UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Data Science Curriculum

Andri Hõbemägi

Sales Forecasting based on Economic Indicators for a Construction Company

Master's Thesis (15 ECTS)

Supervisor: Novin Shahroudi, MSc

Tartu 2024

Sales Forecasting based on Economic Indicators for a Construction Company

Abstract:
The construction sector's economic performance is closely related to macroeconomic conditions, shaping investment decisions and market dynamics. Despite its significance, forecasting sales within this sector is notably challenging due to the sector's high sensitivity to economic fluctuations. Sales forecasting, in practice, traditionally relies on bottom-up approaches, where the prediction process starts from compiling individual business unit-level budgets and aggregates upwards to forecast the company's sales revenue. This method can overlook broader economic trends, failing to incorporate the vital interplay between market dynamics and company activities. Our study addresses this challenge by integrating both historical sales data and macroeconomic indicators into a unified forecasting model. By constructing various feature sets through feature engineering and systematically evaluating them, we demonstrate how each set contributes to enhancing predictive performance. We applied five pre-selected regressors on each dataset, aiming to select the model with the lowest possible error with respective variability. As part of our pipeline, we also conducted optimisation of hyperparameters to fine-tune each regressor's performance. We compared the results against the error threshold of 10%, which is of material importance as set by the Nasdaq Tallinn stock exchange for public companies to assess the relevance of the errors obtained from our models. This approach not only aligns company sales forecasts with external economic indicators but also refines the model's accuracy through targeted data enhancement and parameter optimisation. Conducted for Nordecon AS, a leading publicly listed construction company, this research provides a framework for robust quarterly forecasts that contribute to improving the company's decision-making processes.

Keywords:
Sales forecasting, regression analysis, time series analysis, machine learning, macroeconomic indicators

CERCS:
P176 – Artificial intelligence, S181 – Financial science, S183 – Cyclical economics

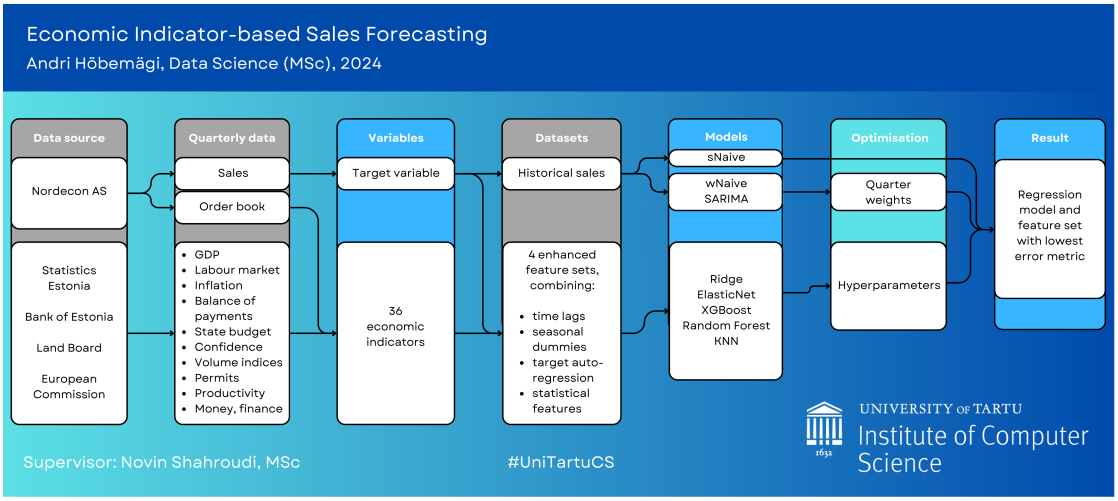
Ehitusettevõtte müügitulu ennustamine majandusindikaatorite põhjal

Lühikokkuvõte:
Ehitussektori tulemused on tihedalt seotud makromajandusliku keskkonnaga, mis mõjutab oluliselt nii ettevõtete investeerimisotsuseid, kui ka sektori üldist turudünaamikat. Hoolimata selle olulisusest on müügitulu prognoosimine ehitussektoris keeruline, kuna sektor on väga tundlik majanduse kõikumiste suhtes. Müügitulu ennustamine toetub traditsiooniliselt alt-üles lähenemisele, kus prognoosimisprotsess algab üksikute äriüksuste eelarvete koostamisest ja liigub ülespoole, prognoosimaks ettevõtte kui terviku kogu müügitulu. See meetod võib aga eirata laiemaid majandustrende, jättes arvestamata turudünaamika ja ettevõtte tegevuste vahelise olulise seose. Meie uuring käsitleb seda väljakutset, integreerides ühtsesse prognoosimismudelisse nii ajaloolised müügiandmed kui ka makromajanduslikud näitajad. Erinevate arvtunnuste komplektide loomise ja süstemaatilise hindamise kaudu näitame, kuidas iga komplekt aitab kaasa ennustustulemuste parandamisele. Igal andmekogumil rakendasime viit eelvalitud regressorit, eesmärgiga valida mudel, mille mõõtmisviga koos vastava variatiivsusega oleks võimalikult väike. Töövoo osana viisime läbi ka hüperparameetrite optimeerimise, et parandada iga regressori ennustustäpsust. Võrdlesime tulemusi Nasdaq Tallinna börsi poolt avalikele ettevõtetele kehtestatud 10%-lise olulisuse lävendiga, et hinnata vea olulisust. See lähenemine mitte ainult ei vii ettevõtte tulemusprognoose kooskõlla väliste majanduskeskkonna mõjudega, vaid parandab ka mudeli täpsust lähteandmete täiustamise ja parameetrite optimeerimise kaudu. Uuring viidi läbi Nordecon AS-i, juhtiva börsil noteeritud ehitusettevõtte müügitulu andmetel. Tulemusena loodi täpsem raamistik müügitulu kvartaalseks ennustamiseks, millele tuginedes on võimalik parandada ettevõtte otsustusprotsesse.

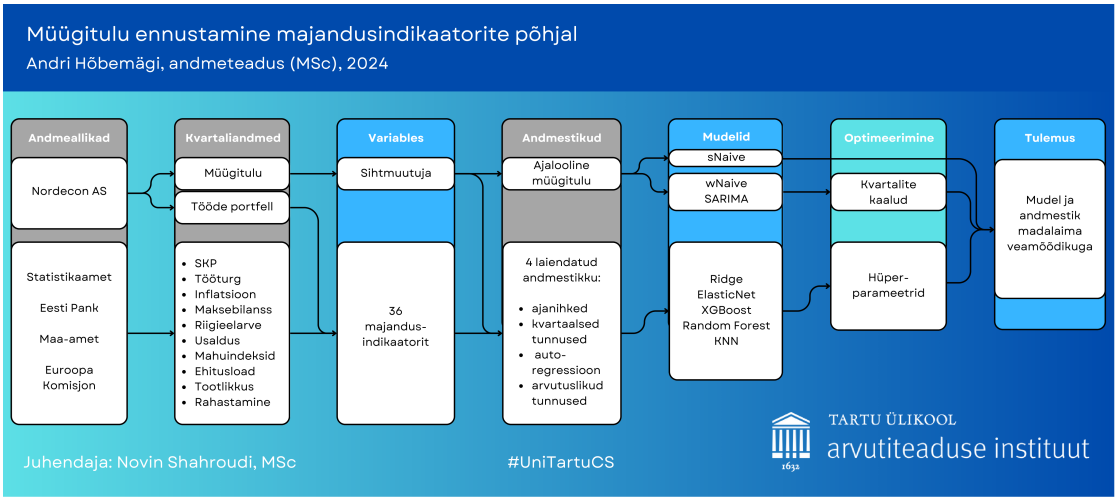
Võtmesõnad:
Müügitulu ennustamine, regressioonianalüüs, aegridade analüüs, masinõpe, makromajandusindikaatorid

CERCS:
P176 – Tehisintellekt, S181 – Rahandus, S183 - Majandustsüklid

Graphical Abstract:



Visuaalne kokkuvõte:



Contents

1 Introduction

The construction sector holds a pivotal and consistent role in the Estonian economy, as evidenced by its share of value added at current prices over the past decade. Its contribution to the country's gross domestic product (GDP) has ranged from 6.1% to 7.0%[1], underlining its significance as a major economic driver.

In the broader economic context, the construction sector is instrumental in enabling and enhancing economic activities through its output of investment goods. This goes beyond their intrinsic value, contributing significantly to the well-being and growth of the economy. Without sustained investment in construction, economic health is likely to deteriorate [Hil00]. In Estonia, investments in the construction of dwellings and other buildings and structures typically account for between 48% and 56% of the total gross capital formation[2], demonstrating a substantial impact on the nation's economic landscape.

Moreover, the construction sector enhances societal well-being by shaping living and working conditions, thereby increasing worker productivity and indirectly fostering economic growth. The built environment also supports the delivery of services by other sectors, collectively representing approximately 45% of the total economy [Joh18].

Construction activity typically lags behind the demand cycle, introducing a delay or inertia in response to economic changes [PRB99]. This delay often leads to over-production during economic downturns, illustrating the sector's significant volatility and sensitivity to broader economic conditions. Such cyclicality not only underscores the construction sector as a major contributor to economic volatility [Bar09], but also highlights its acute reactivity to economic stimuli, which can stem from varied sources such as domestic fiscal policies, international trade dynamics, or broader geopolitical events affecting investor sentiment and construction activity.

The impact of the construction sector's volatility is starkly evident when compared to the general GDP, as demonstrated by the quarter-on-quarter fluctuations shown in Figure 1. Although the average growth rates for GDP and the construction sector are similar—1.7% and 1.5%, respectively—the standard deviations reveal a significant disparity. The construction sector exhibits a higher standard deviation of 5.1% compared to 2.1% for GDP, indicating a greater degree of volatility.

This variance not only illuminates the construction sector's reactivity to economic changes but also its substantial impact on the macroeconomic environment. A heatmap

[1] All data cited on this page are derived from Statistics Estonia (`https://andmed.stat.ee/en/stat`). The percentages and figures presented have been calculated by us based on the raw data obtained from various tables within this source.

[2] Gross capital formation includes outlays on additions to the fixed assets of the economy, such as land improvements; plant, machinery, and equipment purchases; and the construction of roads, railways, and buildings.

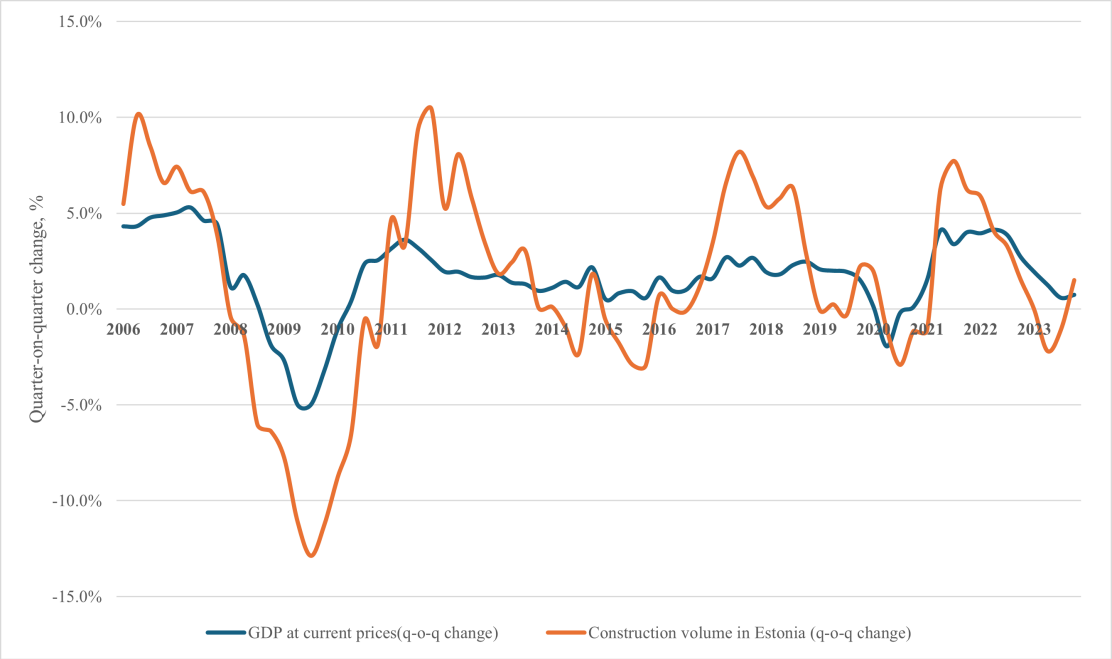


Figure 1. Change in four quarter volumes of Estonian GDP and construction market

of selected macroeconomic indicators[3] as published in [Ees] and exhibited in Figure 2 further illustrates that within the economy as a whole, different sectors exhibit ebbs and flows of varying magnitudes and timelines in their performance.

1.1 Motivation

This initial volatility in the construction sector has profound implications for forecasting methodologies, particularly within companies operating in this space. It necessitates a forecasting approach that not only assimilates the sector's intrinsic variability but also remains agile in the face of its magnified risks and opportunities. This research, conducted for Nordecon AS[4]—a leading publicly listed construction company—aims to provide a framework for robust quarterly forecasts that significantly enhance the company's decision-making processes, ensuring strategic alignment with market realities and improving its competitive stance in the industry. Considering Nordecon AS's average market share of approximately 8%[5] in the Estonian construction sector over the past 18

[3]Macroeconomic indicators are statistics or readings that reflect the economic circumstances of a particular country, region, or sector, used by analysts and governments to assess the current and future health of the economy (`https://www.xtb.com/en/education/macroeconomic-indicators`).

[4]Website: `https://nordecon.com/en/`.

[5]This figure is based on our calculations. For example, in 2023 Nordecon's sales in Estonian market amounted to 272 million euros, whereas total value of construction production in Estonia according to

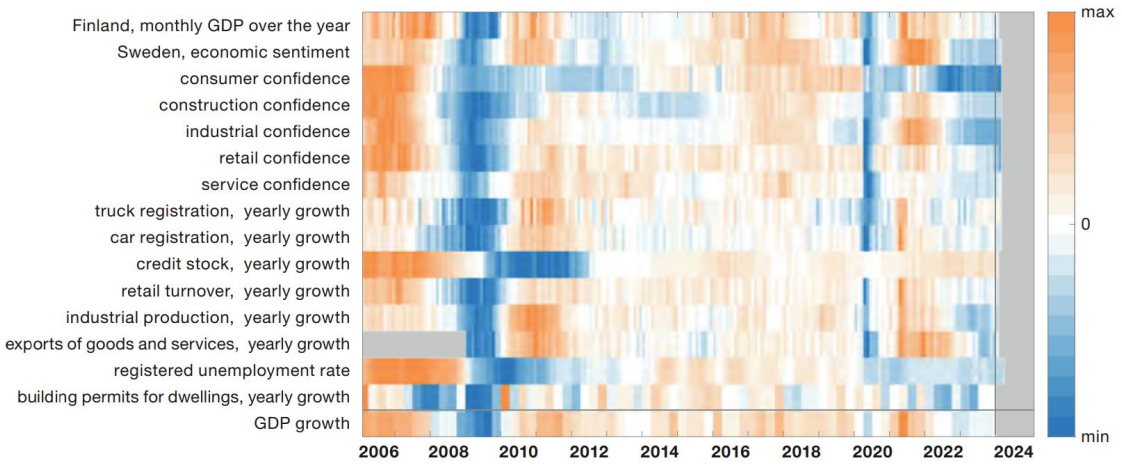


Figure 2. Heatmap of the Estonian economy

years, it is reasonable to infer that the company is broadly exposed to the vicissitudes of market dynamics. Hence, the necessity for enhanced predictive accuracy is not merely a statistical challenge but a strategic imperative.

1.2 Aim of the Work

Traditional statistical forecasting methods primarily extrapolate historical trends and seasonal patterns to predict future sales. However, these methods fall short in anticipating macroeconomic changes that can significantly impact demand [VAD20]. To compensate for these potential shifts, companies often resort to manually adjusting their statistical forecasts or depending on expert predictions. Unfortunately, both strategies tend to introduce biases and inaccuracies, as humans generally struggle with making precise adjustments [FGLN09]. Additionally, these methods are time-consuming, complicating the forecasting process further.

Moving from traditional statistical forecasting methods, which struggle with macroeconomic shifts, machine learning (ML) offers promising alternatives for handling multivariate datasets effectively. For instance, ML regression methods like LASSO and its variants have shown considerable prowess in the context of economic forecasting. These models are capable of processing hundreds of predictors simultaneously, crucial for extracting useful information for accurate predictions [LC14].

