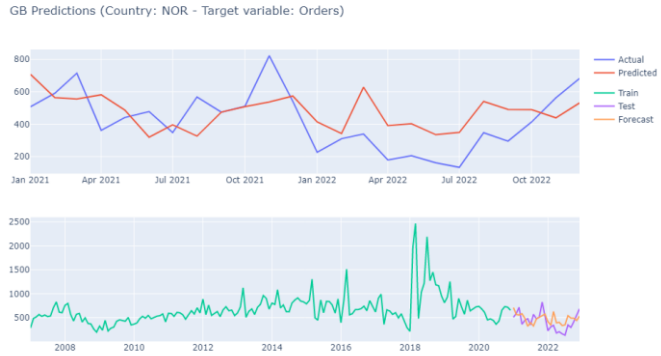
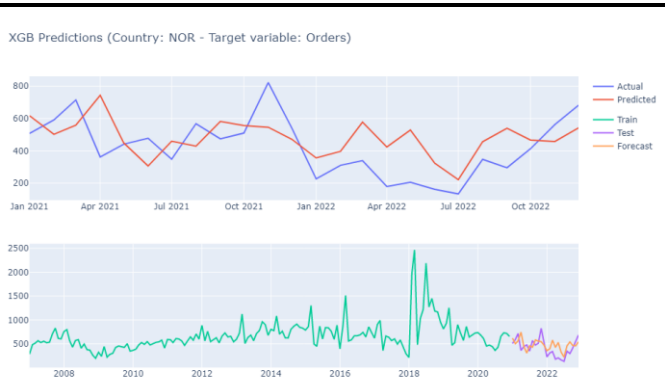


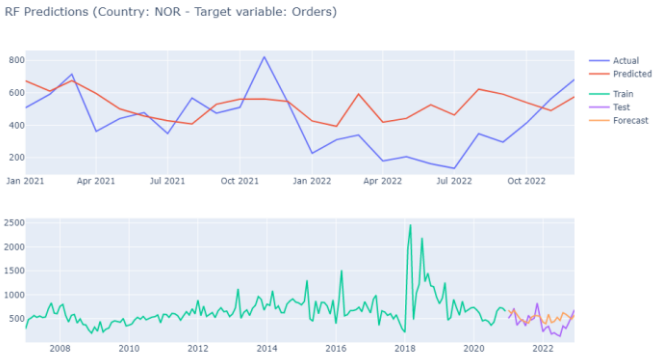
Table 13: Norway Results

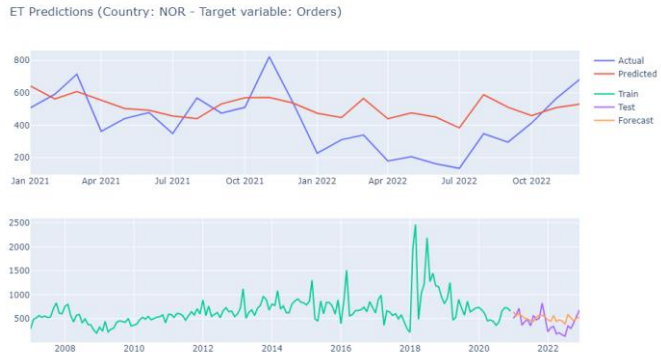

Ada Predictions (Country: NOR - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast


GB Predictions (Country: NOR - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast


XGB Predictions (Country: NOR - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast


LGBM Predictions (Country: NOR - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast


RF Predictions (Country: NOR - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast


ET Predictions (Country: NOR - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast

6.2.4 Sweden

In the case of Sweden, eXtreme Gradient Boosting offers the best results with a MAPE of 24.52%. However, opposite to the rest of the countries, in this case, there is no substantial difference between boosting and bagging techniques with all models showing percentage error results below 30%.

One more time the theory that machine learning algorithms can offer best results than traditional statistical algorithms is supported by the outcomes of this experiment.

