Predicting the Future: Econometrics vs. Machine Learning in Macroeconomic Forecasting

Mads Tomra Nicolaisen — 20194894

June 3rd, 2024

AALBORG UNIVERSITY

STUDENT REPORT

Aalborg University Business School
4th Semester, cand.oecon

Supervisor: Hamid Bekamiri

Number of characters: $88,515 \approx 37$ pages.
ECTS points: 30 points

Abstract

The ability to forecast economic variables plays a crucial role in informing policy decision, making the precision of said forecast a significant concern. With the recent rise in Artificial Intelligence (AI) solutions in commercial products, this paper investigates how such solutions could improve the accuracy of economic forecasting, compared to classical econometric modelling.

This paper investigates how Machine Learning (ML) and AI techniques can improve the accuracy of economic forecasting compared to classical econometric models, using the specific case of forecasting Danish Gross Domestic Product (GDP) in the period from 1990 to 2024. The analysis uses ARIMA, Gradient Boosting, Long Short-Term Memory and Large Language Models to analyse which type is best at forecasting GDP.

The paper finds that the ARIMA models provide the most accurate forecasts when forecasting one-period ahead. On the contrary, the ML models perform best when forecasting 16 periods ahead. The LLMs are generally worse than the other models, but in some scenarios, they did provide the most accurate forecast of all the models.

The paper argues that statistical models like a ARIMA model are tough to beat when trying to forecast the near-future, as the model's strong statistical background ensures mathematical precision. Oppositely, the paper argues that ML and AI models can be able to provide further insights when forecasting longer into the future, as these models are capable of modelling more complex relationships than the standard econometric models. Though, this insight might come at a cost of explainability of how the forecast is computed.

The code repository used to develop the models for this paper is located at https://github.com/Madstn77/MLForecasting

Contents

List of Figures

List of Tables

1 Introduction

The recent rise of Artificial Intelligence (AI) has already begun having an impact on the surrounding society. From hospitals using AI to improve accuracy of X-Rays scans (Christensen et al., 2024), to optimizing the energy use in non-residential buildings (Pedersen, 2021), AI is actively being implemented in regular use cases, improving the service of those using these applications.

From an econometrician's perspective, it is also relevant to wonder whether this technology can be used in the field of econometrics. Specifically, the rise of AI may introduce a paradigm shift into the way econometricians approach forecasting.

Until the past few years, economic forecasts have been built upon econometrical models, using statistical concepts to define those models. This has meant that the models, and any forecast produced of such models, have a solid statistical reasoning. However, these models often provide limited insights beyond basic trend extrapolation, resembling a simple 'random walk' approach.

The question is now whether Machine Learning (ML), or related AI models, will be able to out-predict these classical models? And if so — will these predictions be so 'good' that the econometricians — and the greater society — will accept that this accuracy comes at a cost of less explainability?

This paper aims to evaluate the relevance of AI as a tool of economic forecasting.

Specifically, the paper will:

1. **Describe the evolution of methods used for forecasting time series:** This involves outlining the recent historical development of economic forecasting techniques, from traditional econometric methods to the emergence of advanced AI methodologies.

2. **Analyse forecasting performance:** The study will assess the forecasting accuracy and reliability of both classical econometric models and ML/LLM-based approaches. The focus will be on forecasting Danish Gross Domestic Product (GDP) as a case study.

3. **Discuss advantages and limitations:** A comprehensive discussion will be conducted to identify the respective strengths and weaknesses of each method. This discussion will shed light on the potential benefits and challenges associated with adopting ML/AI into economic forecasting.

2 Literature Review

In the coming chapter, previous literature in the field of time series forecasting will be presented in a structured way, such that each of the four approaches, that will be used in the analysis, first are presented with previous literature for each approach.

2.1 Classical Econometric Forecasting

In the majority of the time that time series forecasting has existed, classical econometric methodology has been used to produce these said forecasts. The modern discipline of finding patterns in past observations, to predict future values, is widely regarded as being started alongside the rise of the modern computer.

The first academic contribution to the field is regarded to be G. E. P. Box and Jenkins (1979), which introduces the concept of an ARIMA model to retrieve the 'optimal forecast' from a time series.

Specifically, Box and Jenkins are attributed with the three-step method proposed in their paper:

1. Model Identification: Analysing the time series to find whether the model is stationary, experiences seasonality, and/or autocorrelation.
2. Model Estimation: Finding the coefficients that best fit the observed data. G. E. P. Box and Jenkins (1979) proposes either to use a likelihood function [1] or Bayes Theorem [2].
3. Model Diagnostic Checking: The primary concern during diagnostic checking is to assess whether the model is overfitted. Overfitting occurs when a model captures too much noise or random fluctuations in the historical data, leading to a poor generalization of the underlying patterns. This means the model fits the training data very well but performs poorly on new, unseen data.

The general methodology is still relevant for modern ARIMA models, as well as other econometrical models. Today, ARIMA modelling is typically implemented as an SARIMA, ARIMAX, or a combination of the two (SARIMAX).

Here, the 'S' in SARIMA denotes a model able to capture seasonal aspects of a time series and the 'X' in ARIMAX denotes a model being able to incorporate eXogenous variables. Both terms are further explained in subsection 4.1.

Maccarrone et al. (2021) has in their article tried to measure the predictive performance of an ARX model and a SARIMAX model, compared to a newer ML model

[1] By maximizing such likelihood function, the model will find the optimal set of model parameters, which will make the observed data point the most probable.

[2] This theorem integrates prior beliefs with the observed data, creating a posterior distribution that Bayesian Method will use to achieve robust estimators.

— more specifically a K-Nearest Neighbour model. They argue that it is important from an econometric standpoint to test these new and compute-intensive methods and see how they forecast compared to the classic econometric approaches.

A similar argument is proposed by Biau and D'Elia (2012, p. 3):

"Since hard data (e.g. GDP) are published with a considerable delay, policy decisions have to rely on more timely information: for example, business tendency survey data, which — due to their early release — are widely used as potential indicators to track economic activity."

Biau and D'Elia (2012) further explains that these survey data typically are scattered across multiple sources and are therefore difficult to incorporate in classical econometric models. Their paper therefore tries to implement a Random Forest (RF) model with these survey data as explanatory variables. To compare the performance, Biau and D'Elia (2012) created a simple AR model.

Both Maccarrone et al. (2021) and Biau and D'Elia (2012) use evaluation metrics, which in general compute the distance between a forecast and the realized values[3], to compare the accuracy of their models' forecasts. At the same time, they graph the forecast such that more intricate behaviours of the models can be analysed.

Maccarrone et al. (2021, p. 2) writes:

"The idea is that the decision-maker should adopt the machine learning as a powerful instrument and should employ it with awareness without regarding it as a "black-box.""

The quote explains that the job of the econometricians is not only to illustrate the performance of comparable models, but also have a communicative job of explaining the abilities and weaknesses of this new methodology compared to econometric models.

Both Maccarrone et al. (2021) and Biau and D'Elia (2012) introduce a line-up of explanatory variables in their papers, used to forecast the endogenous variable. Both papers test the models on different setups where the explanatory variables are combined in different ways. By using different setups of exogenous variables in the testing, they minimize the risk of deeming a model type bad at forecasting, in the case that the underlying variables simply are bad predictors.

It is still important to note that the variables picked in the articles are conventional variables used in econometric forecasting and are picked based on economic rationale.

[3] Further specified in subsection 4.5

In their conclusion, Maccarrone et al. (2021) find that the K-Nearest Neighbour model outperforms their ARX and SARIMAX models. On the other hand, Biau and D'Elia (2012) finds that the RF model does not predict GDP that well — instead he uses the RF model to select relevant variables, and uses these variables in a simple Linear Regression (LR) model. This model, in turn, performs very well.