A body of research is also dedicated to testing multiple regression models simultaneously to ascertain their relative effectiveness. These comparative studies are instrumental in identifying the optimal model for specific forecasting tasks, particularly in fields like real estate where variables and market dynamics can vary greatly. Such an approach

Statistics Estonia database was 3,871 million euros, resulting in Nordecon's market share of 7.03%.

not only allows for a broader evaluation of potential tools but also enhances the understanding of how different models capture and analyze variability within the data [MAP19, MGN20].

When discussing regression models, the importance of hyperparameter optimisation cannot be overstated as it significantly enhances model performance. Tuning the hyperparameters to suit different problems is essential because the optimal configuration directly impacts the effectiveness of ML models [YS20]. By integrating hyperparameter optimisation into the ML pipeline, models can be precisely adjusted to align with the unique characteristics of the data. This calibration not only boosts the model's precision but also substantially improves its predictive accuracy across various conditions.

Feature engineering is a pivotal process in the ML pipeline, enabling data scientists to significantly enhance model accuracy. By transforming raw data into clear, well-defined inputs, feature engineering highlights essential patterns and correlations, thus improving the predictive power of models. Since all data in ML ultimately needs to be represented as numeric features, these transformations are crucial for enabling ML models to operate more effectively. This makes feature engineering not just beneficial but essential for optimising the performance of ML systems across various applications [ZC18].

Building on these approaches and addressing the complexities of forecasting within the construction industry, this research is driven by several critical questions: How can ML improve the accuracy of sales forecasts in the construction industry? Which ML techniques are most effective for managing the time-series data inherent in construction industry sales forecasts? To what extent does the integration of macroeconomic indicators into predictive models enhance forecasting accuracy? And, how do ML forecasts compare with traditional methods in terms of reliability and precision within the construction sector?

To answer these questions, the objectives of this research are to develop and evaluate a variety of ML models specifically tailored to effectively forecast construction sales. This involves investigating the incorporation of macroeconomic indicators into these models to improve predictive accuracy. Additionally, the study aims to compare the performance of these advanced models against traditional forecasting methods and provide strategic recommendations based on the outcomes to support decision-making within the construction industry.

The thesis is organised as follows. Chapter 2 delves into the literature, providing a critical review of principles and relevant methods in forecasting sales as time series using ML. Also, in this section we give a technical overview of modelling preliminaries used in our experiments. Chapter 3 details the methodology, outlining our approach for integrating macroeconomic indicators with historical sales data to refine sales forecasting models. This chapter also describes the design of our experiments and the data preparation process for Nordecon AS. In Chapter 4, we present our findings, highlighting the predictive performance achieved through our models on different datasets. This is

followed by Chapter 5, which offers concluding remarks, reflecting on the implications of our research findings and suggesting future research directions. Each section builds on the previous one, aiming to provide a clear and logical progression of research activities and insights.

1.3 Scope and Limitations

The scope of this research is focused on forecasting sales in the Estonian construction industry, utilising historical sales data and macroeconomic indicators relevant to this market. While the findings may offer valuable insights, their applicability might be limited to similar economic contexts and may not be directly transferrable to other industries or regions with different economic dynamics. Furthermore, the accuracy of the forecasts is contingent upon the quality and granularity of the data available, and the models' performance might be constrained by the inherent limitations of the ML techniques used.

2 Background

2.1 Literature Overview

2.1.1 Sales Forecasting

The significance of forecasting within operations management cannot be overemphasised, encompassing a wide array of functions such as resource allocation, target setting, and synchronising various organisational activities [OW09]. As depicted on Figure 3 from [OW09], through a synthesised approach combining top-down[6], bottom-up[7], and statistical forecasts, a consensus forecast is crafted, which addresses the multifaceted organisational challenges in creating and refining forecasts. The amalgamation of different forecasting methods is not only strategic but also practical in reducing bias and improving overall accuracy [Arm01].

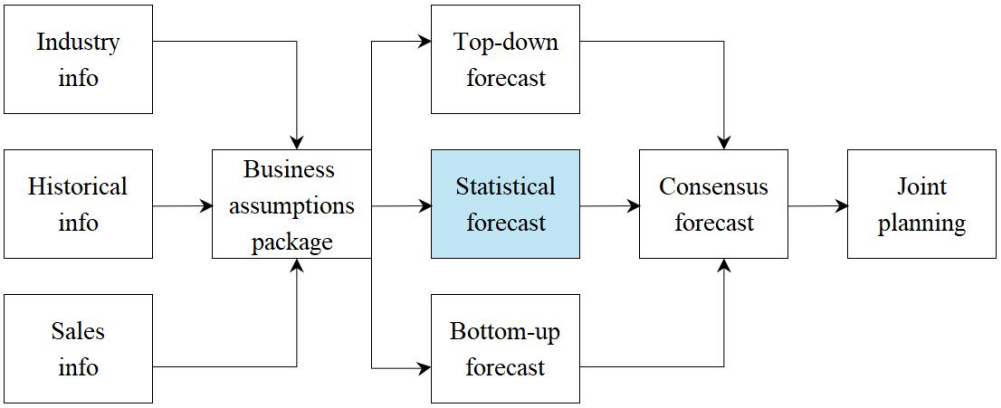


Figure 3. Corporate forecast process

Corporate sales forecasting serves diverse functions, spanning marketing, sales, finance, production, and logistics. However, there is often confusion between forecasting, planning, and target-setting within companies. Forecasting is defined as the projection of expected demand based on a set of environmental conditions, while planning and target-setting involve a suite of managerial actions [MM04].

[6]Top-down forecasting is a method of estimating a company's future performance by starting with high-level market data (like its market share) and working "down" on sales revenue (https://corporatefinanceinstitute.com/resources/financial-modeling/top-down-forecasting/).

[7]Bottom-up forecasting is a method of estimating a company's future performance by starting with low-level company data and working "up" to revenue. This approach starts with detailed customer or product information and then broadens up to revenue (https://corporatefinanceinstitute.com/resources/financial-modeling/top-down-forecasting/).

2.1.2 Economic Indicators as Independent Variables

As illustrated in Figure 3, the process of sales forecasting in corporate settings integrates various data sources, including industry and historical sales information, to form a comprehensive business assumptions package. This synthesised approach facilitates both top-down and bottom-up forecasts, which are then merged into a statistical forecast. Understanding this interconnection highlights the critical role of macroeconomic indicators in refining these forecasts. These indicators not only enhance the statistical forecasting model by providing a broader economic context but also assist in aligning the forecast with real-world economic dynamics. This integration is essential for developing robust, adaptable forecasting models that are capable of interpreting and reacting to economic signals, thereby making them crucial tools for navigating the complexities of market fluctuations and unforeseen economic disruptions.

The relationship between the financial valuation of construction work and macroeconomic indicators is crucial for understanding market dynamics. Research has highlighted that variations in economic indicators, notably the GDP and property price indices, can significantly explain the completion rates of construction work in the private residential sector [SELL15]. This interplay underscores the susceptibility of the construction industry to broader economic trends and the need for adaptive forecasting models that can interpret these economic signals. This necessity often requires lagging the data[8] or forecasting the independent variables[9] to establish a robust predictive foundation [WC16].

Furthermore, the robustness of construction cost forecasting, particularly in the face of unforeseen events such as pandemics and geopolitical conflicts, has been brought into question [Ayd24]. This study examines the impact of such disturbances on forecasting sales performance across selected European countries and reveales a pronounced increase in forecast errors during periods of instability. These findings accentuate the necessity for forecasting methodologies to be versatile and resilient enough to capture the changes in economic trends and incorporate the influence of external factors such as macroeconomic indicators.

[8]Two sequential observations in data can be correlated, ie. the earlier observation contributes to predicting the next [WC16]. Such earlier observations can be defined as lagged data of the current observation.

[9]A variable is considered dependent if it depends on an independent variable. Dependent variables are studied under the supposition that they depend, by some law or rule (e.g., by a mathematical function), on the values of other variables. Independent variables, in turn, are not seen as depending on any other variable in the scope of the experiment in question (`https://en.wikipedia.org/wiki/Dependent_and_independent_variables`).

2.1.3 Time Series in Sales Forecasting

Time series analysis is crucial in sales forecasting, involving the sequential recording of sales data at consistent intervals. This analysis helps in decomposing the time series into systematic and unsystematic components. The systematic components include the level, trend, and seasonality[10], which are predictable and repeatable, and thus, can be modeled for future forecasting. The non-systematic component, also known as noise, consists of random variations that cannot be directly forecasted but are important for understanding data variability [Bro17].

Sales can be forecasted using various methodologies, primarily distinguished as univariate and multivariate approaches. Univariate forecasting focuses solely on the historical data of a single variable to predict future values, utilising techniques such as trend analysis and extrapolation that capitalise on the inherent autocorrelation within revenue data [IHCD22]. Multivariate forecasting methods, on the other hand, incorporate multiple independent variables or series to predict a dependent variable. These approaches require that the values of the independent variables be known or reliably predicted for the forecast period, allowing for a more comprehensive analysis that factors in external influences on sales trends [Auf21].

Forecasting techniques, particularly multivariate methods, are pivotal in identifying the relationships between sales and various independent exogenous variables[11] that influence them [MM04]. These methods are essential for exploring the dynamics that drive sales performance within a changing economic environment.

Further enhancing these multivariate approaches, the combination of regression analysis with time series modeling has proven particularly effective. This integrated approach significantly improves the accuracy of predictions, as demonstrated in the forecasting of construction tender price indexes. By blending these methods, the model achieves superior performance over individual techniques, especially for short-term forecasts spanning one to two quarters [STNW04].

Building on the integration of regression and time series modeling, the utilisation of leading indicators[12] further enhances forecasting methodologies [AAH98]. These indicators, characterised by their strong correlation and predictive relationship with key economic measures, provide essential insights for anticipatory modeling. By effectively identifying and applying leading indicators, forecasters can equip their models with the ability to preemptively adjust to shifts in market dynamics, thus offering a substantial advantage in the accuracy and relevance of forecasts [BAL12].

[10] Specifically, the level refers to the average value in the series, the trend indicates an increasing or decreasing pattern, and seasonality represents the recurring short-term cycles influenced by seasonal factors.

[11] In the context of our work macroeconomic indicators can be viewed as exogenous variables.

[12] A leading indicator is a measurable set of data that may help to forecast future economic activity. Leading economic indicators can be used to predict changes in the economy before the economy begins to shift in a particular direction (`https://www.investopedia.com/terms/l/leadingindicator.asp`).

2.1.4 Machine Learning Considerations in Sales Forecasting

Building upon the advanced forecasting methods discussed earlier, ML introduces a dynamic component to sales forecasting, albeit with varying degrees of success when compared to traditional methods. [MSA18] reports that traditional statistical methods often still outperform ML models in terms of post-sample accuracy[13]. This study highlights not only the effectiveness of statistical techniques across various accuracy measures and forecasting horizons but also points out their relatively modest computational demands in contrast to ML models.

Despite the robust performance of statistical methods, the evolving capabilities of ML in sales forecasting remain significant. For example, the utilisation of the XGBoost ML model [Niu20] has shown promising results, particularly when combined with comprehensive feature engineering, as demonstrated by Walmart's approach to sales forecasting. This suggests that with careful data preparation, ML models can effectively enhance forecasting accuracy.

A critical aspect of deploying both statistical and ML models in sales forecasting is maintaining a balance between pattern recognition and the avoidance of overfitting. Overfitting occurs when a model is too closely fitted to the limited data it was trained on, making it less effective at predicting new, unseen data [PF13]. To mitigate this, techniques such as model regularisation are employed. This includes methods like tree pruning in tree-based models, selective feature inclusion, and the incorporation of complexity penalties into the model's objective function, all aimed at simplifying the model sufficiently to ensure robust performance without unnecessary complexity [PF13].

Furthermore, the application of cross-validation as a method for out-of-sample evaluation is a cornerstone in the use of ML models, helping to prevent overfitting by testing the model's ability to generalise to new data. However, the assumption that data samples are independent and identically distributed (i.i.d.), which underlies traditional cross-validation, does not hold in many real-world scenarios, including time series data [CTM20]. This mismatch can disrupt the temporal order of the data, leading to potentially unrealistic performance estimates. To address this challenge, forecasting practitioners often apply techniques like rolling window evaluation, which respects the chronological sequence of time series data and offers a more realistic gauge of a model's predictive accuracy [HA21].

The integration of ML into sales forecasting is a complex yet promising frontier that requires careful consideration of methodological choices to harness its full potential effectively.

[13] Can be defined as accuracy calculated on samples out of data the model was trained on, such as validation and test sets.

2.1.5 Similar Works

Simpler time series forecasting methods primarily use historical sales data to predict future trends. However, these methods often fail to capture macroeconomic changes that significantly impact demand within the business environment. To address these limitations, some companies manually adjust their statistical forecasts or rely on expert opinions, although such adjustments are subject to bias and can be costly. In response to these challenges, [VAD20] introduced an innovative forecasting framework that effectively incorporates a broad set of macroeconomic indicators for short-term sales forecasting, ie forecast horizon of one year with monthly steps. This novel framework automates the selection of relevant variables and facilitates the prediction of future sales. It combines the seasonal naive method [HA21] for estimating the seasonal component with a LASSO regression technique [Tib18] that utilises macroeconomic indicators to capture the influence of changes in economy. By optimising the size of the independent variables set, the framework aims to maintain simplicity while enhancing forecasting accuracy. The effectiveness of this approach is demonstrated by a significant reduction in MAPE [HA21] compared to naive forecasting methods[14], underscoring the benefits of integrating macroeconomic insights into sales forecasting. The naive forecasting methods serve as benchmarks for these types of studies, providing a foundation for comparing more complex models [PF13].

Similarly, the work [SAKD18] also underscores the significance of macroeconomic indicators in enhancing forecasting methodologies. Their study illustrates the effectiveness of a LASSO regression model that integrates seasonal[15] dummy variables[16], autoregressed[17] sales, and macroeconomic variables, thereby significantly improving short-term sales predictions. The proposed statistical forecast in their methodology not only refines the models but also deepens the understanding of market dynamics impacting sales by automatically selecting both the type of leading indicators and the order of the lead for each selected indicator. Moreover, their approach addresses the inherent uncertainty introduced by the unknown future values of these leading indicators. This uncertainty is managed by employing an unconditional forecasting setup that does not assume prior knowledge of future indicator values, contrasting with a conditional setup where such values are presumed known. This distinction allows for an evaluation of the theoretical loss in forecast accuracy between the two setups, providing a more

[14] In naive forecasting method, all forecasts are set to be the value of the last observation, without attempting to establish correlating factors.

[15] According to Cambridge Dictionary, seasonal means relating to or happening during a particular period in the year.

[16] In regression analysis, a dummy variable (also known as indicator variable or just dummy) is one that takes a binary value (0 or 1) to indicate the absence or presence of some categorical effect that may be expected to shift the outcome (`https://en.wikipedia.org/wiki/Dummy_variable_(statistics)`).

[17] The autoregressive model specifies that the output variable depends linearly on its own previous values (`https://en.wikipedia.org/wiki/Autoregressive_model`).

robust and realistic assessment of the predictive power of their model.

2.2 Modelling Preliminaries

In this section, we delve into the foundational aspects of the statistical modelling techniques employed in this study. Our focus centers on a suite of regression models, each chosen for its unique ability to address specific challenges within data analysis. These models include Ridge, ElasticNet, XGBoost, Random Forest, and K-Nearest Neighbors, which are pivotal in handling issues ranging from multicollinearity[18] to overfitting. Additionally, this section outlines the application of Hyperparameter Optimisation and the Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model for time series analysis, providing a robust framework for both predictive modeling and forecasting. We also discuss the Evaluation Metric used to assess the effectiveness of these models, emphasising our commitment to accuracy and reliability in predictions.

2.2.1 Point Forecasts

In time-series forecasting, we aim to predict future values based on patterns observed in past data. This involves analyzing time-ordered data points to understand trends, seasonality, and other dynamics that influence future outcomes. Within this framework, point forecasts serve as a targeted prediction method where future values are estimated at specific times, making them crucial for decision-making processes. By focusing on point forecasts, we apply the principles of time-series analysis to produce specific predicted values for future periods based on the information available at a given time t.