Country	Metric	ADA	GB	XGB	LGBM	RF	ET	TES	SARIMA	SARIMAX
SWE	MAE	141.84	142.99	141.49	142.63	148.68	134.14	172.78	380.45	157.42
	RMSE	167.66	178.82	172.18	174.22	174.39	163.21	214.25	435.23	188.48
	MAPE	27.39	27.35	24.52	25.23	28.52	25.13	29.61	61.65	31.24

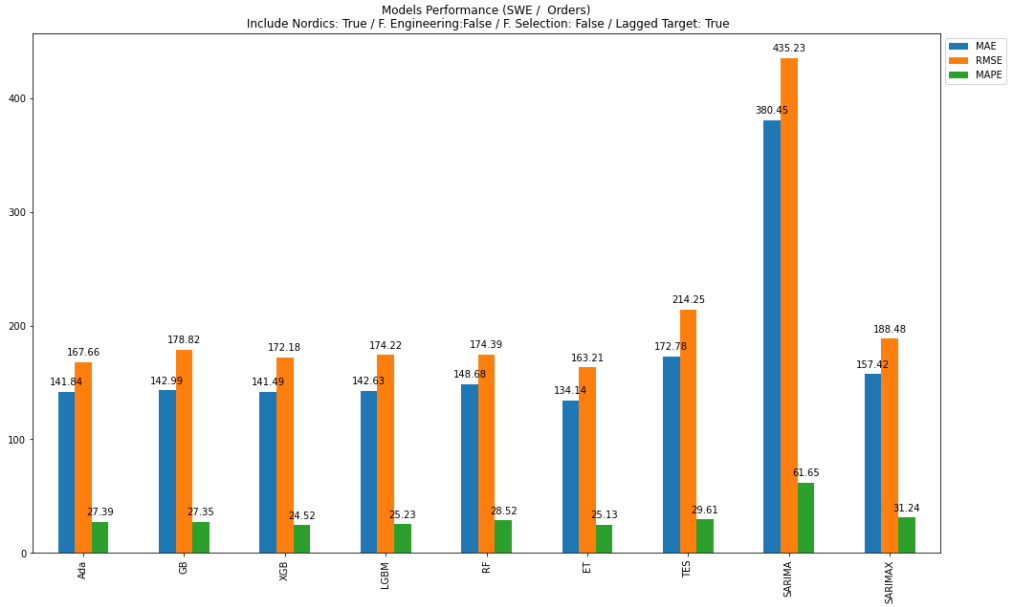


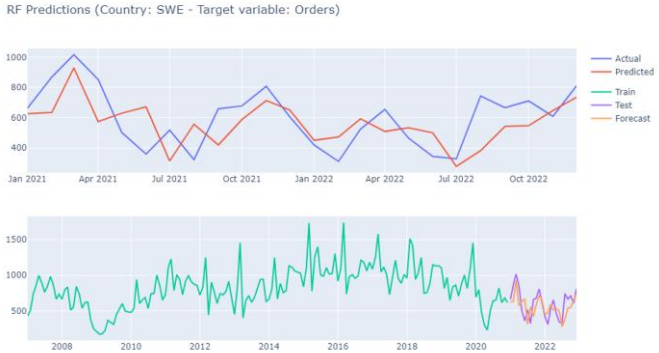
Table 14: Sweden Results

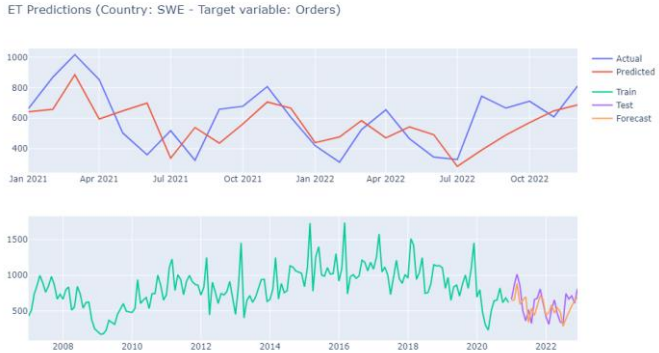

Ada Predictions (Country: SWE - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast


GB Predictions (Country: SWE - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast


XGB Predictions (Country: SWE - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast


LGBM Predictions (Country: SWE - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast


RF Predictions (Country: SWE - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast


ET Predictions (Country: SWE - Target variable: Orders)
Actual
Predicted
Train
Test
Forecast

In summary and, as a conclusion, we have seen that by selecting the right ML algorithm in each case, the results obtained improve in comparison to those provided by traditional algorithms and, despite boosting techniques seem to be, in general, the ones offering better results, bagging techniques shouldn't be kept aside because those can, in some cases and depending on the data, offer as good results as boosting algorithms.

Finally, one important conclusion to extract from these experiments is that, by using the selected socio-economic factors as input features, the learning rate and the accuracy are reinforced as shown by the improvement offered by SARIMAX models versus SARIMA model in all cases. This does not mean that sales volumes are explained just and only by those factors but that they can contribute to the model learning rate.

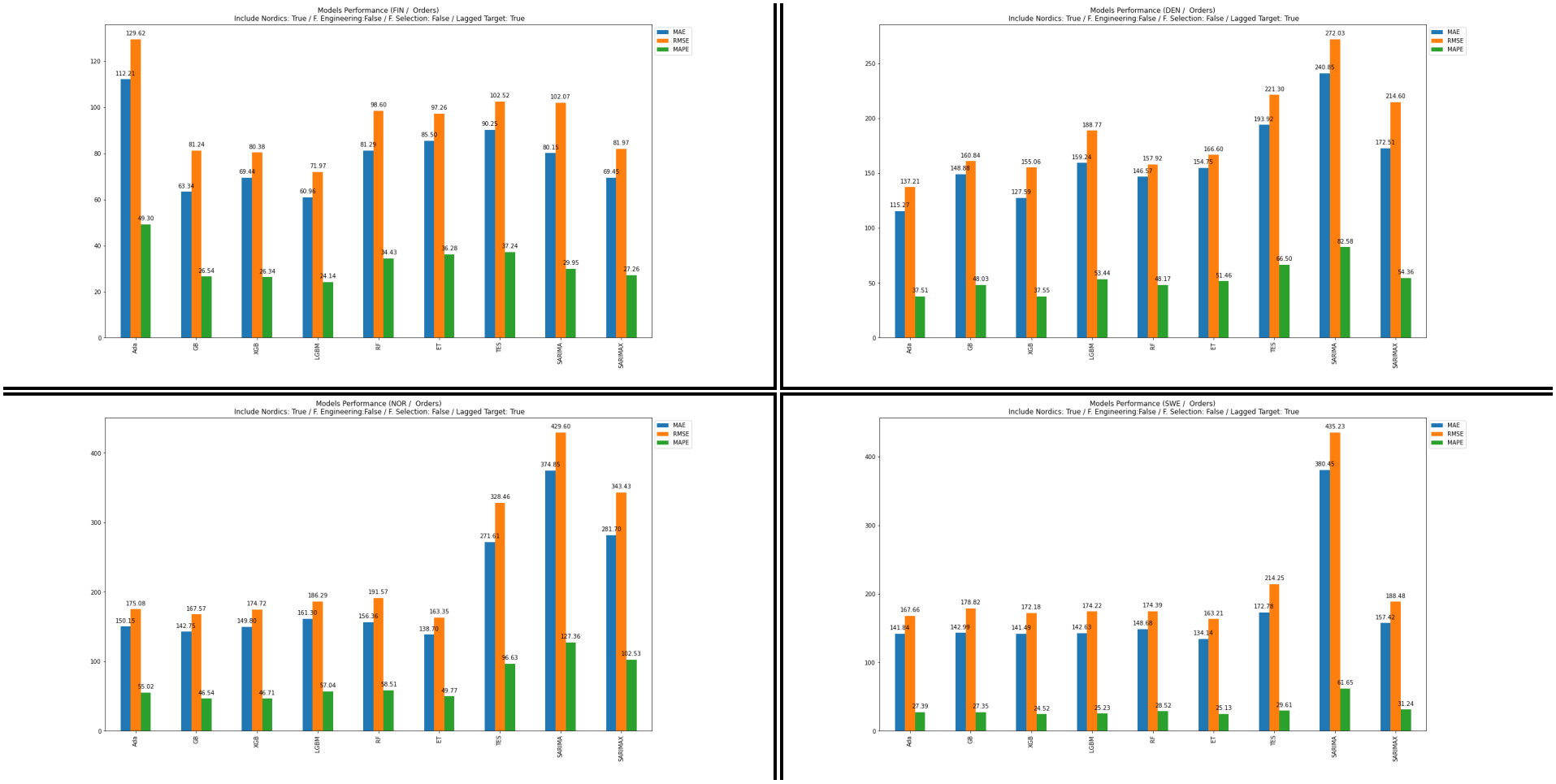


Table 15: Overall accuracy results

7 CONCLUDING DISCUSSION

7.1 Research contributions

The prediction of sales is, in general, a challenging task, and forecasting car sales using artificial intelligence techniques is not exempt from such challenges. However, in this work, the effect of including additional exogenous features to predict the volume of sales in the Nordic countries has been discussed by comparing the results of statistical models with machine learning techniques using a time series dataset including real data with the historical number of sales of a specific carmaker in the Nordic countries.

As part of the research, a preliminary exploratory data analysis was done for every Nordic country providing country-specific insights, and feature engineering and feature selection techniques were applied evaluating the impact of these techniques on the overall model performance.

Opposite to the other studies analyzed that were focused either on traditional models or ML models, a combination of both approaches was used in this research testing a bunch of different algorithms of each type to compare their performance on predicting car sales.

In addition, the performance of all those models for each Nordic market was measured and compared using error metrics such as MAE, RMSE, and MAPE, and the results were analyzed and compared among the different models and among the different countries part of this study.

7.2 Conclusions