2.2 Simple Machine Learning Forecasting

In the past two decades, econometricians have been entering the academic field of machine learning. As written above, the argument for entering this field is to see if ML models could improve econometric forecasts.

Biau and D'Elia (2012) applies an RF model to a dataset that they argue would not be possible to use in a classical econometric model because the data contains unstructured data.

In this paper, they try to predict the Financial Crisis in 2008 on European GDP using survey data from the European Union. They find that the model was not good at predicting the crisis, as no negative GDP growth numbers had not been present until the crisis entered.

Though, they implement another simple LR model, which uses the selected set of variables that the RF model had found to be explaining GDP. This LR model, with variables selected from the RF model, performs well and is comparable to the economic outlook of the European Union from the European Central Bank.

This could suggest that simple implementations of ML models might lack predictive abilities, but can highlight relevant variables from a larger dataset.

On the other hand, Yoon (2021) implements both an RF model and a Gradient Boosting (GB) model on Japanese GDP.

The paper uses a wide-ranging list of macroeconomic variables to predict the GDP and finds that both of the models outperform the economic outlooks of the Bank Of Japan (BOJ) and the International Monetary Fund (IMF). Though the paper does not implement a statistical model, so Yoon (2021) does not give an answer to whether such a statistical model might outperform the economic outlooks as well.

2.3 Advanced Machine Learning Forecasting

In an attempt to improve the forecast accuracy, Sa'adah and Wibowo (2020) applies ML models that in theory should overcome some issues of the simple ML models described above.

Sa'adah and Wibowo (2020) implements both a Long-Short Term Memory model (LSTM) and a secondary Recurring Neural Network model (RNN) in terms of the SimpleRNN from TensorFlow. The researchers chose to investigate these models because of their ability to process sequential data. This should mean that these models should improve in their sequential understanding of a time series, contrary to a simple RF model, that does not consider time in its predictions.

Sa'adah and Wibowo (2020) finds that both the LSTM and RNN model provide an accurate forecast of the Indonesian GDP. Importantly, the models seem to be able to predict the Financial Crisis of 2008, but are less accurate when trying to predict the beginning of the Covid-19 pandemic in 2020.

Hopp (2022) similarly, tries to implement a LSTM model to test the LSTM model's prediction performance against a Dynamic Factor Model (DFM) which represents a statistical model in this paper. Hopp (2022) argues that DFM models are the current standard when nowcasting[4] as it can compensate for data issues, e.g. missing observations. This is relevant when introducing a wide range of non-structured data to a model.

Hopp (2022) found that the LSTM, on average, outperformed the DFM in terms of nowcasting Global Merchandise Trade in both volume and monetary value, as well as Global Services Trade. As Hopp (2022) mentions, this result is positive, not only because of the LSTM outperformed the DFM model. The LSTM model can also handle more model-features before experiencing computer bottlenecking. This means that with the same computing resources, the LSTM can include more features than DFM before experiencing capacity issues.

2.4 Large Language Model Forecasting

In the current rise of AI, most of the current advancements have happened in the realm of Large Language Models (LLM). With the rise in LLMs like OpenAI's GPT models and Google's various similar models, more researchers have speculated in whether these types of transformer models could be implemented for time series analysis. For now, the issue has been, that not enough time series data is available to train a time series pre-trained model. Zhou et al. (2023) finds that the largest time series dataset available for training is 10 GB, which is much less than the datasets the latest LLMs are trained upon.

Zhou et al. (2023) challenges this preconception, by wondering if a model trained on text might be able to perform accurate time series analysis. By using the rather outdated GPT-2 model from OpenAI, Zhou et al. (2023) show that this model can outcompete similar ML models trained on pure time series.

[4] Nowcasting refers to the idea of estimating macroeconomic variables that are normally published with a significant lag, by using more timely indicators.

Zhou et al. (2023) tests the LLM's ability to compute relevant time series analysis tasks. This spans tasks as anomaly detection, and short and long term forecasting, based on relevant datasets generally used to test model performance.

Zhou et al. (2023) concludes their LLM models are on par, or are better than the state-of-the-art models in most of the tested tasks. The tasks that the LLM outperforms in include tasks such as long/short-term forecasts, anomaly detection and imputation. They acknowledge that their zero-shot[5] approach is still lacking, meaning the current models still needs to be tuned upon specific examples to be able to provide meaningful forecast.

The finding of Zhou et al. (2023) lessens the gap between the work done in LLMs and the realm of time series analysis.

As the above chapter has shown, providing the most precise time series forecast is not as simple, as finding the most advanced model to feed the data to. Each forecasting must be carefully analysed, and the model with the best forecast must be picked to improve the chances of an accurate forecast — if the assumption is that the future data will be similar to the historical development.

To determine what is the most accurate forecast is, the modeller must carefully consider each forecast evaluation metric, but at the same time, consider how the different models handle trend and seasonality. If a model's forecast performs well in terms of its evaluation metric, but is incapable of forecasting each period's seasonal aspects, it can be discussed whether the model is appropriate.

[5] A LLM prompting method where the model is not given examples of expected output.

3 Methodology

Based on the aforementioned literature study, the methodology of the following analysis will be addressed.

3.1 Data Selection

The forecast of a country's GDP holds paramount significance in macroeconomic analysis and policymaking. GDP serves as a comprehensive measure of a nation's economic output, reflecting the overall productivity of an economy and a general signifier of an economy's health. Policymakers rely on GDP forecasts to make informed decisions regarding fiscal and monetary policies, as it provides crucial insights into the current state and future trajectory of the economy.

Forecasting GDP accurately is particularly vital in countries adhering to Keynesian economic principles. Keynesian economics advocates for active government intervention in the economy, especially during economic downturns or periods of inflationary pressure. Timely and precise GDP forecasts enable policymakers to enact appropriate measures, such as fiscal stimulus or tightening, to stabilize the economy and mitigate adverse effects on employment, inflation, and thereby enhancing overall economic well-being.

However, one challenge in utilizing GDP as a policy tool is the significant delay in its publication. Economic data, including GDP figures, often undergoes extensive processing and validation before being released to the public. As cited by Biau and D'Elia (2012, p. 3), this delay can hinder the effectiveness of policy responses, as policymakers may not have real-time information on the economy's performance.

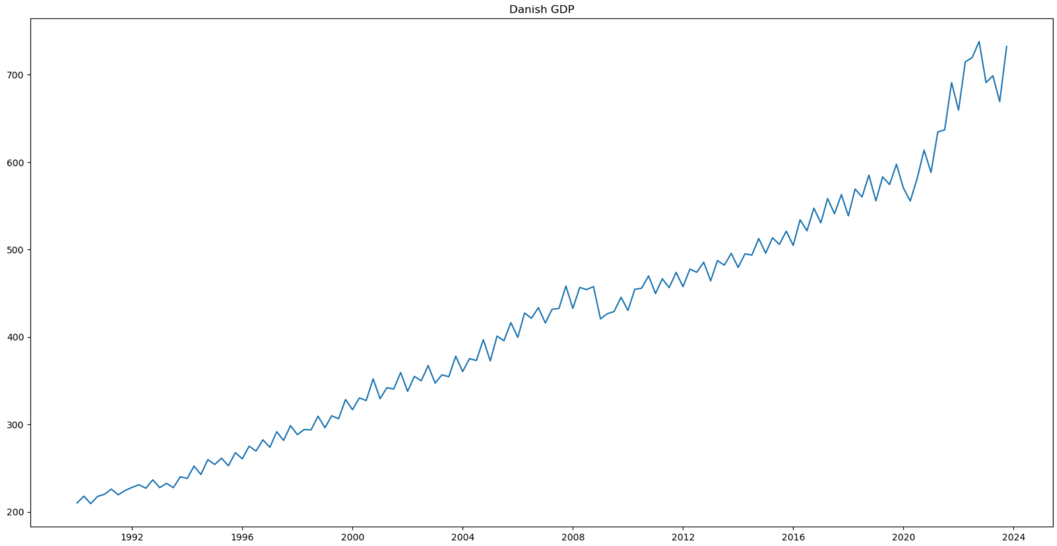
To address this issue, alternative methods that estimate GDP based on more timely indicators have gained prominence. The idea is that these indicators offer insights into economic activity before official GDP figures are available. By leveraging these indicators, policymakers can make proactive decisions to steer the economy in the desired direction, whether it requires stimulus to spur growth or measures to prevent overheating.