As described in [MCM$^+$13], the forecast process involves generating predictions about future values based on the data available at a given time t. A point forecast, denoted as $\hat{y}_{t+k|t}$, represents a single predicted value for a future time $t+k$ based on the information available at time t.

Point forecasts are typically presented as the conditional expectation of the future value, given the current state of knowledge and any relevant parameters. Mathematically, it is expressed as:

$$\hat{y}_{t+k|t} = \mathbb{E}[Y_{t+k}|g, \Omega_t, \hat{\Theta}], \tag{1}$$

where Ω_t encapsulates all relevant information available up to time t, g represents any model, and $\hat{\Theta}$ denotes estimated parameters influencing the forecast.

Although point forecasts provide a specific predicted value, they inherently differ from probabilistic forecasts. Within the framework of stochastic processes, a point forecast provides a precise predictive value but simultaneously acknowledges the inherent

[18]Multicollinearity denotes when independent variables in a linear regression equation are correlated. Multicollinear variables can negatively affect model predictions on unseen data. Several regularisation techniques can detect and fix multicollinearity (`https://www.ibm.com/topics/multicollinearity`).

uncertainty associated with future predictions. This recognition does not measure the uncertainty in numerical terms; instead, it suggests that the forecasted figure should be interpreted as an estimate—an average of potential future scenarios as understood from present data [MCM+13].

2.2.2 Regression Models

Ridge Regression is one of the pivotal linear models in the realm of regression analysis. This method is specifically tailored to address the issue of multicollinearity within a dataset, which it achieves by introducing a penalty term into the ordinary least squares (OLS) objective function. Developed to extend the capabilities of OLS regression, Ridge regression strives to mitigate the risk of overfitting. It does this by applying a constraint to the regression model's coefficients, thereby preventing them from reaching large magnitudes that could otherwise lead to model sensitivity to small fluctuations in the data [KB82].

The objective function of Ridge regression can be expressed mathematically, underscored by the regularisation parameter λ, which is pivotal in adjusting the degree of shrinkage applied to the coefficients. This regularisation term plays a critical role in balancing the trade-off between model variance and bias. The Ridge regression objective function is given by:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{2}$$

where:

- $\hat{\beta}^{\text{ridge}}$ represents the estimated coefficients of the Ridge regression model,
- N denotes the number of observations in the dataset,
- p represents the number of predictors or features in the dataset,
- y_i represents the observed value of the dependent variable for the i-th observation,
- X_i represents the row vector of predictor variables for the i-th observation,
- β represents the coefficient vector to be estimated,
- λ is the regularisation parameter, also known as the penalty term, controlling the amount of shrinkage applied to the coefficients.

Higher values of λ correspond to greater shrinkage, effectively diminishing the variance of the Ridge regression model, albeit at the expense of an increase in bias.

Elastic Net distinguishes itself by harmonising the strengths of both Ridge and LASSO regression through the incorporation of dual penalty terms in the ordinary least squares (OLS) objective function. This technique adeptly manages multicollinearity and executes variable selection, enabling certain coefficients to shrink towards zero while others remain nonzero. The efficacy of the Elastic Net is regulated by two regularisation parameters, λ_1 and λ_2, which dictate the degree of shrinkage. These parameters are instrumental in modulating the model's variance and bias [ZH05].

The objective function of Elastic Net is mathematically formulated to optimise the balance between Ridge and LASSO penalties and can be represented as:

$$\hat{\beta}^{\text{elasticnet}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - X_i\beta)^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| \right\} \tag{3}$$

where:

- $\hat{\beta}^{\text{elasticnet}}$ denotes the estimated coefficients of the Elastic Net regression model,
- N signifies the number of observations in the dataset,
- p indicates the number of predictors or features within the dataset,
- y_i is the observed value of the dependent variable for the i-th observation,
- X_i represents the row vector of predictor variables for the i-th observation,
- β encapsulates the coefficient vector that is to be estimated,
- λ_1 and λ_2 are the regularisation parameters that control the extent of shrinkage applied to the coefficients, corresponding to the Ridge and LASSO penalties, respectively.

Elevated values of λ_1 and λ_2 contribute to greater shrinkage, which effectively curtails the model's variance, thereby incrementing the bias in the model for an enhanced balance and performance.

XGBoost (Extreme Gradient Boosting), as a prominent member of the ensemble learning family, effectively utilises a sequence of weak learners—typically decision trees—to iteratively correct the errors of preceding models. This method is particularly adept at handling various types of predictive modeling challenges, including both regression and classification tasks. XGBoost is distinguished by its inclusion of L1 (LASSO) and L2 (Ridge) regularisation terms in the objective function, which serve to penalise the

complexity of the model and thus prevent overfitting while enhancing the algorithm's generalisation capabilities [CG16].

The mathematical expression for XGBoost's objective function, which seeks to minimise the loss function while also considering the regularisation terms, is as follows:

$$\text{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{4}$$

In this function:

- Obj denotes the objective function that XGBoost aims to minimise.
- n signifies the number of samples in the dataset.
- y_i and $\hat{y}_i$ represent the true and the predicted values of the target variable for the i-th sample, respectively.
- $L(y_i, \hat{y}_i)$ is the loss function that quantifies the discrepancy between the actual and the predicted values.
- K corresponds to the number of trees within the ensemble.
- $\Omega(f_k)$ refers to the regularisation term which penalises the complexity of the k-th tree in the ensemble, thus contributing to the prevention of overfitting.

The XGBoost algorithm, through this well-defined objective function, robustly navigates the trade-off between fidelity to the training data and the simplicity of the model, resulting in a powerful and versatile ML tool.

Random Forest exemplifies an ensemble learning method that operates through the formation of a multitude of decision trees. During training, each tree is independently cultivated on a randomly selected subset of the data. This inherent randomness is pivotal in mitigating overfitting and enhancing the model's ability to generalise to new data. The predictive strength of Random Forest lies in its ability to aggregate the predictions of all individual trees, taking the mode for classification tasks or the mean for regression [Bre01].

Random Forest utilises the bagging technique, which involves random sampling with replacement from the training dataset to generate multiple diverse subsets. This leads to the construction of distinct and decorrelated decision trees. The collective prediction of the Random Forest model for regression tasks is quantified as the average of the predictions from all trees within the ensemble. Mathematically, this is represented as:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{5}$$

where:

- $\hat{y}$ signifies the predicted value of the target variable, obtained by averaging the predictions of the ensemble.
- N corresponds to the number of decision trees included in the Random Forest ensemble.
- y_i is the prediction rendered by the i-th decision tree.

By capitalising on the predictions of numerous trees, Random Forest efficiently addresses the variance that typically affects a single decision tree, thereby enhancing prediction reliability and model robustness.

K-Nearest Neighbors (KNN) is an instance-based learning algorithm that departs from traditional regression methods by utilising the proximity of data points to predict outcomes. Unlike models that infer explicit functional relationships between variables, KNN operates on the premise that the entire dataset is the source of truth for prediction, eschewing any assumptions about the underlying data distribution. It leverages a distance metric—such as Euclidean or Manhattan distance—to ascertain the 'k' closest instances to the query point and bases its predictions on either the average or the majority vote among these neighbors [GCS14].

The KNN Regressor, suitable for regression tasks, identifies the K most similar instances to a given query point and computes predictions by averaging their target values. This non-parametric approach relies on the full breadth of the training data for generating predictions. The parameter K, indicative of the number of neighbors to consider, is central to the KNN algorithm. The prediction yielded by the KNN Regressor for any given query point is mathematically depicted as follows:

$$\hat{y} = \frac{1}{K} \sum_{i=1}^{K} y_i \tag{6}$$

where:

- $\hat{y}$ denotes the predicted value for the query point, derived by averaging the target values of the nearest neighbors.
- K signifies the count of nearest neighbors that have been elected for the prediction process.
- y_i represents the target value associated with the i-th nearest neighbor.

By averaging over the neighbors, KNN incorporates a localised and interpretable decision-making process, which is robust against noisy data and effective for a wide range of applications.

2.2.3 Hyperparameter Optimisation

In this work, Hyperopt serves as a Python library designed for the purpose of hyperparameter optimisation. It employs Bayesian optimisation, a strategy adept at finding the best parameter configuration for a given model, even when dealing with a vast parameter space [Dav20]. Hyperopt's principal components include:

- **Search Space**: Hyperopt allows the specification of stochastic search spaces for input parameters, offering various functions to define parameter ranges.
- **Objective Function**: This function, which Hyperopt aims to minimise, takes hyperparameters as input and returns the associated loss.
- **fmin Function**: The core of the optimisation, which iteratively explores various hyperparameter combinations to minimise the objective function.
- **Trials Object**: Records all hyperparameter values, losses, and related information throughout the optimisation process.

The hyperparameter optimisation process follows these steps:

1. Initialisation of the search space.
2. Definition of the objective function.
3. Selection of the search algorithm.
4. Execution of the optimisation via the `hyperopt` function.
5. Analysis of the trials object to review the outcomes.

2.2.4 SARIMA Model

The SARIMA model [HA21], designed to address time series with seasonality, is extensively utilised for its ability to model and forecast data where seasonal patterns are prominent. SARIMA extends the ARIMA model by integrating both non-seasonal and seasonal factors. Parameters for SARIMA, denoted as $(p, d, q) \times (P, D, Q)_S$, are meticulously selected using an exhaustive grid search over possible combinations provided by itertools.

The SARIMA model is mathematically formulated as follows:

$$\text{SARIMA}\,(p, d, q) \times (P, D, Q)_S \tag{7}$$

where:

- p represents the order of the non-seasonal autoregressive (AR) part,
- d the degree of non-seasonal differencing,
- q the order of the non-seasonal moving average (MA) part,
- P the order of the seasonal autoregressive part,
- D the degree of seasonal differencing,
- Q the order of the seasonal moving average part,
- S the length of the seasonal cycle.

The non-seasonal AR and MA components of the model are given by:

$$\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p, \quad \theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q \tag{8}$$

and the seasonal components are:

$$\Phi(B^S) = 1 - \Phi_1 B^S - \ldots - \Phi_P B^{PS}, \quad \Theta(B^S) = 1 + \Theta_1 B^S + \ldots + \Theta_Q B^{QS} \tag{9}$$

2.2.5 Evaluation Metric

The efficacy of the regression models and baseline forecasting approaches is quantified using the Mean Absolute Percentage Error (MAPE). MAPE is an intuitive metric that measures the average magnitude of errors in a set of predictions, without considering their direction. It calculates the average percentage deviation between the forecasted values and the actual observed figures. This metric is particularly useful in providing a clear, percentage-based evaluation of forecast accuracy, which facilitates comparability across different models and datasets [HA21].

The formula for MAPE is given by:

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \tag{10}$$

In Equation 10, A_t represents the actual value at time t, and F_t denotes the forecast value at the same time point. The absolute value of the percentage error for each predicted point is aggregated across the entire forecast horizon and then divided by the number of predictions n to yield the mean.

3 Methodology

This chapter outlines the methodology employed to forecast sales for Nordecon AS by integrating macroeconomic indicators with historical sales data to enhance prediction accuracy and economic relevance. Specifically, our study focuses on point forecasts rather than probabilistic forecasts. Our approach utilises a mix of more traditional forecasting techniques (univariate models based on historical sales) and advanced ML models (multivariate regression techniques), capturing both micro-level business activities and macro-level economic trends that influence the construction sector. We deploy five pre-selected regression algorithms—Ridge, ElasticNet, XGBoost, Random Forest, and K-Nearest Neighbors—across multiple datasets enriched through feature engineering to minimise forecast error and variance. Additionally, hyperparameter optimisation further refines each model's performance, with results evaluated against a material significance threshold of 10% as defined by Nasdaq Tallinn. By synthesising diverse data sources and optimising model parameters, our methodology aims to provide robust quarterly forecasts that support strategic decision-making for Nordecon AS in a volatile market environment.

3.1 Data

The data utilised in this study encompass Nordecon AS quarterly sales data in Estonia as a target variable and input to historical sales-based modeling, as well as a selection of Estonia's quarterly macroeconomic indicators as attributes to regression modeling. As the company provided its quarterly sales along with order book[19] data (one of the sector specific leading variables), sources of macroeconomic data are Statistics Estonia[20], Bank of Estonia[21], Land Board[22], Estonian Institute of Economic Research (EKI)[23] and European Commission (EC)[24].

The initial dataset starts from the first quarter of 2006 and extends to the fourth quarter of 2023 in the case of Nordecon's sales and the third quarter of 2023 in the case of macroeconomic indicators, respectively. Year 2006 is selected as a starting point because this is the year Nordecon AS was listed on the Nasdaq Tallinn Stock Exchange[25], meaning that the company data is publicly available from this point forward on a comparable basis.

We selected the macroeconomic data to represent the main aspects of a country's

[19] An order book is the aggregate sale value of all construction works that have not yet been performed.
[20] https://andmed.stat.ee/en/stat
[21] https://statistika.eestipank.ee/#/en
[22] https://www.maaamet.ee/kinnisvara/htraru/
[23] https://www.ki.ee/en/index.html
[24] https://economy-finance.ec.europa.eu/economic-forecast-and-surveys_en
[25] https://nasdaqbaltic.com/statistics/en/instrument/EE3100039496/trading

economy as presented in the reports by main state institutions like the Ministry of Finance[26] or Bank of Estonia[27] and based on our domain expertise from financial analysis. We selected 36 indicators from the leading economic topics described in Table 1. In light of the fact that the company's contribution to Estonia's total annual volume of construction work has during last 18 years been in the range of 7-9.6%, we assessed it as sufficiently substantial to justify the exclusion of direct construction sector indicators from the analysis. We made this methodological decision to mitigate any potential bias that the company's financial performance might impart on these indicators. Consequently, the study focused exclusively on macroeconomic indicators that bear an indirect relation to the construction sector, thereby ensuring a more objective evaluation of external economic factors.

All the retrieved data extends back further than the first quarter of 2006 so there is no missing data involved in original historic data. All the data is regularly updated and made available to general public by the respective authorities with maximum delay of approximately two months after end of each quarter. All the data is numerical.