The main conclusion from this study and, in response to our first research question *RQ1: Is there any correlation between the analyzed external factors and the volume of car sales in the Nordic countries?* - we can conclude that the automotive market is not easily explainable with the studied macroeconomic indicators as there are numerous external factors with an impact on the number of car sales and, many of them, are unpredictable, such as pandemics or semiconductors crises and, others are the result of human intervention such as marketing campaigns, brand-specific price repositioning campaigns, dealer incentives or other hidden interests leading to a distorted trend in the volume of sales. Nonetheless, we demonstrated some learning improvement when using them as exogenous features and, therefore, we can conclude that there is some sort of correlation between the selected indicators and the volume of sales but, clearly, those factors are not the only ones driving the automotive market and, probably, a deeper study is needed to expand the obtained results.

Additionally, the presented results showed that the introduction of exogenous factors as inputs for the forecasting methods improved the performance of the models and that, despite traditional models providing good results in some of the analyzed cases, in general, machine learning algorithms out-performed classical statistical methods such TES and SARIMA/SARIMAX answering our *RQ3: Are ML algorithms performing better than traditional algorithms for car sales prediction?*

On the other side, the results also demonstrated that there is no "one-size-fits-all" when it comes to car sales forecasting in the Nordic countries and that there is a need for expert analysis and problem understanding to accurately predict the volume of sales which answers our *RQ2: Can we find a model good enough to predict the volume of sales and the market trend in the Nordic countries?*

Another remarkable conclusion out from this research is that adding more features does not necessarily mean a performance increase and, in some cases, it may just add more noise to the data, and keeping the right balance between model complexity and dimensionality reduction is as important as choosing the right

algorithm when dealing with forecasting problems. However, it was demonstrated based on the obtained results that the usage of the volume of sales in the rest of the Nordic countries as exogenous features contributes significantly to the learning capacity of the machine learning models and, therefore, we can positively respond to our *RQ4: Are the Nordic markets affecting each other in terms of volume of car sales?*

7.3 Future work

This work is focused on the comparison of the impact of different data-pre-processing techniques and different statistical and machine learning algorithms in the accuracy performance to predict the number of car sales in the Nordic countries but to extend this work some ideas could be exploited as part of future research such as:

1. Testing the impact of hyperparameter tuning in the different machine learning algorithms.
2. Evaluating different forecasting horizons.
3. Creating a forecaster framework that can be deployed to production following MLOps (Machine Learning Operations) processes.