The paper will make use of the Consumer and Industry Sentiments statistics collected by Statistics Denmark every month from a representative section of the Danish Consumers and Industry. As Statistics Denmark notes in their analysis, the Consumer Confidence Indicator, does have a significant correlation with the household consumption (Bosanac et al., 2022), and it must be expected that the industry will make future investments based on their current expectations to the future.

As the consequence of trying to forecast GDP through timely indicators, this paper focuses on forecasting the GDP of Denmark using the sentiment indicators. By incorporating consumer and industry survey data from Danish society, the aim is to provide policymakers with a more immediate assessment of economic performance and facilitate timely and effective policy responses.

Data-description

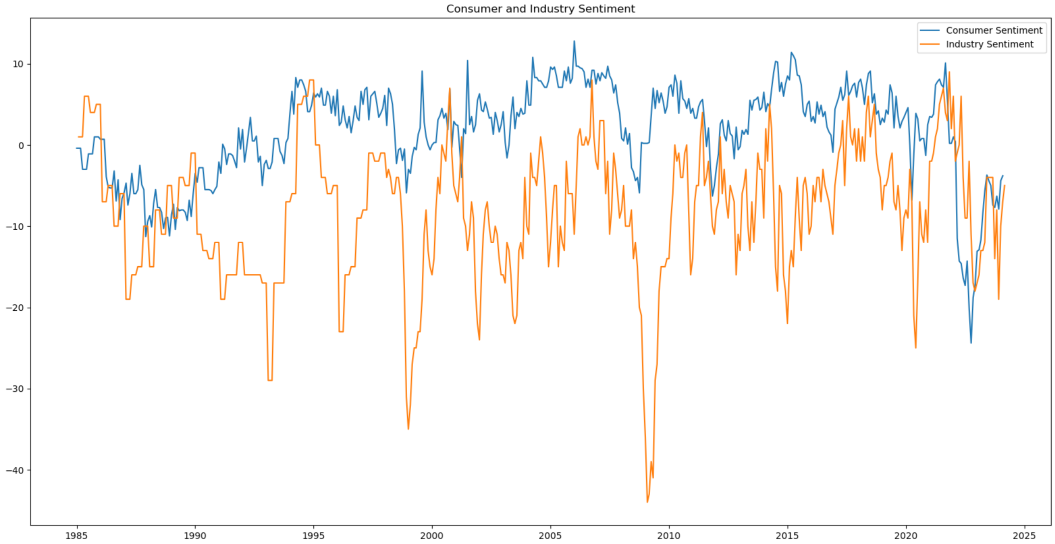
The Danish GDP is found through Statistics Denmark, (StatBank Denmark, 2024) (NKN1), and is a time series running from 1990 Q1, to 2023 Q4. The GDP used in the paper is nominal GDP in DKK (billions), and is not seasonally adjusted. The paper chose the nominal GDP instead of real GDP (with fixed prices), as it can be argued that this variable represents the simplest version of the GDP, while the real GDP is adjusting the GDP retroactively. However, it is important to note that nominal GDP figures can be influenced by inflationary pressures, particularly in periods of higher inflation, such as the post-Covid era examined in this analysis. This is further addressed in subsection 5.2.



(StatBank Denmark, 2024)

Figure 1: Danish Quarterly Gross Domestic Product, 1990–2023

The Danish consumer and industry surveys are conducted by Statistics Denmark and is found through the European Commission's Directorate General for Economic and Financial Affairs (European Commission, 2024a), (European Commission, 2024b). The consumer sentiments consist of twelve questions asked to a representative group of consumers in Denmark, and the industry sentiments consists of 7 questions asked to a representative groups of companies in industry in Denmark. In Figure 2 the representative indicator for the consumer sentiments and industry sentiments are graphed. This representative indicator is an arithmetic mean from a selected set of the questions. This selected set is further described in Appendix A, but it is an OECD standard procedure.



Source: (European Commission, 2024a), (European Commission, 2024b)

Figure 2: Monthly Industry and Consumer Sentiment Indicator, 1985–2024

The time series have been available for every month since 1985. The time series has been filtered in this paper, such that it matches what is available in the dependent variable. Further, the time series has been converted to quarterly values, by taking the arithmetic mean of the monthly values in each quarter. This conversion was necessary such that all input data consisted of the same time format.

3.2 Model Selection

The paper will investigate whether modern machine learning models will perform better at forecasting Danish GDP, than classical econometric models. The paper will therefore continue the structure created in the literature review, by classifying the models into four different levels of machine learning.

Model Type	**Econometric Model**	**Simple Machine Learning**	**Advanced Machine Learning**	**Large Language Model**
Selected Model	ARIMAX & SARIMAX	Gradient Boosting Model	Recurring Neural Network & Long Short-Term Memory	Mistral 7-B Pre-Trained LLM

Table 1: Overview Of Used Models

The first model type will be the classical econometric models. Based on the type of data, with one dependent variable and multiple independent variables, ARIMA modelling will represent the classical econometric genre. Specifically, the paper will test an ARIMAX model and an SARIMAX model[6].

The next model type will be the simple machine learning model. The model, picked to represent this, is the Gradient Boosting model (GB). GB models are ensemble learning methods that combine multiple weak models (typically decision trees) in an iterative manner, where each subsequent model aims to correct the errors made by the previous models. GB models are relatively simple to implement and interpret, making them a suitable choice for representing the class of simple machine learning models.

The third model type will be the advanced machine learning models. Here, the models picked will be a simple Recurring Network Model (RNN) and a Long-Short Term Memory model (LSTM). RNNs and LSTMs are types of neural networks specifically designed to handle sequential data, such as time series. Unlike traditional feedforward neural networks, RNNs and LSTMs have a loop-like structure that allows them to maintain an internal state and model the dependencies between observations at different time steps.

As the last model type, the paper will test relevant transformer models to see whether this newest iteration of the machine learning field can be a relevant addition to the field of time series analysis. In particular, the paper will leverage Mistral AI's open-source LLM called 'Mistral 7-B Instruct v0.2'. The choice of a transformer model is motivated by its ability to capture complex patterns and relationships in sequential data, which could potentially lead to improved forecasting accuracy compared to traditional methods.

[6] In case of Univariate modelling, the proper notation is ARIMA and SARIMA. To align names in graphs, the ARIMAX and SARIMAX will be used in the univariate situations.

3.3 Evaluation Method

To demonstrate the models' reliability, it is essential to evaluate the forecasts accurately. For this task, a holdout validation approach is used to evaluate the models.

A holdout validation method consists of splitting the dataset containing historical observations into a training dataset and a testing dataset. In this paper, we use two scenarios:

1. Test period is from 1990Q1 to 2019Q4 (120 observations). The test period is from 2020Q1 to 2023Q4 (16 observations).
2. Test period is from 1990Q1 to 2015Q4 (104 observations). The test period is from 2016Q1 to 2019Q4 (16 observations).