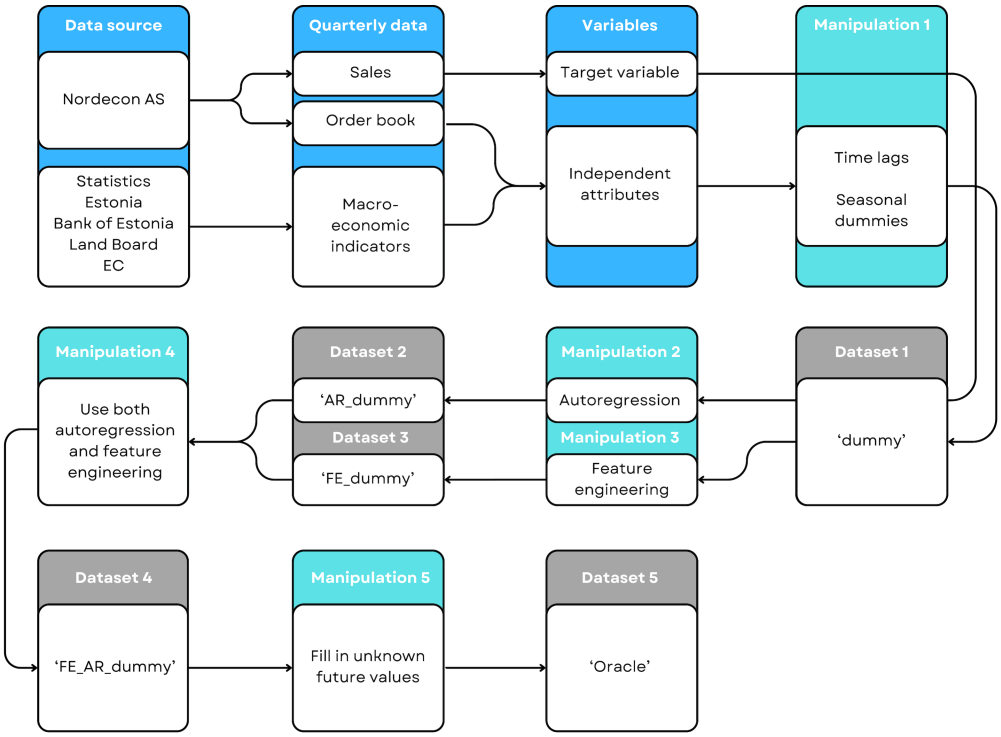


Figure 4. Data preparation process

[26] https://www.fin.ee/en/public-finances-and-taxes/state-budget-and-economy/annual-state-budget-and-economic-forecasts

[27] https://www.eestipank.ee/publications/estonian-economy-and-monetary-policy

Table 1. Data used for modelling

Variable	Unit	Source	Frequency	Availability
Nordecon AS				
Consolidated sales in Estonia	th. euros	Nordecon AS	Monthly	Apr. 2002
Order book	th. euros	Nordecon AS	Quarterly	Q4 2004
GDP at constant prices				
Private consumption	mil. euros	Statistics Estonia	Quarterly	Q1 1995
Government sector consumption	mil. euros	Statistics Estonia	Quarterly	Q1 1995
Gross fixed capital formation	mil. euros	Statistics Estonia	Quarterly	Q1 1995
Export of goods and services	mil. euros	Statistics Estonia	Quarterly	Q1 1995
Import of goods and services	mil. euros	Statistics Estonia	Quarterly	Q1 1995
Labour market				
Unemployment rate	percentage	Statistics Estonia	Quarterly	Q1 2000
Employment rate	percentage	Statistics Estonia	Quarterly	Q1 2000
Average gross monthly wage	euros	Statistics Estonia	Quarterly	Q1 2002
Inflation				
Consumer price index	percentage	Statistics Estonia	Monthly	Jan. 1998
Producer price index	percentage	Statistics Estonia	Monthly	Jan. 2002
Balance of payments				
Current account balance	mil. euros	Bank of Estonia	Quarterly	Q1 1993
Current account balance to GDP	percentage	Bank of Estonia	Quarterly	Q1 2000
Foreign direct investments	mil. euros	Bank of Estonia	Quarterly	Q1 1993
State budget				
State budget surplus/deficit	mil. euros	Statistics Estonia	Quarterly	Q1 1999
Government budget balance to GDP	percentage	Statistics Estonia	Quarterly	Q1 1999
Government sector investments	mil. euros	Statistics Estonia	Quarterly	Q1 1999
Confidence indicators				
Industrial enterprises' confidence	index	EKI	Monthly	Jan. 1995
Retail trade enterprises' confidence	index	EKI	Monthly	Jan. 1995
Services enterprises' confidence	index	EKI	Monthly	Jan. 1995
Consumer confidence	index	EKI	Monthly	Jan. 1995
Economic confidence index	index	EKI	Monthly	Jan. 1995
Demand in industry as limiting factor	index	EC	Quarterly	Q1 1995
Demand in services as limiting factor	index	EC	Quarterly	Q3 2003
Volume indices				
Retail trade volume index	Index	Statistics Estonia	Monthly	Jan. 2001
Industrial production volume index	Index	Statistics Estonia	Monthly	Jan. 2000
New car registrations	pcs	Statistics Estonia	Monthly	Jan. 1993
Sale of apartments	pcs	Land Board	Monthly	Jun. 2003
Building permits				
Residential building permits	th. sqm	Statistics Estonia	Quarterly	Q1 1996
Non-residential building permits	th. sqm	Statistics Estonia	Quarterly	Q1 2000
Productivity				
Labor productivity per person	percentage	Statistics Estonia	Quarterly	Q1 1995
Labor productivity per worked hour	percentage	Statistics Estonia	Quarterly	Q1 1995
Money and corporate financing				
Total corporate profit	th. euros	Statistics Estonia	Quarterly	Q1 2001
Corporate investments into buildings	th. euros	Statistics Estonia	Quarterly	Q1 2001
Loans granted to non-fin enterprises	mil. euros	Bank of Estonia	Quarterly	Q1 1997
M1 (money supply indicator)	mil. euros	Bank of Estonia	Quarterly	Q1 2004

The flow chart illustrating the creation of datasets after variable selection to accompany methodology description that follows is presented in Figure 4.

3.2 Time Series Cross Validation

Given the nature of the exercise, it is important to acknowledge that the data under study are time series variables and, therefore, are not independent and identically distributed (i.i.d.). Each data point can be influenced by previous values within each attribute. To accommodate this, we employed a quarter-by-quarter expanding window approach. This technique ensures temporal consistency in the training and testing of the models, respecting the sequence of both the target variable and independent attributes.

As described in Figure 5, for cross validation purposes we divided each dataset into training, validation, and test splits of eight folds. To capture the seasonal patterns within the year, both the validation and test splits span a duration of four quarters. This structure allows the models to be evaluated against the backdrop of a full year's data, thereby integrating the intra-year seasonality into the assessment of model performance.

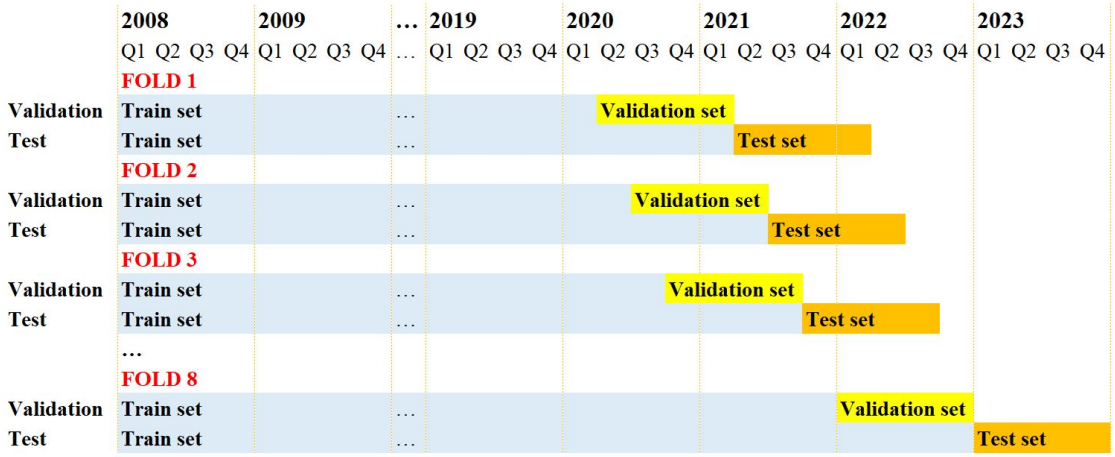


Figure 5. Time series cross validation setup

The selection of eight folds for cross-validation was informed by the need to ensure equal representation of all quarter types - from the first through to the fourth quarter - across the splits, thus providing a balanced view of the data across the two-year period covered by both the validation and test sets.

The cross-validation process commences with the validation set for the first fold beginning in the second quarter of 2020 and the corresponding test set starting in the second quarter of 2021. For the eighth and final fold, the validation set initiates in the first quarter of 2022, with the test set following in the first quarter of 2023.

3.3 Creation of Datasets

The goal of this research is to devise predictive models that can forecast future trends. To this end, we construct time lags for the independent variables, extending up to eight quarters—or two years—into the future. This period is chosen based on the understanding that macroeconomic influences typically do not extend beyond this timeframe, allowing us to focus on the most relevant periods for potential impact on sales outcomes[28]. We employ cross-correlation analysis to examine the relationships between sales and various time-shifted versions of each macroeconomic indicator. By assessing the correlation over multiple lag periods for each indicator, we identify the specific time lags that most strongly correlate with sales. Figures 6 and 7 visually depict these correlations and clarify how different macroeconomic indicators exhibit their strongest correlations with sales data at varying lag intervals, thus uniquely influencing the forecast within distinct temporal contexts.

3.3.1 Unknown Future Values

The models' primary goal is to forecast future, yet-to-be-observed periods, akin to real-world prediction scenarios. For instance, estimating fourth quarter 2023 sales using actual data up to third quarter 2023 requires dealing with the issue of data gaps created by the generation of lagged variables. For example, while third quarter 2023 data serves as the lag one value for fourth quarter 2023, there's no available lag one value for first quarter 2024, since the actual fourth quarter 2023 figures are unknown at the time of forecasting, leading to a unknown future value. Additionally, the reason behind this unknown future value primarily stems from the way macroeconomic data is published, which is reflected in our dataset. Typically, to fill these unknown future values, one would need to forecast them based on available data, a task that extends beyond the scope of this study. Therefore, we utilise the 'Oracle' dataset described below, which presupposes access to future data, thus sidestepping the challenge of forecasting unknown future values and directly addressing the potential impacts of data incompleteness on our analysis.

To address this challenge and simulate realistic forecasting conditions, we intentionally constructed the datasets include unknown future values within both the validation and test sets. This ensures the models are trained, validated, and tested under conditions that reflect actual data availability at the point of forecasting. Subsequently, when forecasting horizon step one, the indicators from lag one up to lag eight can be used, but when forecasting horizon step four, only indicators with lag four to lag eight can be used. In contrast, the 'Oracle' dataset includes actual values for these steps, providing a complete

[28] In our analysis, we consider lags up to eight quarters, corresponding to two years, as macroeconomic events seldom influence beyond this duration. This assumption streamlines our analysis by focusing on time frames most likely to affect sales outcomes.

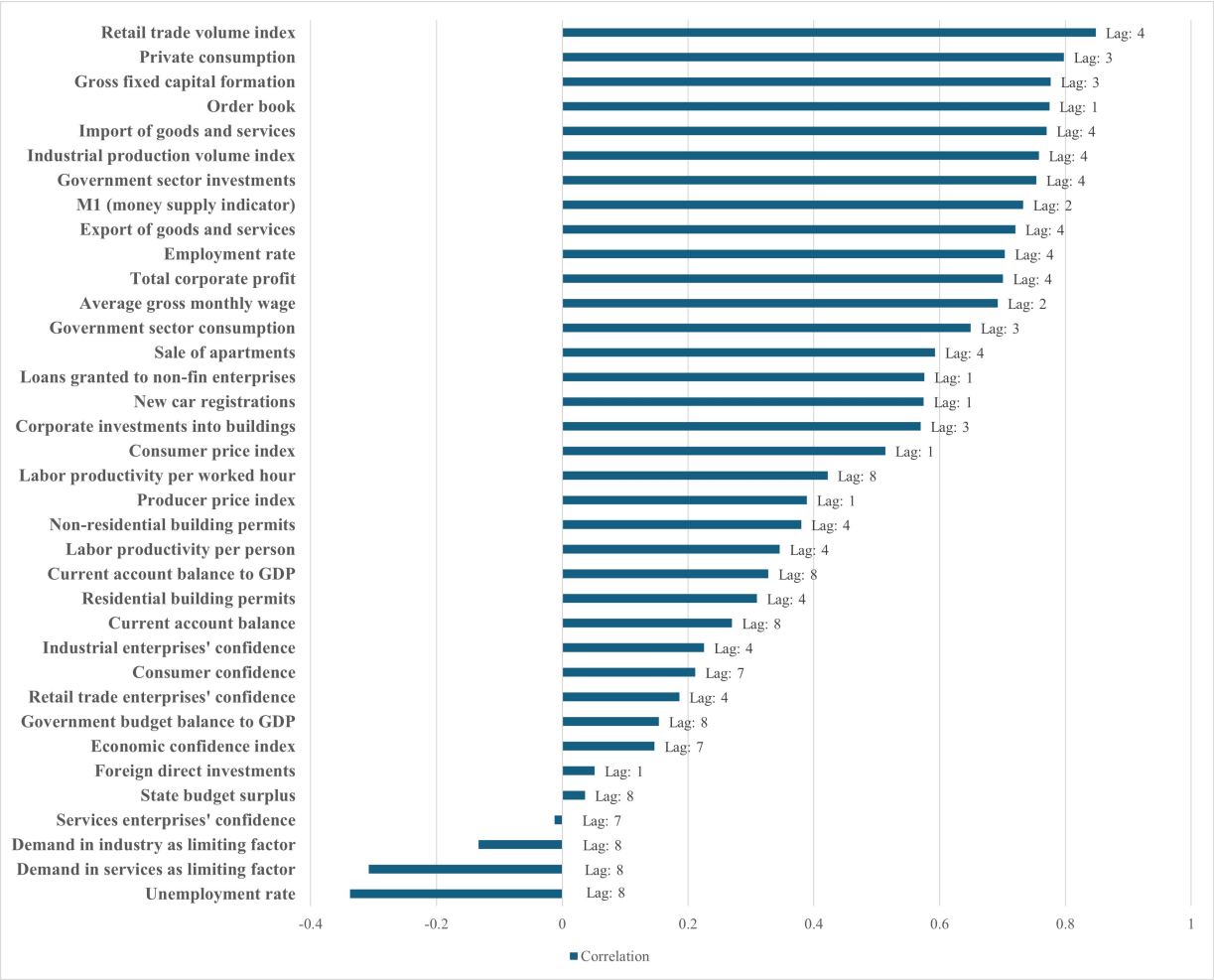


Figure 6. Correlation of the target variable with best lags of independent variables

dataset that bypasses the typical forecasting challenges associated with unknown future values.

3.3.2 Dataset with Encoded Seasonal Variables

The purpose of creating the first dataset is to capture both the short and medium-term effects of economic changes on sales, while aligning the data to reflect the impact of cyclical economic patterns that influence the sales performance.

Consequently for each original attribute we created eight lagged attributes, shifting values forward from one quarter (i.e., Order_book -> Order_book_lag_1) to eight quarters (i.e., Order_book -> Order_book_lag_8). As we used actual data from the first quarter of

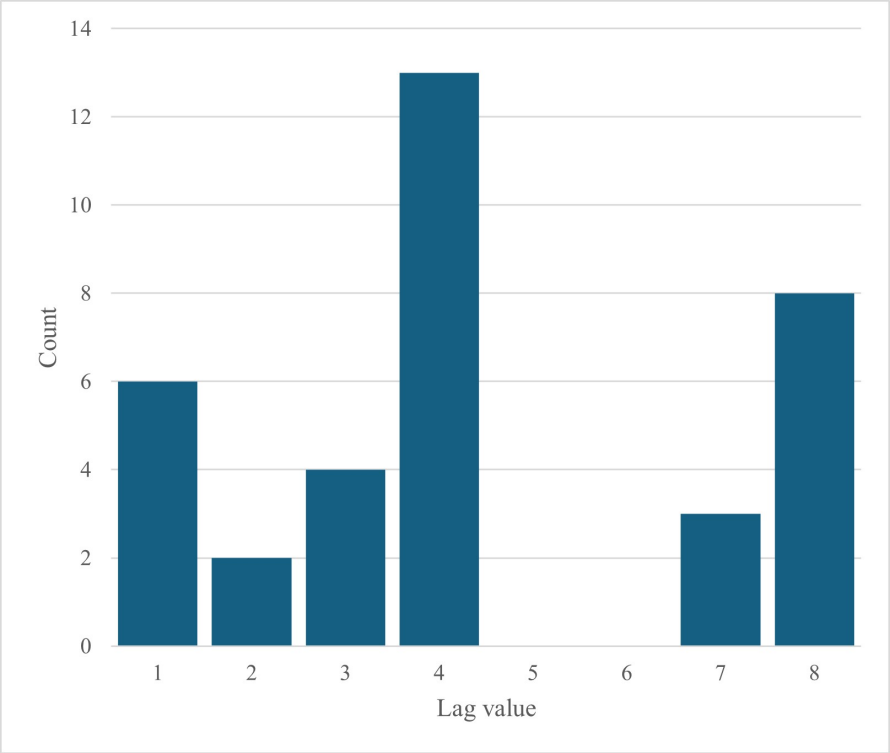


Figure 7. Histogram of best lags

2006 to create these lags, it created missing lag values for the period from quarter one 2006 to quarter four 2007. Subsequently, we removed these missing lag values, so the dataset's starting point is the first quarter of 2008. As the previous economic downturn of 2007-2008 was preceded by intensive economic growth followed by a remarkable decline from 2008 onward, the start of the dataset covers both the intense downfall in sales as the target variable as well as developments in the economic climate from the period 2007 to 2008.

Also, to amplify the effect of seasonality by adding the time feature, we created four encoded attributes by one hot encoding, one for each quarter [HA21]. After removing the original attributes, the first dataset (hereinafter we will refer to this dataset as 'dummy') was created, with eight lagged attributes from each of the original attributes (36 * 8 = 288), seasonal encoded variables (4), and the target variable.

However, the availability of attributes decreases with each forecast horizon step over four quarters due to emergence of unknown future values, as detailed below:

- at the first forecast horizon step, the complete set of 292 attributes is available,
- by the second step, the number has reduced to 256 attributes,

- the third step sees a further reduction to 220 attributes,
- by the fourth step, only 184 attributes remain accessible.