REFERENCES

1. Allwright, S. (28 de August de 2022). *What is a good MAE score? (simply explained)*. Obtenido de Stephen Allwright: https://stephenallwright.com/good-mae-score/

2. Baržić, M., Munitić, N.-F., Bronić, F., Jelić, L., & Lešić, V. (2022). Forecasting Sales in Retail with XGBoost and Iterated Multi-step Ahead Method. *2022 International Conference on Smart Systems and Technologies (SST)* (pp. 153-158). Osijek: IEEE. Retrieved from https://ieeexplore.ieee.org/document/9954658/

3. Bojer, C. S. (2022). Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities. *International Journal of Forecasting*, 1555-1561. Retrieved from https://doi.org/10.1016/j.ijforecast.2021.11.003

4. Breiman, L. (2001). Random Forests. *Machine Learning*, 5-32. Retrieved from https://link.springer.com/article/10.1023/a:1010933404324

5. Brownlee, J. (28 de August de 2020). *How to Develop Multivariate Multi-Step Time Series Forecasting Models for Air Pollution*. Obtenido de Machine Learning Mastery: https://machinelearningmastery.com/how-to-develop-machine-learning-models-for-multivariate-multi-step-air-pollution-time-series-forecasting/

6. Brühl, B., Hülsmann, M., Borscheid, D., Friedrich, C., & Reith, D. (2009). A Sales Forecast Model for the German Automobile Market Based on Time Series Analysis and Data Mining Methods. *Advances in Data Mining. Applications and Theoretical Aspects*, 146-160. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-03067-3_13

7. Drucker, H. (1997). Improving Regressors Using Boosting Techniques. *Proceedings of the 14th International Conference on Machine Learning*.

8. Frost, J. (s.f.). *Autocorrelation and Partial Autocorrelation in Time Series Data*. Obtenido de Statistics By Jim: Making statistics intuitive: https://statisticsbyjim.com/time-series/autocorrelation-partial-autocorrelation/

9. Geurts, P., Ernst, D., & Wehenkel, L. (2006). *Extremely randomized trees.* Springer Science + Business Media, LLC .

10. Guestrin, C., & Tianqi, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

11. Homolka, L., Vu Minh, N., Drahomíra, P., Bach, T., & Dehning, B. (2020). Short- and medium-term car registration forecasting based on selected macro and socio-economic indicators in European countries. *Research in Transportation Economics*, Volume 80, 100752.

12. Islam, R., Bashawir Abdul Ghani, A., Kusuma, B., & Teh Yew Ho, E. (2016). An Analysis of Factors that Affecting the Number of Car Sales in Malaysia. *International Review of Management and Marketing (IRMM)*, 872–882. Retrieved from https://www.econjournals.com/index.php/irmm/article/view/2726

13. Junjie, G., Yanan, X., Xiaomin, C., Han, Y., & Feng, G. (2018). Chinese automobile sales forecasting using economic indicators and typical domestic brand automobile sales data: A method based on econometric model. *Advances in Mechanical Engineering*, 10(2).

14. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T.-Y. (2017). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree.* Retrieved from https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf

15. Kobayashi, M., Tomino, T., Shintaku, J., & YoungWon, P. (2017). Demand Fluctuation and Supply Chain Integration: Case Studies of Japanese Firms. *Perspectives on Global Development and Technology*, 564-586. Retrieved from https://brill.com/view/journals/pgdt/16/5/article-p564_564.xml?language=en

16. Kumar, D. (25 de December de 2018). *Introduction to Data Preprocessing in Machine Learning*. Obtenido de Towards Data Science: https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d