Each model will be trained upon the data from the training dataset and will be tasked with providing a forecast in the period of the test dataset. When the forecast has been obtained, the forecast can then be compared to the realized values in the test dataset by using specified evaluation metrics.

To select these metrics used to evaluate, it is important to consider what aspects of the model's forecast — and specifically the forecast' error terms — that the paper needs to empathize.

In this paper, a combination of the Mean Average Percentage Error (MAPE) and Root Mean Squared Error (RMSE) will be used to evaluate the forecasts' error terms.

The theory behind the evaluation metrics will be considered in subsection 4.5, but the argument for using the two metrics in combination is as follows;

MAPE is a simple calculation of how 'wrong' a prediction is percentage-wise, summed across the whole prediction period. This makes the metric simple and interpretable. Though, the metric is biassed towards periods with low realized values, as this value is used as the denominator of the equation, before summing.

To counter this issue, RMSE simply squares the error term to equalize the positive and negative error terms. This means positive and negative errors both count towards an accurate evaluation measure. The squaring of the errors, further, means that large error terms are weighted heavier than small error terms in the metric.

Because of both of the metrics' advantages and disadvantages, both needs studying to consider whether a model has provided an accurate forecast.

4 Model Descriptions

The theory section aims to provide a comprehensive understanding of the various models employed in this analysis, catering to the background knowledge and familiarity of economists. While the theoretical foundation of the ARIMA models will be relatively straightforward for those well-versed in econometrics, the subsequent explanations of the ML models may venture into territory typically associated with data science and computational techniques.

The goal of this section is to illustrate the theoretical underpinnings of both classical econometric models and cutting-edge ML techniques. By bridging the gap between these two disciplines, the goal is to enable econometricians and data-scientist to critically evaluate the strengths, limitations, and applicability of each approach in the context of time series forecasting.

4.1 ARIMA Process

To represent a rather simple econometric model, the paper selected the ARIMA process. This type of process was specifically selected, as the paper is trying to forecast a single time series based on a factor of exogenous variables. Furthermore, this model type represents a simple and transparent model, where every prediction will be calculated on rather simple maths.

The specific ARIMA process used in this paper is the ARIMAX(p,d,q) [M] [7] model defined as

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{m=1}^{M} \beta_m X_{mt} + \varepsilon_t$$

(Artley, 2022)

This model considers three types of inputs when it, in the following analysis, returns a forecast value. In the following section, the Autoregressive (ϕ), Moving Average (θ) and the Exogenous (β) aspects will be discussed, before addressing the model as a combined model.

Autoregressive Process

The first aspect of the ARIMAX model is the Autoregressive process (AR). This process investigates the relationship between an observation and its own past observations. The parameter p specifies how many past observations are included in the specific model, and the coefficient ϕ notates the impact of the lagged value

[7] A SARIMAX model is also used in the analysis for handling seasonality. It is described later, but is not part of the following equation.

Y_{t-i} on the current observation. This coefficient is normally calculated through an Ordinary Least Squares (OLS) method.

The OLS method is a fundamental technique in econometrics and statistics used to estimate the parameters of a linear regression model. In the context of time series analysis, the OLS method aims to find the coefficients that minimize the sum of squared residuals, which are the differences between the observed values of the dependent variable and the values predicted by the linear model.

Integration Process

The job of the integration process is to ensure the dependent variable is stationary. In practice, this process uses a Unit-Root test like the Dickey-Fuller test. The Dickey-Fuller test, tests the null-hypothesis that there is a unit root in a simple AR(1) process imposed on the time series (Wooldridge, 2016, p. 575). If a unit root is found present, the time series is not stationary, making it temporarily incompatible with the ARMA aspects. To remove any unit roots, the integration process will take the difference of the time series, to make the time series stationary. d denotes how many 'Orders of Integrations' needed to make the dependent variable stationary.

Moving Average Process

While the objective of the AR process is to capture the relationship between current and past observations, the moving average (MA) process incorporates the relationship between the current observation and past error terms. The parameter p specifies the number of lags considered in the process, while θ estimates the coefficient of each lag ε effect the prediction Y_t. Unlike the AR process, the MA process does not directly model the observations themselves but instead focuses on the residual errors, which can capture any remaining patterns or shocks that are not accounted for by the AR component. OLS is again used to estimate these coefficients.

The MA process can be thought of as an error correction mechanism, as it aims to account for the influence of past errors on the current observation. By incorporating these past errors, the MA process can potentially improve the accuracy of the forecast by adjusting for any systematic patterns or shocks that were not captured by the AR component alone. This approach is particularly useful when dealing with time series data that may exhibit temporary deviations or irregularities not fully explained by the autoregressive component.

Exogenous Variables

As the addition to the original ARIMA model, exogenous variables are a relevant process to consider, especially when trying to forecast future values. Included exogenous variables should be tested for correlation before use, such that we know that the variables does have the ability to explain each other. If they do not, irrelevant exogenous variables introduce complexity and noise to the model.

When estimating the coefficients β, OLS is once again used to estimate the relationship between the dependent variable Y and the m number of exogenous variables X.

Handling of Seasonal Aspects

The following analysis introduces a second ARIMA model in the form of a SARIMAX model. This model handles seasonal aspects separately from the general AR, I, and MA processes. Specifically, a seasonal ARIMA model can be denoted as SARIMA$(p,d,q)(P,D,Q)[s]$:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{k=1}^{P} \Phi_k Y_{t-ks} + \sum_{l=1}^{Q} \Theta_l \varepsilon_{t-ls} + \varepsilon_t$$

(Artley, 2022)

The addition of Φ and Θ elements is the seasonal Autoregressive and Moving Average aspects, where s denotes the length of the seasonal cycle. In terms of the data used in this paper, quarterly observations results in a seasonal cycle of 4. The uppercase P, D and Q values denote the number of lags included in the seasonal elements.

The concept of adding the seasonal aspects of a time series, ensures that the processes consider how previous values in the same season influences current values.

Akaike Information Criteria

When dealing with multiple potential models or model configurations, it is essential to have a criterion for selecting the most appropriate model. The Akaike Information Criterion (AIC) is a widely used metric that helps strike a balance between model fit and model complexity, preventing issues of overfitting or underfitting.

The objective of the AIC is to find the model that best explains the data, while it penalizes models with increasing model complexity (number of parameters). The model with the lowest AIC value is considered the optimal choice, as it achieves the best trade-off between goodness-of-fit and simplicity.

The AIC is expressed by the following equation:

$$AIC = N \times ln\left(\frac{SS_e}{N}\right) + 2K$$

(Manikantan, 2021)

In this equation, N represents the number of observations the model is trained upon, and SS_e represents the sum of squared errors of the model, which captures how well the model fits the data. The term $N \times ln\left(\frac{SS_e}{N}\right)$ is the goodness-of-fit component, with lower values indicating better fit.

The term $2K$ acts as a penalty term, where K is the number of parameters in the model. As the model complexity increases (more parameters), the penalty term increases, counterbalancing the potential improvement in goodness-of-fit due to overfitting.

By minimizing the AIC, the objective is to find the model that achieves the best balance between explaining the data (low SS_e) and model simplicity (low K). Any addition of extra parameters must be justified by a significant drop in the sum of squared errors to compensate for the increased penalty term.

Autocorrelation

Autocorrelation is a fundamental concept in time series analysis that describes the correlation between a time series variable and its own lagged values. In simpler terms, it measures the degree of similarity or dependence between observations at different time points within the same time series (Newbold et al., 2020, p. 712).

Autoregressive (AR) models, such as the AR component in ARIMA models, are specifically designed to capture and model this autocorrelation structure. By including lagged values of the time series as predictors, AR models can effectively capture the inherent dependencies and patterns within the data, leading to better forecasting performance.