This means that by the time we reach the final step of the forecasting horizon, there's a decrease in a number of available attributes of approximately 37% from the initial data volume available at the first step.

3.3.3 Dataset with Autoregressive Values

The purpose of the second dataset is to enhance the selection of the attributes by capturing the effect of the past values of the target variable, sales. We created dataset, hereinafter called 'AR_dummy', by adding the best autoregression order of the target variable to the 'dummy' dataset as a new independent attribute. Although the Akaike Information Criterion (AIC) is widely used for finding the best order[29] [HK08], we used the autocorrelation function (ACF) to ensure comparability with the SARIMA model. As a result of this analysis, we established the best autoregressive order of four quarters. Hence, 'AR_dummy' had 293 independent attributes plus a target variable available for the forecast horizon step one, which gradually decreases to 185 independent attributes by step four due to generation of unknown future values.

3.3.4 Dataset with Statistically Engineered Features

We created the third dataset by adding engineered features to the 'dummy' dataset. For this, we identified the lags with the best correlation to the target variable. Then, for each of these best correlated lag attributes, we created five new attributes capturing:

- characteristics of time series (moving average of last four quarters, moving standard deviation of last four quarters),
- log transform of the values (where applicable, i.e., not for attributes with zero or negative values), and
- growth indicators (absolute, i.e., the difference between consecutive values, relative, i.e., the ratio between consecutive values).

This dataset is hereinafter called 'FE_dummy' and has 459 independent attributes plus a target variable available for the forecast horizon step one, which gradually decreases to 294 independent attributes by step four due to generation of unknown future values.

[29]The order of an autoregression is the number of immediately preceding values in the series that are used to predict the value at the present time.

3.3.5 Dataset with Autoregressive Values and Engineered Features

To experiment with the effect of both engineered features and the best autoregression value of the target variable, i the fourth dataset, hereinafter called 'FE_AR_dummy', we combined the attributes from the 'AR_dummy' and 'FE_dummy' datasets. 'FE_AR_dummy' has hence 460 independent attributes plus a target variable available for forcast horizon step one, decreasing to 295 independent attributes by step four.

3.3.6 Oracle Dataset

The dataset named 'Oracle' was specifically designed by us to measure the impact of unknown future values by addressing gaps that usually occur when forecasting future values based on past data. In typical scenarios, these gaps—represented by unknown future values—would need to be forecasted or estimated using available data, introducing a degree of uncertainty in predictions. However, for the 'Oracle' dataset, we fill these gaps with actual future data that would only be known in hindsight. This method simulates a scenario where forecasters have complete knowledge of future outcomes, effectively eliminating any uncertainty typically associated with forecasting such lagged values. By doing so, the 'Oracle' dataset serves as a benchmark to understand the maximum potential accuracy of our predictive models if all future inputs were known in advance, helping to delineate the impact of unknown future values on forecast reliability.

'Oracle' is created using both autoregressive and engineered features and has 460 independent attributes with no unknown future values plus a target variable for each forecast horizon step.

3.4 Fitting and Optimising Regression Models

3.4.1 Preliminary Setup

Model Selection: Initially, we assessed a diverse suite of 14 regression algorithms, including Linear Regression, LASSO, Ridge, ElasticNet, Decision Tree, Random Forest, Gradient Boosting, K-Nearest Neighbors, XGBoost, LightGBM, Extra Trees, Huber Regressor, AdaBoost, and Bayesian Ridge. Based on initial tests assessing their performance and their ability to represent different regression strategies, we chose five key algorithms—Ridge, ElasticNet, XGBoost, Random Forest, and K-Nearest Neighbors—for in-depth analysis.

Search Space Definition: To each selected model we assigned a specific search space for hyperparameters, which are detailed in the Appendix I on page 60. These search spaces are designed to cover the most impactful parameters for each model.

3.4.2 Data Preparation and Initial Evaluation

Data Handling: Then we prepared data for time-series cross-validation. The datasets underwent initial splitting into eight folds for training, validation, and testing, reflecting the temporal sequence essential for forecasting.

Baseline Training: We initially trained each model using default hyperparameters across all datasets, employing 10 different random seeds to achieve prediction stability and establish a baseline performance metric using MAPE.

3.4.3 Hyperparameter Optimisation (HPO) Setup

Configuration: Hyperparameter optimisation was then configured by us to run for 300 iterations using the Hyperopt library with a consistent random seed to ensure reproducibility. The optimisation process aimed to minimise the validation MAPE.

Runtime Tracking: A timer was set to track the duration of the HPO process for each model, starting the timer before the optimisation loop.

3.4.4 Optimisation Loop

Iterative Optimisation: For each model, we used the optimisation function fmin from Hyperopt to explore the predefined search space. Each potential set of parameters was evaluated by:

- Training the model on the training split.
- Validating the model on the validation split.
- Tracking if the new parameter set offered an improvement over the best-found parameters in terms of MAPE.

Performance Tracking: Throughout the optimisation, updates on improved scores and parameters were printed to monitor progress and adjustments.

3.4.5 Post-Optimisation Analysis

Best Parameters and Scores: After completing the iterations, we recorded the best parameters and their corresponding scores.

Testing Phase: We then tested each model using the best-found parameters against the test splits to assess generalisation capability. This involved:

- Applying the same data preprocessing steps as in training/validation.

- Evaluating the model performance using MAPE on each test split.
- Calculating and recording the average and standard deviation of the MAPE across all test folds.

3.4.6 Results Compilation and Reporting

Results Aggregation: All results, including the training, validation, and testing MAPEs along with runtime information and best parameters, were compiled into a dataFrame for easy access and comparison.

Comprehensive Summary: A detailed summary highlighting the best training and validation scores, the average test scores, and the time taken for the optimisation was provided at the end of the process for each model.

3.4.7 Final Evaluation

Performance Evaluation: We compared the performance of all optimised models against the baseline models to gauge the improvement. We used the results to decide which model and parameter combination yielded the most accurate forecasts, considering both the average MAPE across folds and the robustness of the model as indicated by the standard deviation of MAPE.

3.5 Handling Unknown Future Values

In the forecasting process, we gave special attention to managing unknown future values, which vary across different forecast horizons. Given that the lengths of X_{val} and X_{test} correspond to four quarters, our model must adapt to each quarter's specific data availability. To accommodate this, we adopted the following approach:

Iteration Over Forecast Horizons: We conducted the model fitting and prediction processes separately for each quarter. This segmentation is crucial because the presence of unknown future values can differ from one quarter to the next, affecting which variables are available for model training and prediction.

Handling Unknown Future Values:

- For each forecast horizon (i.e., each quarter in X_{val} and X_{test}), we isolated the dataset for that specific quarter.
- We identified any columns (predictors) in the dataset that contain unknown future values for that particular quarter.

- We excluded these columns from both the training and validation/test datasets to ensure that the model only uses complete cases for making predictions.

Model Fitting and Prediction:

- With the cleaned dataset (i.e., absent the columns with unknown future values), we fitted the model using the corresponding training data for that forecast horizon.
- The model then makes predictions for the validation or test set of that specific quarter.
- This process ensures that each model's fit and subsequent predictions are tailored to the data characteristics of each specific forecast horizon, accommodating the dynamic nature of available information across different quarters.

Aggregation of Predictions: After the model is separately processed for each quarter, we aggregated the predictions for the four individual quarters. This ensures that each quarter's forecast remains distinct, allowing for a detailed quarter-by-quarter evaluation. We then calculated MAPE for these predictions in comparison to the actual values of each respective quarter, providing a precise measure of the model's accuracy over the entire four-quarter period.

By fitting the regression model individually for each forecast horizon, we effectively address the challenges posed by unknown future values in time-series forecasting. This method ensures that our model's performance remains robust, providing reliable forecasts even when faced with incomplete data due to the inherent time lags in economic reporting.

3.6 Baseline Forecasting Models

In addition to the regression models previously discussed, our analysis employs two naive forecasting approaches and the SARIMA model for comparative purposes. All methods utilise the expanding window cross-validation technique outlined earlier to ensure a consistent and robust evaluation across different model configurations. This method systematically increases the size of the training dataset while maintaining a fixed-sized testing set, allowing us to comprehensively assess each model's predictive power and stability over time.

3.6.1 Naive Forecasting Approaches

Given the strong quarterly seasonality observed in past sales data, we customised the simple naive approach (sNaive) to project future sales based specifically on the sales

from the corresponding quarter of the previous year. This method capitalises on the predictable seasonal patterns that are characteristic of the construction industry.

In contrast, the weighted naive approach (wNaive) that we developed for this study, builds on the sNaive method by incorporating sales data from the same quarter over the last three years. This approach recognises that while the most recent corresponding quarter may hold the greatest relevance, earlier corresponding quarters can still provide valuable insights by capturing longer-term trends in historical sales development. To optimise this method, we fine-tuned weights using a grid search technique to find the best combination that accurately reflects these trends. The optimal weights are determined based on validation splits and subsequently applied to the test set to assess their effectiveness in capturing the inherent seasonality of the data.

3.6.2 SARIMA Model

For the SARIMA model configuration, the ranges of parameters, including the non-seasonal components (p, d, q) and seasonal components (P, D, Q), are first defined. These parameters are critical as they determine the model's ability to capture underlying trends, seasonality, and noise in the data. A grid search approach is employed to optimise these parameters by finding the combination that minimises forecasting errors on the validation set.

During the grid search, MAPE is calculated for each validation period to evaluate the accuracy of the model's forecasts. Additionally, the standard deviation of the MAPE across different validation splits is computed to assess the consistency of the model's performance. This dual metric evaluation helps in ensuring that the model not only fits well on average but also performs consistently across different time periods.

The next phase involves selecting the SARIMA model configuration that offers the lowest MAPE across all validation splits. This model configuration is identified as the optimal one for reliable forecasting. Once selected, this optimal model is then used to forecast future quarters using the test split. The predictions made by the model are compared against the actual sales data, and the MAPE is computed for these test splits. This step quantifies the accuracy of the model's predictions, providing a final assessment of its performance.

3.7 Evaluation Criteria

In assessing the performance of our forecasting models, it is crucial to determine the materiality of the forecasting error in a manner that adheres to both industry standards and specific regulatory requirements. According to regulations set by the Nasdaq Tallinn stock exchange [Nas22], if an issuer's publicly disclosed financial forecast deviates from actual financial results by more than 10%, a prompt adjustment and detailed explanation

of the deviation are required. This regulation highlights the importance of this threshold as a measure of material significance.

For public companies like Nordecon AS, this threshold of 10% serves as a critical benchmark. By comparing our models' MAPE against this threshold, we assess whether the errors in our forecasts are significant enough to potentially impact financial decisions and reporting obligations. This comparison is not merely a measure of model accuracy but is also a compliance check against regulatory expectations that govern financial disclosures.

Furthermore, the regulation [Nas22] stipulates that any forecast issued must clearly outline the underlying assumptions and circumstances. Should these assumptions prove inaccurate, leading to a forecast deviation beyond the 10% threshold, the issuer must provide immediate clarification and update the forecast accordingly. This requirement underscores the necessity for transparency and accountability in financial forecasting practices.

Incorporating these regulatory guidelines into our evaluation criteria ensures that our findings not only align with market standards but also comply with the legal obligations of financial reporting. This approach allows us to benchmark the predictive accuracy of our models in a context directly applicable to Nordecon AS and similar publicly listed companies, ensuring our forecasts are robust, transparent, and regulatory compliant.

3.8 Implementation Details

We conducted data preparation, including min-max normalisation, and model development within a Python environment. The experimental setup for data min-max normalisation, regression models and the computation of performance metrics leveraged the functionalities of the Sklearn and XGBoost libraries, while hyperparameter optimisation was carried out by us using the Hyperopt library. We generated baseline models and SARIMA forecasts using the Statsmodels package, employing the Itertools library for exhaustive grid search optimisation. Statistical analysis was performed with the aid of Scipy libraries. The Jupyter notebooks detailing the complete process are hosted on a GitHub Repository.

Throughout the study, we utilised ChatGPT for smoothing linguistic structures and for code inspection, improving the clarity and integrity of both the written and programming aspects of the research.

4 Results and Discussion

This section presents a comprehensive analysis of the outcomes derived from employing a range of regression models and baseline methods for forecasting the company's sales, emphasising the practical implications and the interpretability of the predictive capabilities. We investigate the performance of baseline models, inclusive of the SARIMA model, and delve into the efficacy of diverse regression algorithms across multiple datasets. Additionally, we assess the impact of hyperparameter optimisation on model performance, offering insights into its role in enhancing predictive accuracy.

4.1 Target Variable

The time series data representing Nordecon's sales in the Estonian market spans from the first quarter of 2008 to the fourth quarter of 2023. Analysing the trend over the years, as exhibited on the Figure 8, we observe fluctuations in sales figures, reflecting the dynamic nature of the construction industry.

In the early years of the dataset, particularly in 2008 and 2009, there appears to be a period of decline in sales, likely influenced by the global economic downturn during that time. However, from 2010 onwards, there is a noticeable recovery and growth trend, with intermittent peaks and troughs. Notably, the years 2011 to 2013 show a period of sustained growth, possibly driven by increased construction projects or favourable economic conditions.

The years following 2018 show a relatively stable trend in sales, with fluctuations reflecting typical market dynamics. However, the dataset also includes the impact of the COVID-19 pandemic, evident in the sales figures for 2020, where there appears to be a dip in sales during the second and third quarters, coinciding with the peak of the pandemic. Nonetheless, sales rebound in subsequent quarters, indicating resilience in the construction sector despite external challenges, but exhibit declining trend from second half of 2022 possibly due to worsening economic conditions.

As shown on the Figure 9, throughout the dataset, there are instances of seasonality evident in the sales data. For instance, there seems to be a pattern of higher sales figures in the second and third quarters of each year, possibly indicating increased construction activity during the warmer months in Estonia. Conversely, the first and fourth quarters exhibit lower sales figures, which may be attributed to seasonal factors such as harsh weather conditions affecting construction projects.

Overall, the fluctuation in Nordecon's sales data captures the cyclical essence of the construction sector, influenced by both external economic forces and seasonal factors. This variability accentuates the necessity of developing a robust forecasting model—one that can adapt to diverse economic climates and deliver reliable predictions across different seasonal cycles and economic conditions.

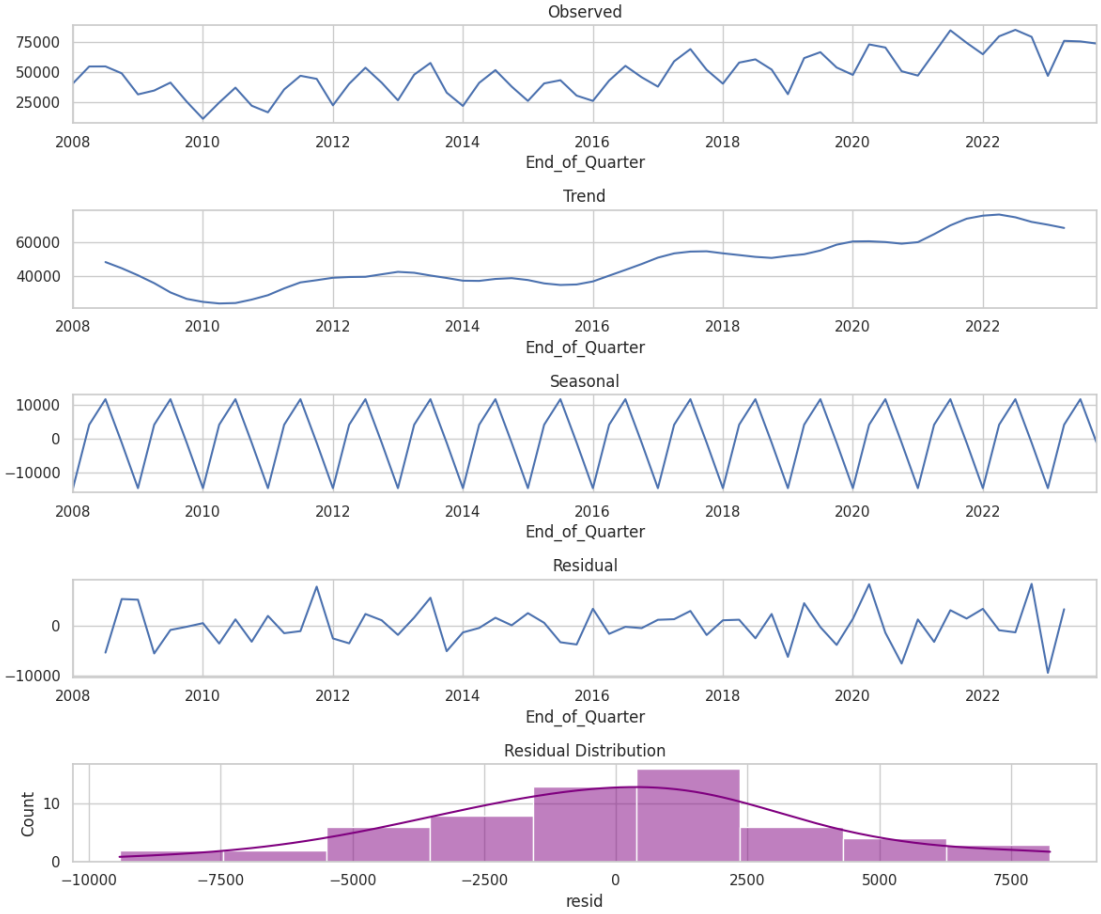


Figure 8. Seasonal decomposition of the target variable

4.2 Baseline Models and SARIMA

In this section, the performance of two naive forecasting approaches—simple naive (sNaive) and weighted naïve (wNaive) – along with the SARIMA model, is compared to generate the baseline prediction purely from time series evaluation. We used MAPE to evaluate the prediction accuracy of these models.