17. Makatjane, K., & Ntebogang, M. (2016). Comparative Study of Holt-Winters Triple Exponential Smoothing and Seasonal Arima: Forecasting Short Term Seasonal Car Sales in South Africa. *Governance & Control: Financial Markets & Institutions*, Vol. 6. Retrieved from https://virtusinterpress.org/COMPARATIVE-STUDY-OF-HOLT-WINTERS.html

18. Mariga, G., & Kamiri, J. (2021). Research Methods in Machine Learning: A Content Analysis. *International Journal of Computer and Informatrion Technology*, 2279-0764. Retrieved from https://www.ijcit.com/index.php/ijcit/article/view/79

19. Masui, T. (2022, January 20). *Towards Data Science*. Retrieved from All You Need to Know about Gradient Boosting Algorithm − Part 1. Regression: https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502

20. OECD. (2023). Inflation (CPI) (indicator).

21. OECD. (2023). Long-term interest rates (indicator).

22. OECD. (2023). Unemployment rate (indicator).

23. Prastiwi, A., & Tahi Ulubalang, D. (2020, November 23). *Algoritma Technical Blog*. Retrieved from Boosting Algorithm (AdaBoost and XGBoost): https://algotech.netlify.app/blog/xgboost/

24. Roberts, A. (2023, February 2). *Mean Absolute Percentage Error (MAPE): What You Need To Know* . Retrieved from Arize: https://arize.com/blog-course/mean-absolute-percentage-error-mape-what-you-need-to-know/

25. Saha, S. (2023, January 25). *Neptune Labs*. Retrieved from XGBoost vs LightGBM: How Are They Different: https://neptune.ai/blog/xgboost-vs-lightgbm

26. Sa-ngasoongsong, A., Bukkapatnam, S., Kim, J., S. Iyer, P., & Suresh, R. (2012). Multi-step sales forecasting in automotive industry based on structural relationship identification. *International Journal of Production Economics*, 875-887. Retrieved from https://www.sciencedirect.com/science/article/pii/S0925527312003167

27. Sanjog, J., & Shoaib, W. (2022). Automobile Sales Forecasting and Correlation with Economic Indicators: A Comprehensive Intra-Region Case Study. *Advances in Manufacturing Technology and Management. Lecture Notes in Mechanical Engineering* (págs. 978-981). Singapore: Springer. Obtenido de https://link.springer.com/chapter/10.1007/978-981-16-9523-0_50

28. Schapire , R., & Freund, Y. (1997). A desicion-theoretic generalization of on-line learning and an application to boosting.

29. Shin, T. (24 de May de 2022). *Predicting the future: Time series analysis with simple exponential smoothing in Python*. Obtenido de The Operator: The Official Census Blog: https://www.getcensus.com/blog/predicting-the-future-time-series-analysis-with-simple-exponential-smoothing

30. Stoll, F. (2020). *A Comparison of Machine Learning and Traditional Demand Forecasting Methods.* Retrieved from https://tigerprints.clemson.edu/all_theses/3367

31. Suomen Pankki. (2021, June 1). *Vehicle loans account for almost one fifth of consumer credit granted by credit institutions.* Retrieved from https://www.suomenpankki.fi/en/Statistics/mfi-balance-sheet/older-news/2021/vehicle-loans-account-for-almost-one-fifth-of-consumer-credit-granted-by-credit-institutions/#_ftn2

32. Yasaman , E., Saman , A. H., Guoqing, Z., & Bharat , S. (2022). Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. *International Journal of Information Management Data Insights*, Volume 2, Issue 1. Retrieved from https://doi.org/10.1016/j.jjimei.2022.100058

33. Zhang, C., Tian, Y.-X., & Fan, Z.-P. (2022). Forecasting sales using online review and search engine data: A method based on PCA–DSFOA–BPNN. *International Journal of Forecasting*, 1005-1024. Retrieved from https://www.sciencedirect.com/science/article/pii/S0169207021001217

34. Zhang, L., Bian, W., Qu, W., Tuo, L., & Wang, Y. (2021). Time series forecast of sales volume based on XGBoost. *Journal of Physics: Conference Series*, 1873. Retrieved from https://iopscience.iop.org/article/10.1088/1742-6596/1873/1/012067