On the other hand, if a time series exhibits low or no autocorrelation, it implies that the current value of the variable is largely independent of its past values. In such cases, relying solely on an autoregressive model may not be sufficient for accurate forecasting. Instead, it becomes crucial to incorporate exogenous variables (external factors or related time series) that may have a stronger correlation with the target variable.

The degree of autocorrelation can be quantified using various statistical measures, such as the autocorrelation function (ACF) or the partial autocorrelation function (PACF). These tools help identify the number of lags at which significant autocorrelation exists, providing guidance for selecting the appropriate order of the autoregressive and moving average components in ARIMA models.

4.2 Gradient Boosting Model

The GB model is a simple iterative prediction model, where the iteration process consists of making an estimated prediction and learning from the difference between the prediction and an actual value given to the model. The approach is therefore a simple machine learning approach, where the model will iterate sequentially to minimize a specific loss function. Typically, this loss function will be to minimize the squared residuals between past observations and the models estimates (Masui, 2022).

The GB is an algorithm, that iteratively improves its estimators. To initialize the GB algorithm, a model with a constant value is used:

$$F_0(x) = \underset{\gamma}{argmin} \sum_{i=1}^{n} L(y_i, \gamma)$$

(Masui, 2022)

$F_0(x)$ represents the initial forecast model and γ is the constant value that minimizes the loss function $L(y_i, \gamma)$ for all observations y_i.

The next step is to improve the prediction made from this constant model. The algorithm iterates M times, and the number of iterations is a parameter that can be adjusted during model building. The iterative process involves the following three steps:

1. **Calculate the residuals of the current model**: The residuals, which are the differences between the actual values y_i and the current model's predictions $F_m(x_i)$, help the model identify areas where the current predictions are inaccurate.

2. **Fit a new model to the residuals**: Train a new model on these residuals. The goal is to find a function $h_m(x)$ that minimizes the errors in the residuals, thus improving the overall model $F_m(x)$.

3. **Update the current model**: Update the model by adding the new model's predictions, scaled by a learning rate parameter ρ. This ensures that changes are incremental:

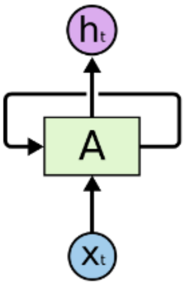
$$F_m(x) = F_{m-1}(x) + \rho h_m(x)$$

The structure of the GB means that the parameters of the models are not defined through economic theory — but are instead only optimized to represent the given time series as best as possible.

A good GB model strikes the balance between encoding the features of a time series and overfitting the parameters. A model, fitted correctly, will understand a time series' general trend, seasonality, and correlation between exogenous variables, giving it a good chance of predicting future values.

4.3 Long-Short Term Memory Model

A LSTM model is a special kind of Recurring Neural Network (RNN). RNN's represent a computational cell-state, which has the inherent ability to 'remember' information over time — opposite to the models described up until now, which need to consider all information from scratch, at all given times. A cell in an RNN can be viewed in figure 3. The idea of the cell is that input at a given time t (x_t), affects the memory cell A, which loops this persistent information between times. The output h_t is then affected both by the given input x_t and the persistent information in A (Olah, 2015).

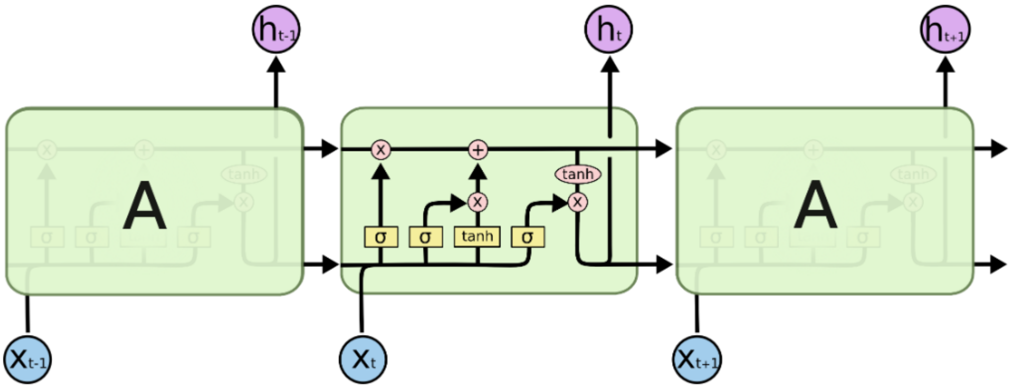


(Olah, 2015)

Figure 3: A Cell In A Recurring Neural Network

The problem with a simple RNN cell is that it does not consider what information should be 'remembered'. This means that at each loop, the information cell A is simply a product of the past inputs represented by x_{t-1}. As Olah (2015) mentions, this feature makes the process simple and intuitive, but creates problems when applying an RNN on real life data. For humans, it is obvious that some information is more relevant than other information — no matter if the information comes in the form of a sentence or a sequence of numbers.

This problem is tried solved in the LSTM model. The model uses the same input, memory and output cells, but expands on how the information is saved in the memory cell.



(Olah, 2015)

Figure 4: A Cell in a LSTM model

In a LSTM memory cell, a set of gates (the yellow boxes) are introduced to restrict the flow of information to and from the cell's memory. Looking from left to right in Figure 4, three gates determine what information is forgotten, what information is updated, and what is outputted:

1. **Forget Gate:** By considering previous outputs h_{t-1} and the new input x_t, a sigmoid neural layer will determine what information the memory cell needs to forget. The sigmoid neuron determines a value between 0 and 1. A value close to zero means that the information is almost entirely forgotten, while a value close to one means the information is remembered.

2. **Input Gate:** A second sigmoid neural layer will determine what information needs updating, where zero means no update necessary and one means to update the value fully. The hyperbolic tangent (tanh) layer takes the input x_t and scales the value to a range of -1 to 1. The combination of the sigmoid and tanh layers determines the weight of the new information in the memory cell.

3. **Output Gate:** Based on the product of the sigmoid coefficient on h_{t-1} and x_t, and a tanh coefficient from the LSTMs memory cell, a given output is returned in the form of h_t.

Based on the above memory process, an optimized LSTM model is capable of pertaining only relevant information from a sequence of data, which in theory should make it good at extrapolating relevant underlying attributes of a time series, and forecasting those attributes.

4.4 Pre-Trained Large Language Models

The latest contribution to the field of time series forecasting, is introduced in papers like Zhou et al. (2023). The fundamental issue with creating a pre-trained time series model, like those trained on large amounts of text, called Large Language Models (LLMs), is that the quantity of time series data available, is significantly less than the quantity of text available for training the LLMs on.

But as mentioned in subsection 2.4, the assumption of papers like Zhou et al. (2023), is that LLMs, trained on large quantities of primarily cohesive pieces of texts, must inherently also be able to return accurate time series forecasts as well. The following section will explain the mechanics of the attention mechanisms in the used LLMs.

Ganesh (2019) explains that the strategy of model building until now has been to forecast exclusively based on the inputted variables. While this strategy is compute-efficient, as it only considers information relevant at the time of the forecast, it begs the question; what if the model was better at understanding the context which it predicts in, by learning seemingly irrelevant information to a specific issue?

To explain how a model can understand a broader context, three important layers of a LLM model should be presented.