We validated the performance of all three models on eight folds of expanding window validation data (four quarters each) and tested on eight folds of expanding window test data (four quarters each).

The weighted naive model underwent weight optimisation utilising the grid search. This process involved leveraging training data from the same quarters of the previous three years to validate the model on the validation split. We determined the resulting best weights, subsequently applied to the test split, as follows: 0.833 for the quarter one year

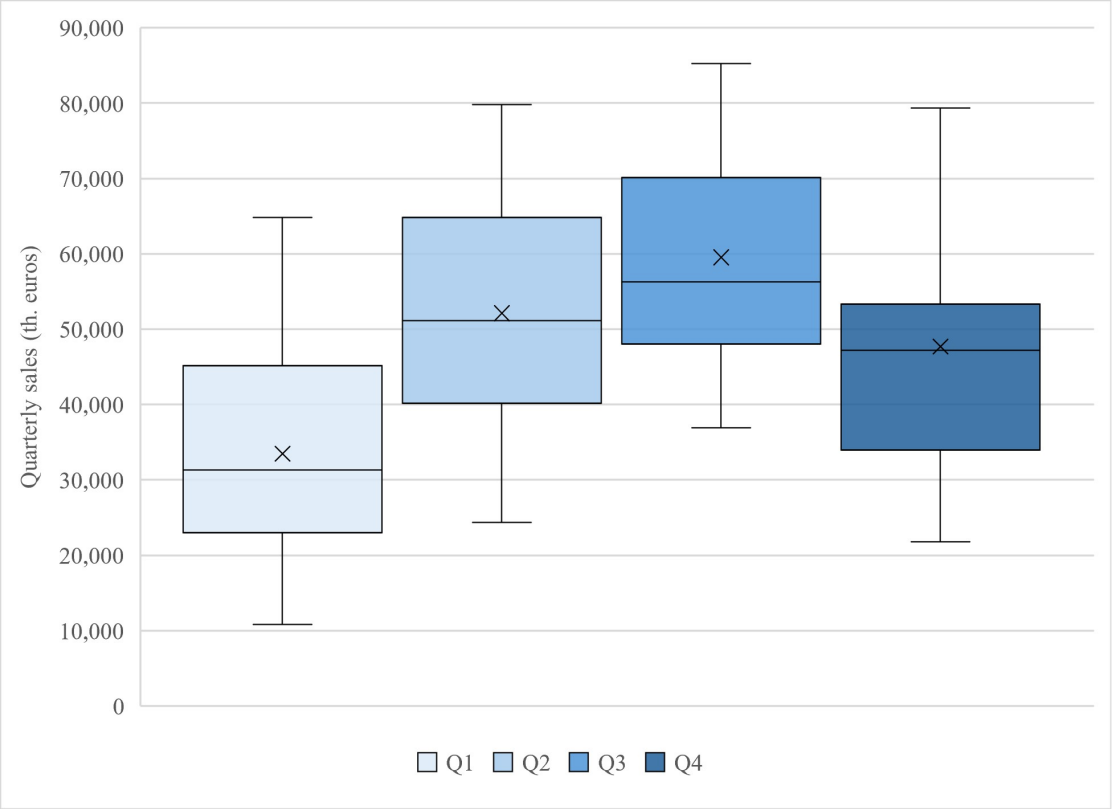


Figure 9. Distribution of quarterly values of the target variable

back, 0.083 for the quarter two years back, and 0.083 for the quarter three years back.

This suggests that the sales performance from the immediate past year holds the most significant influence on the current quarter's prediction, making it rather similar to the performance of sNaive model.

We found the SARIMA model best parameters (p, d, q, P, D, Q, S) also using grid search and these resulted in (4, 1, 5, 3, 1, 2, 8). The parameters represent the seasonal and non-seasonal components of the time series, which are crucial for capturing the more complex patterns observed in the sales data.

The non-seasonal components of the SARIMA model are represented by the parameters (p, d, q), where:

- p (autoregressive term) is 4, indicating that the current value of the sales data is linearly dependent on the previous four time points,
- d (differencing term) is 1, suggesting that the series required first-order differencing to achieve stationarity,

- q (moving average term) is 5, implying that the current value of the sales data is influenced by the previous five forecast errors.

The seasonal components of the SARIMA model are represented by the parameters (P, D, Q, S), where:

- P (seasonal autoregressive term) is 3, indicating a seasonal pattern that repeats every 3 time points,
- D (seasonal differencing term) is 1, suggesting that seasonal differencing was applied to achieve stationarity,
- Q (seasonal moving average term) is 2, implying a seasonal pattern in the forecast errors that repeats every 2 time points,
- S (seasonal period) is 8, indicating that the seasonal pattern repeats every 8 time points, which aligns with the quarterly nature of the sales data.

Overall, the SARIMA model with these parameters appeared to capture both the short-term fluctuations and the seasonal patterns present in the company's sales data.

As presented in Table 2, based on the analysis of Nordecon's sales data, the baseline models, including simple Naive (sNaive) and weighted Naive (wNaive), achieved relatively high MAPE values on the test set. The sNaive model produced a MAPE of 17.21% with a standard deviation of 3.72%, while the wNaive model yielded a slightly lower MAPE of 17.11% with a standard deviation of 4.12%. These results indicate that both baseline models struggled to accurately forecast sales, likely due to their simplistic nature and inability to capture the underlying patterns and seasonality present in the data.

Table 2. Test MAPE results of baseline and SARIMA models

Model	**Test MAPE**	**Standard deviation**
sNaive	17.21%	3.72%
wNaive	17.11%	4.12%
SARIMA	11.51%	5.80%

In contrast, the SARIMA model outperformed the baseline models on MAPE, achieving a lower value of 11.51% on the test set. However, exhibiting a higher standard deviation of 5.8%, we cannot conclude that SARIMA model demonstrated superior predictive accuracy compared to the baseline approaches. As illustrated on the Figure 10, we can observe that error bars based on one standard deviation from baseline and SARIMA models overlap. Hence, superiority of SARIMA is not statistically significant and baseline models are difficult to beat with SARIMA modelling only which had lower bias, but still higher variance.

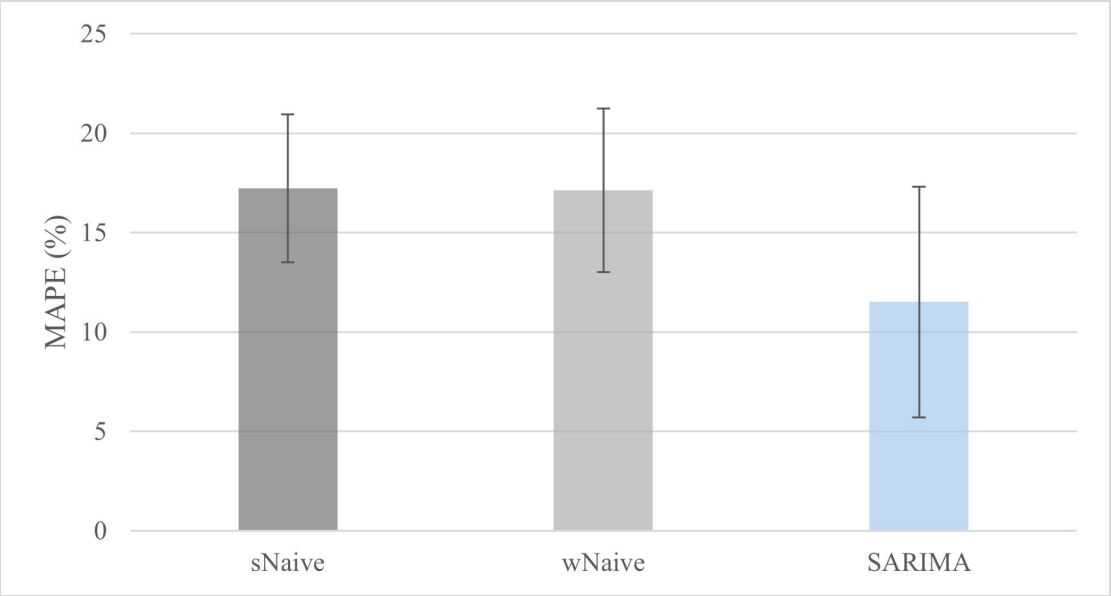


Figure 10. Test MAPE's of baseline and SARIMA models

This underscores the need for more advanced forecasting techniques in capturing the complex dynamics of Nordecon's sales in the Estonian market. Leveraging models like SARIMA, which can incorporate seasonality and temporal dependencies, is crucial first step for obtain more reliable sales forecasts, but there is a need for adopting more sophisticated regression models to improve forecasting accuracy and facilitate informed decision-making in managing operations.

4.3 Optimisation and Performance of Regression Models

This subsection presents the performance metrics of the five regression algorithms – Ridge, ElasticNet, XGBoost, Random Forest, and K-Nearest Neighbors (KNN) – applied to the prepared datasets. In this section we discuss results of first four of them – dummy, AR_dummy, FE_dummy and FE_AR_dummy – whereas the performance of regressors as Oracle model is evaluated separately. MAPE is used as the evaluation metric to assess the accuracy of the models.

Important part of the pipeline, however, is hyperparameter optimisation of regressors, which results are also presented subsequently.

4.3.1 Hyperparameter Optimisation

As a part of the pipeline, we carried out hyperparameter optimisation as described in section 3 so that models would not be reliant on their respective default hyperparameters. It was done separately of each dataset, and it covered all five regressors.

The optimisation results across various regression models reveal substantial improvements in predictive accuracy following hyperparameter optimisation as presented in Table 3. Test MAPE referred to in this table is the average value of respective regressor MAPEs over datasets 'dummy', 'AR_dummy', 'FE_dummy' and 'FE_AR_dummy'.

XGBoost exhibited the most notable enhancement, with its MAPE decreasing from 18.7% to 8.56%, representing a significant improvement of 54%. Similarly, Elastic Net experienced a considerable improvement, reducing its MAPE from 13.6% to 9.88%, marking a 27% enhancement. Ridge regression also demonstrated a notable enhancement, with its MAPE decreasing from 9.79% to 8.89%, reflecting a 9% improvement. Although the K-Nearest Neighbors model showed a slight increase in MAPE post-optimisation (18.13% to 19.17%), the Random Forest model exhibited minimal change (19.22% to 19.06%, a 1% improvement).

Table 3. Improvement by hyperparameter optimisation of regressors

Regressor	**Test MAPE pre-optimisation**	**Test MAPE post-optimisation**	**Improvement**
XGBoost	18.70%	8.56%	54%
Ridge	9.79%	8.89%	9%
Elastic Net	13.60%	9.88%	27%
Random Forest	19.22%	19.06%	1%
K-Nearest Neighbors	18.13%	19.17%	-6%

In case of K-Nearest Neighbors model, we observed during training the consistent pattern of a perfect training score and poor validation and test scores. This indicates that the K-Nearest Neighbors model is overfitting, which implies that hyperparameter optimisation, in this case, is not effectively improving the model's generalisation to new data.

Significantly, the Wilcoxon signed-rank test, a non-parametric statistical hypothesis test used when comparing two related samples or repeated measurements on a single sample to assess whether their population mean ranks differ, confirmed the efficacy of hyperparameter optimisation. The test yielded a statistically significant improvement in the models' performance with a p-value of 0.0296, indicating that the optimisation of hyperparameters reliably enhanced the forecasting accuracy.

4.3.2 Performance of Regressors

The consolidated results presented in Table 4 elucidate the impact of different feature engineering scenarios on the performance of the various regression models.

Table 4. Test MAPE development over different datasets

Regressor	'dummy' dataset	'AR_dummy' dataset	'FE_dummy' dataset	'FE_AR_dummy' dataset
Ridge	9.75%	9.87%	8.00%	7.92%
XGBoost	8.62%	8.61%	8.61%	8.41%
Elastic Net	10.40%	10.55%	9.39%	9.19%
K-Nearest Neighbors	18.67%	18.66%	20.16%	19.19%
Random Forest	18.83%	18.35%	19.51%	19.55%

We can draw several insights regarding the performance of the regression models across different feature engineering scenarios:

- **Feature Engineering Impact:** All regressors, except K-Nearest Neighbors, show improved performance with feature engineering (FE_dummy and FE_AR_dummy), indicating the value added by incorporating relevant features beyond raw data.
- **Autoregressive Terms:** The addition of autoregressive terms (AR_dummy and FE_AR_dummy) generally provides a marginal benefit over no autoregressive terms (dummy and FE_dummy). However, this benefit is most pronounced in the Ridge regression model, which suggests that historical values play a significant role in predicting future sales for this model.
- **XGBoost Consistency:** XGBoost demonstrates remarkable consistency across all scenarios, maintaining a test MAPE within a tight range (8.41% - 8.62%). This uniform performance implies that XGBoost can handle various feature combinations without drastic changes in effectiveness.
- **Optimal Scenario:** For Ridge and Elastic Net models, the FE_AR_dummy scenario (incorporating both feature engineering and autoregressive terms) provides the best performance, emphasising the importance of both past values and engineered features in capturing sales trends.
- **K-Nearest Neighbors and Random Forest:** Both models show the highest MAPE across all scenarios, suggesting a lesser ability to capture the sales dynamics compared to other models. Moreover, the FE_dummy scenario seems particularly detrimental to the K-Nearest Neighbors model.

- **Overall Dataset Performance:** The lowest dataset-wide MAPE is achieved with the FE_AR_dummy scenario, confirming the collective benefit of feature engineering and autoregressive terms across models.
- **Top Performers:** When considering the three best-performing models (Ridge, XGBoost, Elastic Net), the FE_AR_dummy scenario again proves to be the most effective, achieving the lowest combined MAPE of 8.507%.

The data indicate a clear trend: incorporating both feature engineering and autoregressive terms generally leads to improved forecasting accuracy, although the degree of improvement varies by model. K-Nearest Neighbors and Random Forest do not follow this trend, which could be indicative of model-specific limitations or sensitivities. The results also suggest that while sophisticated models like XGBoost are robust to feature variations, traditional models like Ridge and Elastic Net benefit more distinctly from the careful selection of features and the inclusion of past sales data in their predictive processes.

During the experiments we observed that the runtimes of the regression models varied significantly across different datasets and algorithms. Among the regressors, K-Nearest Neighbors consistently demonstrated the shortest runtimes, with most executions completing in less than a second. In contrast, Random Forest and XGBoost exhibited the longest runtimes, particularly when applied to datasets with extensive feature engineering. These models required around an hour or more to process the data and optimise hyperparameters over 300 iterations. Ridge regression and Elastic Net generally fell in between, with moderate runtimes ranging around few minutes depending on the complexity of the dataset. It's crucial to consider runtime constraints when selecting a regression model for real-time prediction tasks or applications requiring rapid model deployment.

4.4 Comparison of Model Predictions to Actual Sales

An examination of the distribution of test set predictions made by the Ridge and XGBoost models against actual sales values offers insight into the models' predictive performance. Kernel Density Estimation (KDE) plots provide a visual representation of the prediction densities and how closely they align with the actual sales figures.