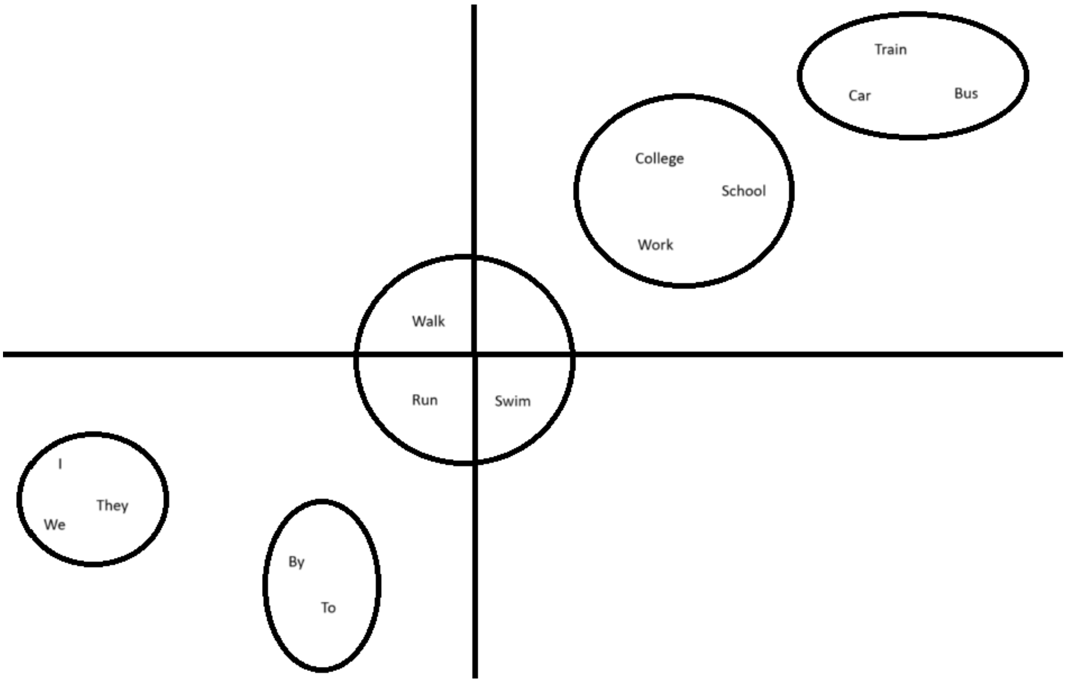
Tokenization and Embedding Layer

To train a LLM, a vast amount of coherent text pieces is needed. An example of such could be to feed the entirety of the Wikipedias encyclopedia to a model.

Tokenization is the process of breaking down the text into individual tokens. In its simplest form, each word can be considered a token. However, for efficiency and handling of subword structures, it is common to break down words into smaller pieces called subword tokens. This allows the model to handle rare and out-of-vocabulary words more effectively. Though, for the following explanation, consider a single word as a token.

Once the text is tokenized, each unique token is assigned a unique vector, known as an embedding. Initially, these vectors consist of fully random numbers. These embeddings are typically represented in a N-dimensional space, where N is the size of the embedding dimension. In simpler terms, N defines how many arbitrary attributes that can be related to each token. As an example, the word 'Stewardess' might be attributed with words like 'woman' and 'plane'. Though, it is important to note that these relationships are not known to the embedding initially. These relationships are first found in when backpropagating.

Backpropagation describes the idea of training the embedding in the LLM. The general idea is to iteratively try and predict the next word in a sentence in a prewritten piece of text. At first, the guess is pure random, as the embedding does not contain any valuable relationships yet. Over time, embedding will begin to provide a probable guess on which word statistically will follow certain word-combinations. By updating the values in each token's vector, the embedding will begin to establish statistical relationships between certain words and context. In the end, such a word embedding can be visualized in 2D in Figure 5.



Inspired by (Madhumita, 2023)

Figure 5: A Representation Of A 2D Word Embedding

Figure 5 visualises how similar words are clustered, symbolising that the embedding have understood the semantic relation between the words inputted. The quality and quantity of the input text determine the proficiency of the model's embedding layer in understanding the general usage of words in a sentence. Depending on one's definition, this may be considered a sort of intelligence.

Attention Layers

For the next layer, the idea is to use this general embedding knowledge in a more specified context. Typically, the LLM will use the knowledge of the embedding to generate new text. Before generating this text, the LLM needs further specific context. The LLM needs this context to consider whether it should use e.g. the word

’interest’ in terms of ’having an interest in painting’ or ’the central bank’s interest rate has risen’(Madhumita, 2023).

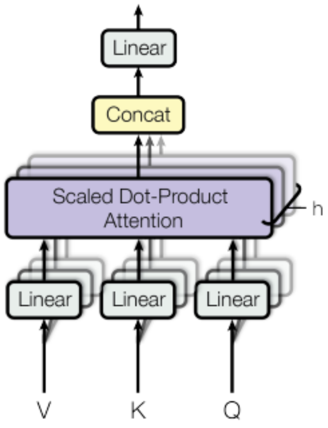
To make this decision, additional specified data is input into the model to provide context. This input could be in the form of a prompt written to the model, giving it concrete instructions on what the current context is.

Output Layer

To output a readable response to what is inputted, the model needs to figure out what to respond with. This concept can be represented as a statistical problem, as the model — based on previous words — needs to figure out the most probable next word in a sentence. As the model knows the general relation between words from the embedding, and the more contextualized relation from the attention layers, it can now predict the most probable next word in a sentence[8].

This predictive process resembles the workings of the LSTM, RNN, and even the econometric models to some degree. If the LLM process is generalised, it too simply predicts unknown values, based on previous values.

The key additional differentiator comes from the revolutionary work introduced by Vaswani et al. (2023) - first introduced in 2017. In their paper, the attention mechanism is introduced, which functions like the LSTM memory cell. Though, the most important changes are that a LLM can process words instead of numbers — through the tokenization process, and a computational ability to run the learning process in parallel, instead of sequentially. This parallel ability is made possible through the multi head attention cell shown in Figure 6.



(Vaswani et al., 2023)

Figure 6: Multi-Head Attention

[8] Often an extra attention layer is used in the output, to decide if the most probable word is the most relevant for the wider model response.

The scaled dot-product attention mechanism enables the parallel computation of the query (Q), key (K), and value (V) matrices through matrix multiplication. This parallelism reduces the computational cost per attention head, as Vaswani et al. (2023, p. 5) explains:

> “Due to the reduced dimension of each head, the total computational cost is similar to that of single-head attention with full dimensionality.”

In practice, this computational advantage allows LLMs to be trained on the vast amounts of input data, giving them the ability to better understand the world.

4.5 Evaluation Methods

For each of the models presented above, a standard approach to evaluating the models’ predictions has been chosen. A combination of Mean Absolute Square Errors (MAPE) and Root Mean Squared Errors (RMSE) will be used as evaluation metrics.

Mean Absolute Squared Error

Mean Absolute Square Error (MAPE) represents the absolute sum of each error at a given time ($\hat{y}_t - y_t$), divided by the actual value (y_t). By multiplying with 100, the metric represents the error as a percentage (Vandeput, 2019):

$$MAPE = \frac{1}{n}\sum\left|\frac{\hat{y}_t - y_t}{y_t}\right| \times 100$$

By using the absolute values in the equation, the metric ensures that an equal positive and negative error term will not even out the average evaluation metric across the two periods.

This metric is widely used in evaluating a model’s performance, as it represents the ’deviation’ from the realized values as a simple percentage value. While this simplicity means that the metric is easy to interpret, it also means that, it — in some cases — can present too simplistic a view of a model’s performance.

The disadvantage of the MAPE metric is that each error-term is divided with the realized observation before being summed. This results in a situation where the same numerical sized error term is penalized differently, depending on the current realized observation. Specifically, in periods with low realized values, a specific error term will return a relative high MAPE metric, compared to another period where the same error term is divided by a higher realized value. This issue causes the MAPE to be biased towards periods with low realized values, as these periods’

MAPE scores will affect the overall MAPE score of the forecast comparably more when summed.

In general, it should also be noted, that a result of dividing with the actual values (y_t), the MAPE will become undefined if evaluating actual values equal to zero. This is especially important to note when using growth or differenced values, as a tool to make variables stationary.

The MAPE should therefore be carefully examined per forecast, and is at most use when used as a rather simple performance metric. Because of the metric's biasness, it will be especially inaccurate in situations with great variance in the forecast. As GDP has significant seasonal aspects, the MAPE metric will not be the sole evaluation metric in the following analysis.

Root Mean Squared Error

The above problem of MAPE is tried addressed in the Root Mean Square Error (RMSE) metric. In this metric, the error term $(\hat{y}_t - y_t)$ is squared to ensure that a positive and negative error both equally count in the final metric (Vandeput, 2019):

$$RMSE = \sqrt{\frac{1}{n}\sum(\hat{y}_t - y_t)^2}$$

When the squared error terms have been summed, the square root is applied to return the value to the original scale.

This metric is therefore better at measuring both positive and negative error terms, but it does lose some interpretability. While the scale is returned to the original, it is now impossible to tell whether the predictions do have a bias towards negative or positive error terms because of the squaring.

Furthermore, the operation of squaring the error terms means that large values are penalized relatively more than small error terms. So while the scale of the RMSE metric represents the original scale of a specific time series, the results of the RMSE calculation are not linearly correlated with the original values, because of this squaring operation.

A combination of these two evaluation metrics will be used to deem whether each model performs accurately.

5 Testing Framework

To test each model effectively and as thorough as possible, a combination of scenarios is introduced in the forecasts, as done in (Maccarrone et al., 2021). This framework ensures that all models are evaluated in such a way that a model's advantages in some scenarios can be highlighted, while its disadvantages can as well. The models presented in subsection 3.2 will be implemented for each of the following scenario-combinations.

5.1 One-Step Ahead vs. Multi-Step Ahead Forecasting

The task of both scenarios is to forecast 16 periods, based on previous training data.

The multi-step ahead models will try to predict the entire test period in one compute, based on this training data. A result of computing multi-steps ahead is that, besides the first forecast-period, each subsequent forecast is based on a previous forecast. This will realistically result in the model's forecast drifting further from the actual observations each period, as the model does not have a chance to evaluate whether its forecast is correct iteratively.

The one-step ahead models will instead implement a sliding window, such that the model computes 16 individual forecasts — one step into the future, based on the training data, plus any observations from the test dataset that are from before the given prediction period. This gives the models the ability to reset their forecast such that drifting forecasts will not be an issue.

The reason for introducing this difference in scenarios is that economists and policymakers have typically been most interested in the short-term forecasts. The argument has been that long-term forecasting of variables like the GDP is impossible to predict, as unknown variables historically have impacted GDP in ways economists and alike have not been able to foresee. On the contrary, policymakers will always be interested in getting a head start on the coming period's economic indicators. If a forecast can estimate these indicators, it would improve the policymaker's time to act.

A relevant test to introduce to the more advanced ML and AI models is therefore, to see, whether these model types are better than classical econometric models in either one-step ahead or multi-step ahead forecasting scenarios.

5.2 Pre-Covid vs. Post-Covid Test Period

In the past four years, society has adapted its way through a global pandemic. While the pandemic has caused issues in all sorts of ways, it too has affected the Danish GDP. The plot in Figure 1 shows how the attributes of the time series of Danish GDP has fundamentally changed following the Covid-19 Pandemic. Specifically after year 2020, the time series cannot be described as following a certain trend line, with some seasonal aspects causing some short-term deviances from the trend line.

The paper is not meant to analyse the cause of the change in the underlying attributes of Danish GDP. But, referencing an analysis from the European Commission (2023, p. 2), the immediate decline in GDP at the start of 2020 is largely attributed to Covid-19. The following strong growth periods, is because of a strong political will to reimburse companies that held on to their workers under the pandemic. This made it easy for Danish companies to restart their production as soon as possible.

The European Commission contributes the strong growth rates of 2022 to the temporary spike in sea freight rates, meaning Danish companies like Maersk and DFDS saw significant profit growth. Similarly, Olsen et al. (2024, pp. 14–17) attributes the growth of the Danish company Novo Nordisk, as the primary growth engine of the entire Danish economy in 2023.

It is also important to note that the GDP in the period 2021-2023 has been correlated with higher-than-normal inflation rates. As this paper is using nominal GDP (measured in current prices), the rising inflation directly impacts the GDP figures during this period. Inflation can cause nominal GDP to increase, even if the real economic output (the actual quantity of goods and services produced) remains relatively stable or grows at a slower pace. This is because higher prices for goods and services lead to an increase in the value of economic transactions, inflating the nominal GDP figures.

These above examples mean that to evaluate the models properly, the models should be tested on both periods. The pre-covid test scenario represents a period with 'normal' growth — without significant external disturbances. The post-covid test scenario oppositely represents a period with many external factors impacting the endogenous variable significantly. It is therefore relevant to test each model to see whether the type is relatively better in one scenario than the other scenario.

5.3 Univariate Forecasting vs. Consumer & Industry Sentiment Indicators

As mentioned briefly in subsection 3.1, the paper tries to test whether it is possible to forecast GDP based on more timely indicators, like the consumer and industry sentiment indicators. To test this, the models are again divided into multiple scenarios. The sentiment indicators are twelve monthly questions asked a representative set of consumers, and seven[9] monthly questions asked a representative set of companies in the industry sector.

A scenario is therefore created where each model is given all eighteen answers to each question as individual exogenous variables[10]. The hypothesis is that this approach will let the advanced models decide which of the individual responses that correlate with the endogenous variable, and then use these exogenous variables in the forecasting task.

An alternative hypothesis could oppositely argue that feeding 18 exogenous variables into some models, especially the simpler models, could introduce noise to the system, resulting in the forecast not being able to utilize the exogenous variables effectively. To overcome this theory, a second scenario is built where the sentiment indicators are simplified into two exogenous variables, one for consumer sentiment and one for industry sentiment. The two simplified indicators are available in the same datasets, where the European Commission has found a representative indicator based on selected responses, further described in Appendix A.

Lastly, to overcome the possibility of these sentiment indicators are only introducing noise to the models, a third scenario is created, where no exogenous variables are used to predict GDP. Only past observations of the endogenous variable are used in the forecasts.

In the end, these scenarios mean that all models are tested on $2 \times 2 \times 3 = 12$ individual scenarios. With the six models presented in Table 1, this means that the below analysis has 72 evaluations metrics to analyse.

[9] This paper utilizes six of the seven questions, as question no. 6 has not been asked in the period 1990-1998.

[10] As denoted in European Commission (2024a), each indicator is the calculated as: % of positive responses, subtracting % of negative responses, returning an indicator between -100 and 100.

6 Results

Each model type has been implemented to maximize performance based on the specified scenarios. This implementation will be introduced before further presentation of the actual results.

For the ARIMAX(p, d, q) and the SARIMAX$(p, d, q)(P, D, Q)$, this means that the specific combination of (p, d, q) values has been selected based on which combination minimizes the AIC value on the training data. By selecting the model with the lowest AIC value, the goal is to select the model balancing the goodness-of-fit and simplicity, as further described in section 4.1.

By simply letting the lowest AIC value select the model used in the analysis, it can be critiqued by not incorporating the empirical understanding of the analysed time series. For the used variables in this paper, it is given that the time series data consist of quarterly observations. Some econometricians would hence argue that it would be logical to include approximately four lags in both the p and q values, and any AIC value must be ignored in the selection process.

However, optimizing the ARIMAX models based on the lowest AIC value ensures they are fitted optimally by design. This process also resembles the following model types, as finding the lowest AIC value is an iterative process, resulting in a relative higher compute cost.