In Figure 11, the Ridge model's test set predictions are shown in blue, while the actual sales values are represented in orange. The Ridge model exhibits a distribution that closely follows the actual sales, with a central peak that closely aligns with the peak of the actual sales distribution, suggesting a better calibrated model with predictions that are consistent with the true values.

Conversely, as depicted in Figure 12, while the XGBoost model captures the general trend of the sales data, its test set predictions are more widely dispersed around the actual values, and the central peak is notably shifted away from the actual sales peak. This

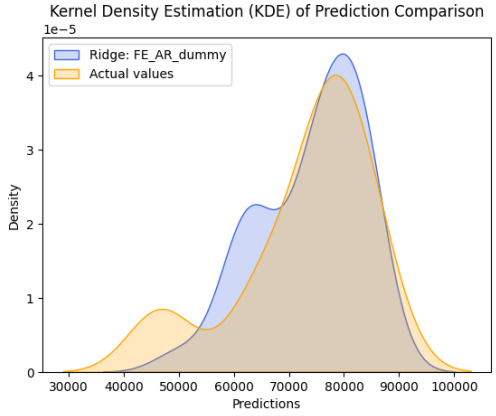


Figure 11. Ridge Model KDE

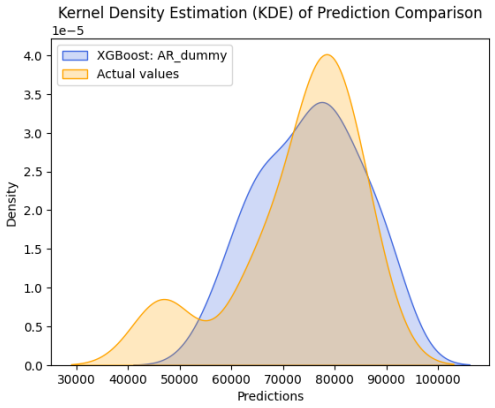


Figure 12. XGBoost Model KDE

indicates a potential bias in the XGBoost model's predictions, leading to a systematic underestimation.

Despite the absence of statistically significant differences in test set MAPE values between the two models, the KDE plots suggest that the Ridge model may offer more accurate and consistent forecasts for this particular dataset, which is an important consideration for businesses seeking reliable sales predictions in the construction industry.

4.5 MAPE over Test Folds

Based on the MAPE behaviour over different test folds of the best regressors for each dataset, we observed that the performance fluctuates in response to changes in the underlying sales data, whereas by default, one might have expected the error to remain stable over these fluctuations. For instance, XGBoost with the AR_dummy dataset maintains relatively low MAPE during periods of sales increase (e.g., 2022-Q2), but the MAPE tends to increase as sales begin to decline (e.g., 2023-Q1). This fluctuation is indicative of the challenges inherent in predictive modeling with non-stationary data, where the model's ability to adapt to underlying data trends directly influences accuracy. As data trends shift, the performance of static models typically varies, reflecting their limited capacity to accommodate such changes without additional adjustments or more sophisticated modeling techniques. Similarly, Ridge regression exhibits comparable trends with the FE_AR_dummy dataset, showing lower MAPE values during sales growth and higher values during downturns. This observation underscores the need for future research into models that can dynamically adjust to data drift[30] and maintain consistent performance across varying economic conditions.

[30]Data drift is a change in the statistical properties and characteristics of the input data (`https://www.evidentlyai.com/ml-in-production/data-drift`).

When comparing the MAPE behaviour between XGBoost (AR_dummy) and Ridge (FE_AR_dummy), both models generally follow the same trend, with fluctuations corresponding to changes in sales volume. However, there may be slight differences in the magnitude of MAPE values between the two models due to variations in their underlying algorithms and parameter configurations.

The visual representation of these observations is displayed in Figure 13, which plots Nordecon's four-quarter rolling sales data alongside the MAPE values of XGBoost (AR_dummy) and Ridge (FE_AR_dummy) across various test folds. In this context, four-quarter rolling sales are calculated by summing the sales data for a given quarter with the sales from the three preceding quarters, providing a cumulative figure that smooths out seasonal fluctuations and reveals underlying trends. This graph helps visualise how the predictive accuracy of the models correlates with changes in sales volume over time, highlighting patterns or trends in their performance.

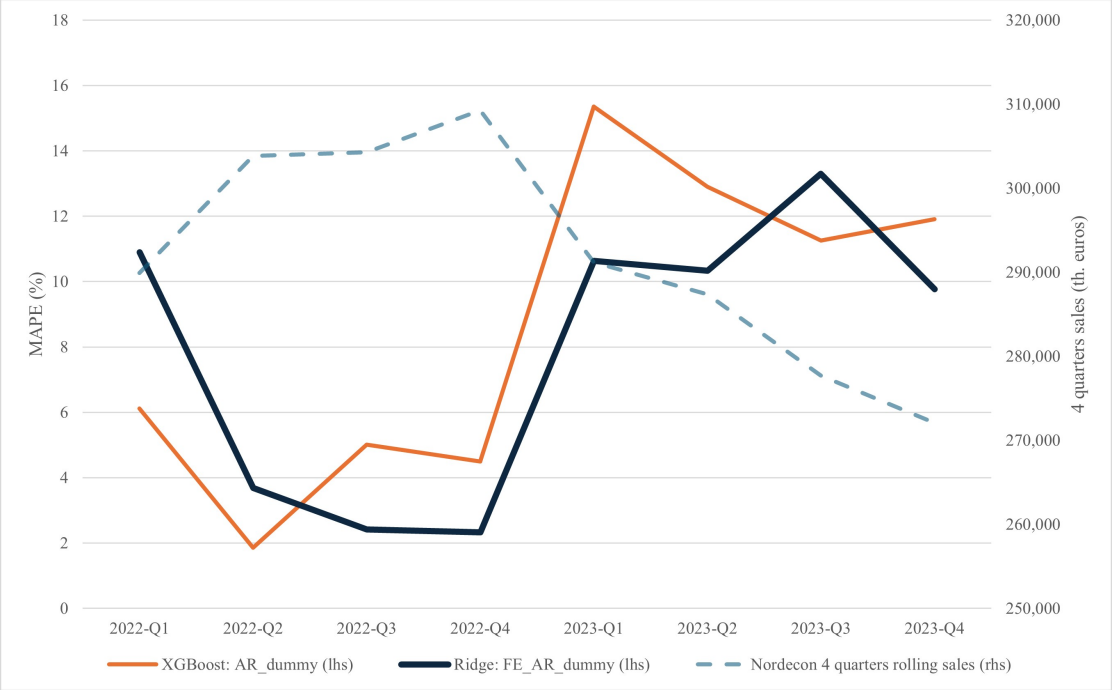


Figure 13. Sales trend and MAPE over test folds

Addressing the challenge of maintaining predictive stability amidst fluctuating sales is paramount for developing reliable forecasting models. As demonstrated by the observed fluctuations in MAPEs relative to changes in sales volumes, the predictive accuracy of models is impacted by shifts in the data, not just declines. This highlights the need for models that are robust and adaptable to any market conditions, whether facing upward or

downward economic trends. Such adaptability ensures the effectiveness of models in maintaining accuracy across all phases of economic cycles, reflecting their capability to handle data drift effectively. Acknowledging this challenge emphasises the importance of ongoing model refinement and adaptation to evolving market dynamics, ultimately enhancing the reliability and utility of forecasting efforts.

4.6 MAPE over Forecast Horizon

Looking at the forecast horizon performance shows how different regressors perform as we get farther into the future. This is illustrated on the Figure 14. While MAPEs over the forecast horizon show expected increasing trend (there is more uncertainty the further the horizon extends), it is noteworthy that the second step often demonstrates lower error compared to the first step. This indicates an improvement in predictive capability as the forecast horizon progresses, suggesting that the models are learning and adapting to the underlying patterns in the data. One possible explanation for this phenomenon is that when predicting the second step, the lag one variables are unavailable, which may indicate that these features have lower predictive power. Consequently, model accuracy could be more dependent on lag two variables and beyond, suggesting a shift in the data features that contribute most effectively to forecasting accuracy. However, it's essential to thoroughly investigate this phenomenon further to ensure that it is not due to anomalies or biases in the dataset or model evaluation process. Understanding the reasons behind this trend can lead to further enhancements in model performance and contribute to more accurate and reliable forecasts across all forecast horizons.

4.7 Oracle Prediction

We created 'Oracle' dataset specifically to assess the impact of unknown future values on predictive accuracy.

The degree of data loss at each step of the forecast horizon of four quarters in the 'FE_AR_dummy' dataset provides valuable context. For the first step, all 460 attributes are available. However, as the forecast horizon progresses, the data loss becomes more pronounced. For example, by the fourth step, only 295 attributes are available, representing a 36% reduction compared to the first step. This reduction highlights the challenge of maintaining data integrity and completeness in forecasting scenarios.

As presented in the Table 5, in the evaluation of predictive models using the 'Oracle' dataset, it's evident that filling unknown future values significantly improves predictive accuracy. Across the regression models used, the test MAPE is lower in the 'Oracle' dataset compared to the 'FE_AR_dummy' dataset: XGBoost test MAPE improved by 25%, Ridge 12%, Elastic Net 24%, K-Nearest Neighbors 6%, and Random Forest 1%.

Furthermore, the average improvement in MAPE across all regression models is noteworthy, standing at 10%. This underscores the importance of addressing unknown

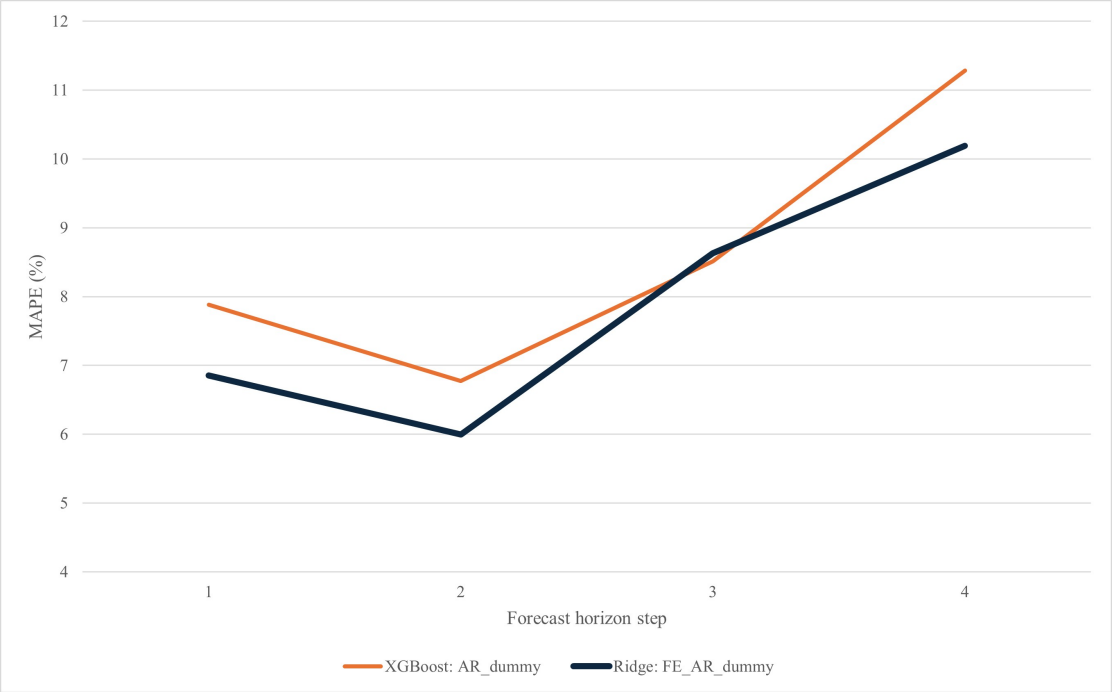


Figure 14. Regressor MAPE over forecast horizon

future values effectively to enhance the accuracy and reliability of predictive models. However, it's essential to interpret these results with caution and consider other factors that may influence model performance, such as the nature of the dataset, feature engineering techniques, and model complexity. Overall, the 'Oracle' dataset serves as a valuable tool for understanding the implications of unknown future values on forecasting accuracy and guiding strategies to mitigate its effects. One effective way to address these gaps is by forecasting the unknown future data/values themselves. The insights gained from the Oracle analysis help determine whether investing resources into forecasting unknown future values is justifiable. In this case, we observe that the best model can achieve a reduction in error of 25%, which is significant. This suggests that as a future direction, efforts to forecast unknown future values could be highly beneficial, potentially leading to considerable improvements in model accuracy.

4.8 Summary of Results

The comprehensive analysis of Nordecon's sales data using various regression models and forecasting techniques provides valuable insights into the factors influencing sales performance and the effectiveness of predictive modeling in the construction industry. Overview of best model for each dataset is presented in Table 6 and illustrated on the

Table 5. Oracle model improvement over 'FE_AR_dummy' dataset

Regressor	'FE_AR_dummy' test MAPE	'FE_AR_dummy' st dev	'Oracle' test MAPE	'Oracle' st dev	MAPE improvement
XGBoost	8.41%	4.92%	6.34%	3.30%	25%
Ridge	7.92%	4.09%	6.99%	3.04%	12%
Elastic Net	9.19%	3.28%	7.00%	3.05%	24%
KNN	19.19%	3.25%	17.95%	5.12%	6%
Random Forest	19.55%	2.91%	19.43%	2.56%	1%
Average	**12.85%**	**3.69%**	**11.54%**	**3.41%**	**10%**

Figure 15.

Table 6. Results of the best models per dataset

Model	Dataset	Test MAPE	Test st dev
XGBoost	Oracle	6.34%	3.30%
Ridge	FE_AR_dummy	7.92%	4.09%
Ridge	FE_dummy	8.00%	4.09%
XGBoost	AR_dummy	8.61%	4.52%
XGBoost	dummy	8.62%	4.52%
SARIMA	historical sales	11.51%	5.80%
wNaive	historical sales	17.11%	4.12%

Firstly, the baseline models, including simple naive (sNaive) and weighted naive (wNaive), along with the SARIMA model, provided a foundational understanding of sales trends and seasonality. However, their relatively high MAPE values indicated limitations in accurately forecasting sales, especially when confronted with the dynamic and cyclical nature of the construction market. The SARIMA model demonstrated better average predictive accuracy, leveraging both non-seasonal and seasonal components to capture complex patterns in the sales data effectively. However, exhibiting a higher standard deviation, we were not able to conclude that SARIMA model demonstrates better compared to the naive baselines statistically significant predictive accuracy.

Furthermore, the application of regression models, including Ridge, Elastic Net, XGBoost, Random Forest, and K-Nearest Neighbors (KNN), demonstrated substantial improvements in predictive accuracy following hyperparameter optimisation. While all models showed varying degrees of performance across different datasets and feature engineering scenarios, Ridge regression consistently emerged as one of the best-performing regressors, especially when applied to the FE_AR_dummy dataset, which is the largest dataset used in this study. This dataset's complexity and feature density potentially allow Ridge, with its inherent regularisation properties, to effectively select important features, thereby enhancing its predictive accuracy and stability. This advantage, combined with

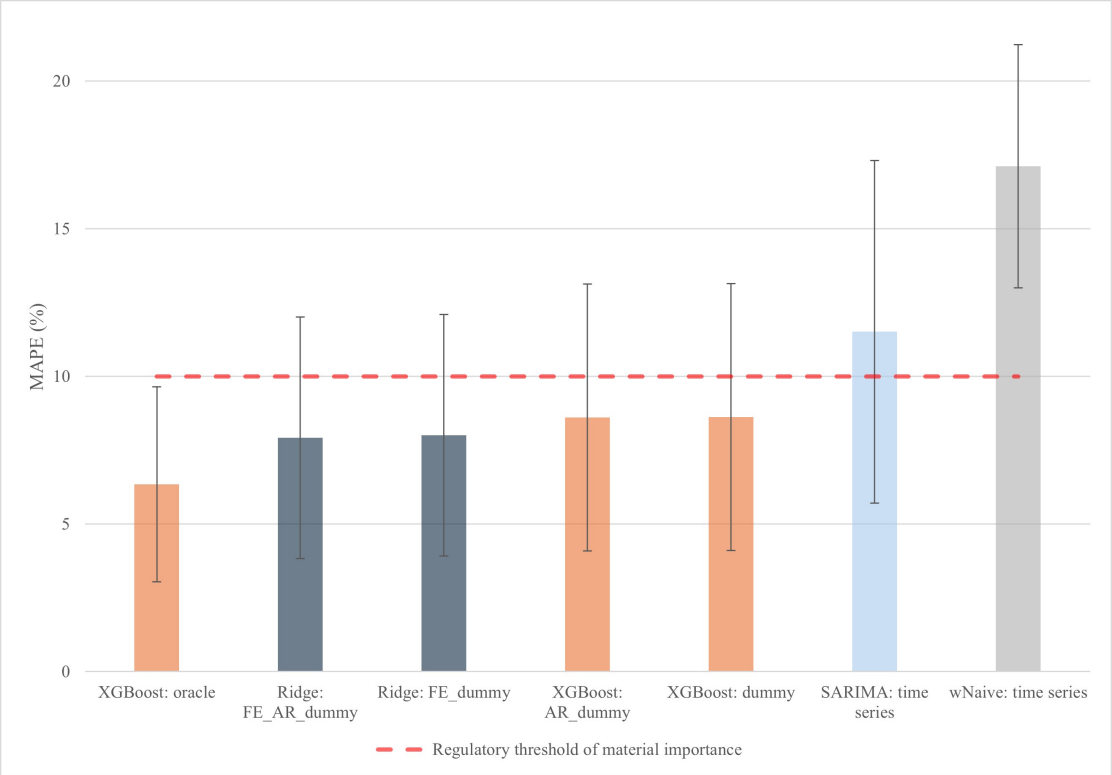


Figure 15. Best model per dataset

Ridge's significantly lower runtime compared to more computationally intensive models like XGBoost, positions it as a particularly effective tool for dealing with the large-scale, feature-rich data typically found in economic analysis. This analysis distinctly shows that integrating lag features and domain-specific indicators via feature engineering can significantly tighten forecast accuracy and stabilise model predictions, directly addressing the sequential nature and lagged correlations found in sales data.

The analysis of MAPE behaviour across different test folds provided insights into the varying performance of regression models like XGBoost and Ridge in response to changes in sales volume. This analysis showed that these models' performance fluctuated significantly with market momentum and economic cycles, with periods of rapid market expansion or contraction impacting predictive accuracy. This highlights the challenge of ensuring that models remain stable and reliable in different market conditions, as they currently do not inherently adapt to shifts in market trends and may exhibit deteriorating performance under changing economic scenarios.

Moreover, the creation and evaluation of the 'Oracle' dataset provided critical insight into impact of forecasting future values of macro-economic indicators. In scenario like

this where unknown future values are addressed, there was a notable enhancement in the models' forecasting accuracy, emphasising the critical role of data quality in predictive analytics. The 'Oracle' scenario represents an idealised benchmark, demonstrating the potential upper limits of forecasting sales performance when limitation of unknown future values for the future is alleviated, thus providing a best-case reference for what could be achieved with complete information.

In our study, the 10% threshold established by Nasdaq Tallinn stock exchange serves as a crucial benchmark for assessing the materiality of forecasting errors in financial results. According to the regulation, if an issuer's forecasted financial outcomes deviate by more than this threshold from actual results, an immediate revision and explanation are mandated to maintain transparency and regulatory compliance.

The analysis presented in this study critically evaluates the performance of various forecasting models against this regulatory benchmark. As illustrated in the Figure 15, the models employing advanced ML techniques, specifically Ridge: FE_AR_dummy, on average maintain MAPE values well below this 10% threshold. This performance not only demonstrates their high accuracy and reliability but also indicates their strong alignment with the regulatory standards that govern financial forecasting in public markets.

Conversely, traditional models like SARIMA and the weighted naive approach occasionally exhibit MAPE values that approach or exceed this threshold, highlighting potential risks in their use for financial forecasting under the regulatory framework. Such instances underscore the necessity for models that can reliably predict within the bounds of material significance, thereby ensuring that the forecasts remain within acceptable limits as defined by regulatory authorities.

This stringent comparison to the 10% threshold underscores the importance of selecting and refining predictive models that not only provide accurate forecasts but also conform to the regulatory expectations of financial reporting. For Nordecon AS, employing models that consistently perform within this threshold ensures that they adhere to market regulations and uphold standards of fiscal responsibility and transparency in their financial predictions.

5 Conclusion

This thesis has successfully developed and validated a practical pipeline for forecasting sales in the construction industry, leveraging both macroeconomic indicators and ML techniques. The study's findings substantiate that integrating diverse data streams, particularly through the application of advanced regression models like Ridge regression enhanced with feature engineering, significantly improves forecast accuracy. Specifically, the most effective models demonstrated a MAPE well below the 10% regulatory threshold of material importance set by the Nasdaq Tallinn stock exchange. This performance not only highlights the models' compliance with industry standards but also their reliability in dynamic economic environments, where traditional non-statistical methods may struggle to provide accurate forecasts. This comparison with the regulatory threshold underscores the practical and regulatory relevance of adopting advanced statistical methods in financial forecasting within the construction sector.

The research has demonstrated that ML models, particularly those incorporating detailed feature engineering, provide not only higher accuracy but also greater stability in forecasting. This is critical for strategic planning in industries like construction where economic sensitivity is high. The models developed herein consistently delivered more stable predictions with reduced variability, which means it could be employed to enhance decision-making processes, resource allocation, and market response strategies for Nordecon. These improvements are pivotal for fostering sustainable growth and enhancing profitability in the fluctuating sphere of the construction industry.

Furthermore, the creation of an 'Oracle' dataset as part of the methodology specifically addressed the critical issue of unknown future values in time series forecasting. By simulating a more complete data scenario, this dataset provided insights into the potential maximum performance of our models under ideal conditions. These insights reveal that forecasting unknown future values of macroeconomic indicators could lead to significant improvements in model accuracy, potentially enhancing results by up to 25%. This finding underscores the importance of developing methods to accurately forecast unknown future values as a promising direction for future research, aiming to push the envelope on forecasting accuracy expectations in the field.

For practical deployment, the models and methodologies developed in this study could be integrated into decision-making processes through interactive dashboards or embedded systems that provide ongoing insights and updates. This would enable real-time strategic adjustments in response to market changes, thereby amplifying Nordecon's adaptive capacities.

Looking forward, the approach outlined in this thesis lays a robust foundation for further research. Future studies could explore the incorporation of additional predictive variables and the application of even more advanced ML techniques, such as neural networks or ensemble methods, to enhance the predictive accuracy and robustness further. Additionally, experimenting with real-time data ingestion and adaptive learning models

could provide a pathway to even more responsive and dynamic forecasting systems. This ongoing evolution in predictive analytics will continue to revolutionise decision-making processes in the construction industry and beyond, driving towards more data-driven, precise, and strategic business operations.

Acknowledgements

I would like to extend my heartfelt gratitude to my supervisor, Novin Shahroudi, whose methodological rigour, ever-present support, and detailed feedback were instrumental in guiding and enriching this research.

My deepest appreciation also goes to my wife and family, whose inspiration and patience have been my cornerstone throughout this journey. I am profoundly grateful for the encouragement and understanding they have shown me.

Additionally, I owe a special thanks to my employer Nordecon AS for facilitating this opportunity and supporting me in balancing my professional and academic commitments. Your support has been invaluable.

And finally, I am very thankful to the academic staff and my peers at the University of Tartu for their pivotal role in developing and managing a curriculum that equips graduates to bring valuable insights and innovations to a wide range of sectors, not limited to information technology. This program has not only shaped my academic endeavors but also demonstrates the universal benefits of integrating advanced data science practices across various traditional industries.

Together, these individuals and groups have played a pivotal role in my studies, and I am immensely grateful for their contributions to my academic and personal growth.

References

[AAH98] Paul Bowen Akintola Akintoye and Cliff Hardcastle. Macro-economic leading indicators of construction contract prices. *Construction Management and Economics*, 16(2):159–175, 1998.

[Arm01] J. Scott Armstrong, editor. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. International Series in Operations Research & Management Science. Springer-Verlag US, New York, NY, 1 edition, 2001. XII, 850 pages.

[Auf21] Ben Auffarth. *Machine Learning for Time-Series with Python: Forecast, Predict, and Detect Anomalies with State-Of-the-art Machine Learning Methods*. Packt Publishing, 2021.

[Ayd24] Serkan Aydinli. Impact of unexpected conditions on construction cost forecasting performance: evidence from europe. *Construction Management and Economics*, pages 1–15, 2024.

[BAL12] Seyed Mohsen Shahandashti Baabak Ashuri and Jian Lu. Empirical tests for identifying leading indicators of enr construction cost index. *Construction Management and Economics*, 30(11):917–927, 2012.

[Bar09] Richard Barras. *Building cycles: growth and instability*. John Wiley & Sons, 2009.

[Bre01] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[Bro17] Jason Brownlee. *Introduction to Time Series Forecasting with Python*. Jason Brownlee, 1.12 edition, 2017.

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[CTM20] Vitor Cerqueira, Luis Torgo, and Igor Mozetič. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11):1997–2028, 2020.

[Dav20] Davis David. Hyperopt: The alternative hyperparameter optimization technique you need to know. `https://www.analyticsvidhya.com/blog/2020/09/alternative-hyperparameter-optimization-technique-you-need-to-know-hyperop` 2020. Accessed: 23-Feb-2024.

[Ees] Eesti Pank. The estonian economy and monetary policy, 1/2024. Quarterly Review by Bank of Estonia. Available online: https://www.eestipank.ee/en/publications/estonian-economy-and-monetary-policy.

[FGLN09] Robert Fildes, Paul Goodwin, Michael Lawrence, and Konstantinos Nikolopoulos. Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 25(1):3–23, 2009.

[GCS14] Rinkaj Goyal, Pravin Chandra, and Yogesh Singh. Suitability of knn regression in the development of interaction based software fault prediction models. *IERI Procedia*, 6:15–21, 2014. 2013 International Conference on Future Software Engineering and Multimedia Engineering (ICFM 2013).

[HA21] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 3 edition, 2021.

[Hil00] Patricia M. Hillebrandt. *Economic Theory and the Construction Industry*. Palgrave Macmillan London, 3 edition, 2000.

[HK08] Rob J. Hyndman and Yeasmin Khandakar. Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, 27(3):1–22, 2008.

[IHCD22] Daniel W. Williams Il Hwan Chung and Myung Rok Do. For better or worse? revenue forecasting with machine learning approaches. *Public Performance & Management Review*, 45(5):1133–1154, 2022.

[Joh18] Johan Skytte Institute of Political Studies. The analysis of productivity, value added and economic impact of construction industry, 2018. Available online: https://skytte.ut.ee/sites/default/files/2022-05/ehitussektori_tootlikkuse_lisandvaartuse_ja_majandusmoju_analuus_uuendatud.pdf.

[KB82] Jeannie S. Kidwell and Lynn Harrington Brown. Ridge regression as a technique for analyzing models with multicollinearity. *Journal of Marriage and Family*, 44(2):287–299, 1982.

[LC14] Jiahan Li and Weiye Chen. Forecasting macroeconomic time series: Lasso-based approaches and their forecast combinations with dynamic factor models. *International Journal of Forecasting*, 30(4):996–1015, 2014.

[MAP19] CH. Raga Madhuri, G. Anuradha, and M. Vani Pujitha. House price prediction using regression techniques: A comparative study. In *2019 International Conference on Smart Structures and Systems (ICSSS)*, pages 1–5, 2019.

[MCM+13] Juan M. Morales, Antonio J. Conejo, Henrik Madsen, Pierre Pinson, and Marco Zugno. *Integrating Renewables in Electricity Markets: Operational Problems*. International Series in Operations Research & Management Science. Springer New York, NY, 1 edition, 2013.

[MGN20] J Manasa, Radha Gupta, and N S Narahari. Machine learning based predicting house prices using regression techniques. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pages 624–630, 2020.

[MM04] John T. Mentzer and Mark A. Moon. *Sales Forecasting Management: A Demand Management Approach*. SAGE Publications, Inc, 2 edition, 2004.

[MSA18] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, 13(3):e0194889, 2018.

[Nas22] Nasdaq Tallinn. *Requirements for Issuers*, 2022. Accessed: 2023-05-03.

[Niu20] Yiyang Niu. Walmart sales forecasting using xgboost algorithm and feature engineering. In *2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, pages 458–461, 2020.

[OW09] Rogelio Oliva and Noel Watson. Managing functional biases in organizational forecasts: A case study of consensus forecasting in supply chain planning. *Production and Operations Management*, 18(2):138–151, 2009.

[PF13] Foster Provost and Tom Fawcett. *Data Science for Business*. O'Reilly Media, Inc., 1 edition, July 2013. First Edition.

[PRB99] Stephen Pyhrr, Stephen Roulac, and Waldo Born. Real estate cycles and their strategic implications for investors and portfolio managers in the global economy. *Journal of real estate research*, 18(1):7–68, 1999.

[SAKD18] Yves R. Sagaert, El-Houssaine Aghezzaf, Nikolaos Kourentzes, and Bram Desmet. Tactical sales forecasting using a very large set of macroeconomic indicators. *European Journal of Operational Research*, 264(2):558–569, 2018.

[SELL15] Michael C. P. Sing, D. J. Edwards, Henry J. X. Liu, and P. E. D. Love. Forecasting private-sector construction works: Var model using economic indicators. *Journal of Construction Engineering and Management*, 141(11):04015037, 2015.

[STNW04] Martin Skitmore S. Thomas Ng, Sai On Cheung and Toby C.Y. Wong. An integrated regression analysis and time series model for construction tender price index forecasting. *Construction Management and Economics*, 22(5):483–493, 2004.

[Tib18] Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 12 2018.

[VAD20] Gylian Verstraete, El-Houssaine Aghezzaf, and Bram Desmet. A leading macroeconomic indicators' based framework to automatically generate tactical sales forecasts. *Computers & Industrial Engineering*, 139:106169, 2020.

[WC16] Daniel W Williams and Thad D Calabrese. The status of budget forecasting. *Journal of Public and Nonprofit Affairs*, 2(2):127–160, 2016.

[YS20] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020.

[ZC18] Alice Zheng and Amanda Casari. *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.", 2018.

[ZH05] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Appendix

I. Hyperopt Search Space

Ridge

- 'alpha': hp.uniform('alpha', 0.01, 200.0),
- 'fit_intercept': hp.choice('fit_intercept', [True, False]),
- 'solver': hp.choice('solver', ['auto']),
- 'max_iter': hp.choice('max_iter', range(100, 2000, 10)),
- 'tol': hp.uniform('tol', 1e-6, 1e-3),
- 'random_state': random_state

Elastic Net

- 'alpha': hp.uniform('alpha', 0.01, 200.0),
- 'copy_X': hp.choice('copy_X', [True, False]),
- 'l1_ratio': hp.uniform('l1_ratio', 0, 1),
- 'fit_intercept': hp.choice('fit_intercept', [True, False]),
- 'max_iter': hp.choice('max_iter', range(100, 2000, 10)),
- 'tol': hp.uniform('tol', 1e-6, 1e-3),
- 'random_state': random_state

XGBoost

- 'learning_rate': hp.uniform('learning_rate', 0.001, 0.1),
- 'reg_lambda': hp.uniform('reg_lambda', 0.01, 1),
- 'reg_alpha': hp.uniform('reg_alpha', 0.01, 1),
- 'updater': hp.choice('updater', ['shotgun', 'coord_descent']),
- 'feature_selector': hp.choice('feature_selector', ['cyclic', 'shuffle']),
- 'booster': hp.choice('booster', ['gblinear']),

- 'random_state': random_state

Random Forest

- 'max_depth': hp.choice('max_depth', range(1, 40)),
- 'min_samples_split': hp.uniform('min_samples_split', 0.0, 1.0),
- 'max_leaf_nodes': hp.choice('max_leaf_nodes', range(2, 50)),
- 'n_estimators': hp.choice('n_estimators', range(50, 250, 1)),
- 'random_state': random_state

K-Nearest Neighbours

- 'n_neighbors': hp.choice('n_neighbors', range(2, 20, 1)),
- 'p': hp.choice('p', [1, 5]),
- 'weights': hp.choice('weights', ['uniform', 'distance']),
- 'algorithm': hp.choice('algorithm', ['auto', 'ball_tree', 'kd_tree', 'brute']),
- 'leaf_size': hp.choice('leaf_size', range(10, 40, 1)),
- 'metric': hp.choice('metric', ['minkowski', 'euclidean', 'manhattan'])

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Andri Hõbemägi**,
*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Sales Forecasting based on Economic Indicators for a Construction Company,
   *(*title of thesis)

   supervised by Novin Shahroudi.
   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Andri Hõbemägi
15/05/2024