For the GB models, a grid search approach has similarly been implemented. The goal of the grid search is to minimize the RMSE value of a model fitted on the training data. A grid search approach can be compared to the above ARIMAX method, in such that a grid search iterates over a set of hyperparameters, finding the specific combination of these parameters that will minimize the RMSE value. For the GB models, the following hyperparameters were optimized using grid search:

- Max Depth: Controls the depth of each decision tree. A high tree depth will make the model complex and likely to overfit.
- Min Child Weight: Specifies the minimum sum of instance weights needed in a child. Lower values allow the model to learn more detailed patterns, which might include noise.
- Gamma: Decides whether a tree splitting will result in a significant drop in the loss function, compared to the increase in complexity. A high gamma value will require a large decrease in the loss function to make a tree split.
- Subsample: Sets the percentage of a training dataset used in each round. A low subsample rate will make the model less sensitive to specific observations, but can also result in underfitting.

- Column Subsampling: Sets the percentage of columns used in each round. If the model is not using all columns, it can better understand the difference each column provides.

For the RNN and LSTM models, Mean Squared Error (MSE) was used as the loss function during compilation. The Adaptive Moment Estimation (Adam) optimizer was used to optimize the models, offering an efficient approach for minimizing the loss function. The optimization algorithm effectively ensures that a model converges towards minimizing the loss function.

Finally, the paper has used the Mistral-7B-Instruct-v0.2 model to represent the LLM model's ability to forecast. The model has been selected, as the model at the time of writing, is the best performing open-source LLM model, meaning it has outperformed the previous best model, Meta's Llama2 model, on all evaluated benchmarks (Jiang et al., 2023).

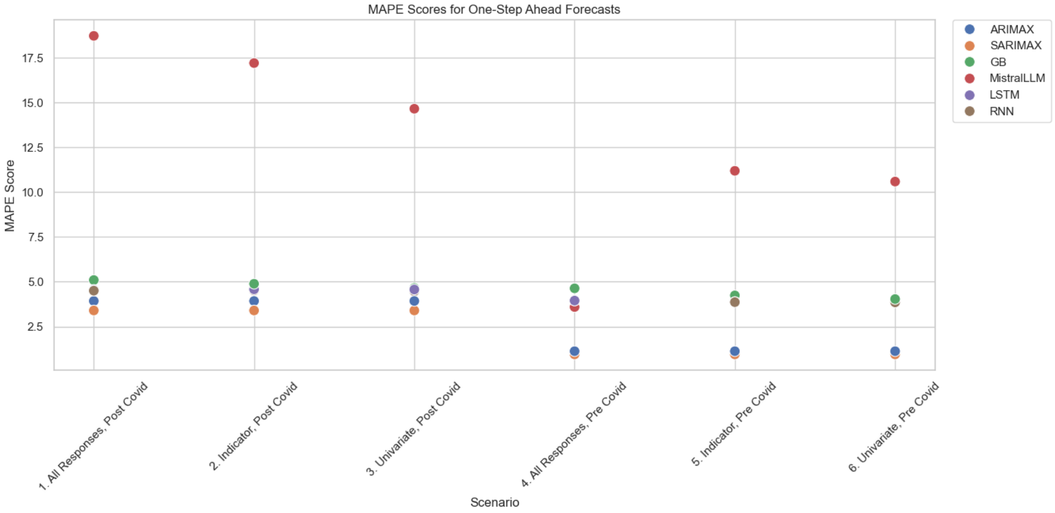
Due to the LLMs lack of transparency in its prediction process, further described in section 9, the process of returning a prediction from this model is less clinical. A general prompt structure has been used across the scenarios, but the specific prompt has been fine-tuned for each scenario, such that a cohesive forecast has been returned each time.

6.1 MAPE Results

The Mean Average Percentage Error (MAPE) scores of each model type in the one-step ahead scenarios are displayed in Figure 7. The figure is the first introduction of how the models compare — both in terms of each other, but also how they differ when subjected to different scenarios.

Figure 7 displays that every model type is better at predicting in the pre-covid scenario, compared to post-covid. This underscores the importance of testing models across multiple time scenarios, as discussed in subsection 5.2.

Similarly, it seems that are a general trend of the forecasts is that they are more accurate, the less exogenous variables are included. This trend is far more significant in the post-covid scenario, suggesting that the consumer and industry sentiments might have lost correlation to GDP in this period. So while Bosanac et al. (2022) finds that consumer sentiments do correlate with GDP in the case of Denmark, this correlation does not seem significant enough to improve GDP forecasts beyond what a univariate model can achieve.



Source: Author's calculations. Code available through Github.

Figure 7: One-Step Ahead MAPE Metrics

Looking at the difference in models in Figure 7, the results are pretty clear.

For all scenarios, the SARIMAX model outperforms the rest — with the normal ARIMAX model close behind. The GB, RNN and LSTM models are below 5% MAPE for all Scenarios.

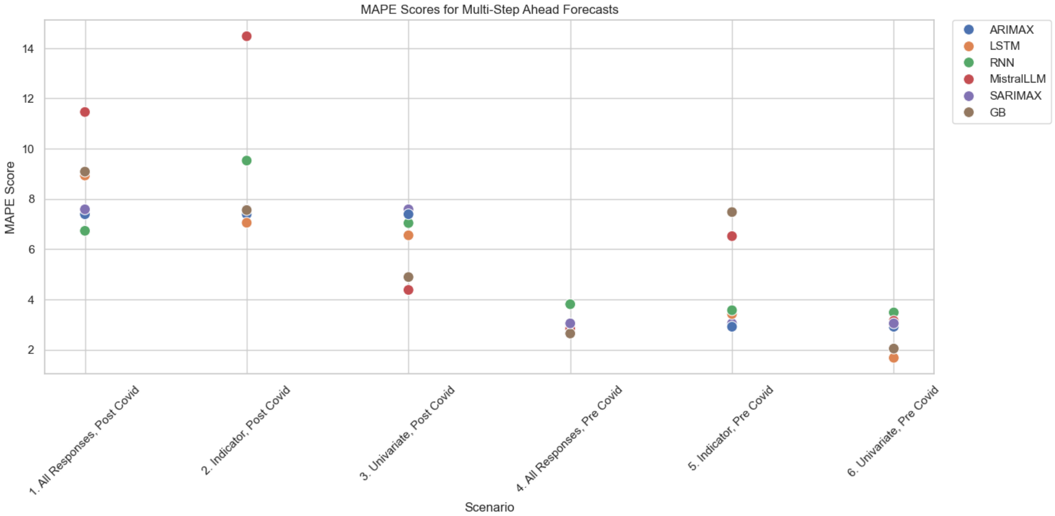
Lastly, The Mistral 7B LLM model is in most scenarios way off, with a MAPE score that much higher than the rest of the models. The only scenario where this is not the case, is in scenario 4, where it outcompetes the LSTM, RNN and GB models.

While the SARIMAX and ARIMAX models are best in every scenario, it is important to acknowledge that these models see a performance boost in the pre-covid scenarios, while they are somewhat clustered with the ML models in the post-covid scenarios.

In Figure 8 the MAPE scores of the multi-step ahead scenarios mapped. When comparing to Figure 7, important distinctions becomes visible.

In terms of the models' ability to predict across scenarios, the same trends as in the one-step ahead forecast is found. The forecasts are more precise in the pre-covid scenarios and the models perform better, the less exogenous variables that are included — in general.

Though, it appears that the ML models prefer to forecast based on all the responses to the sentiment queries, compared to the single indicator. This can suggest that these models have isolated the relevant individual responses. This learning process is hindered when the sentiments are generalised into a single indicator.



Source: Author's calculations. Code available through Github.

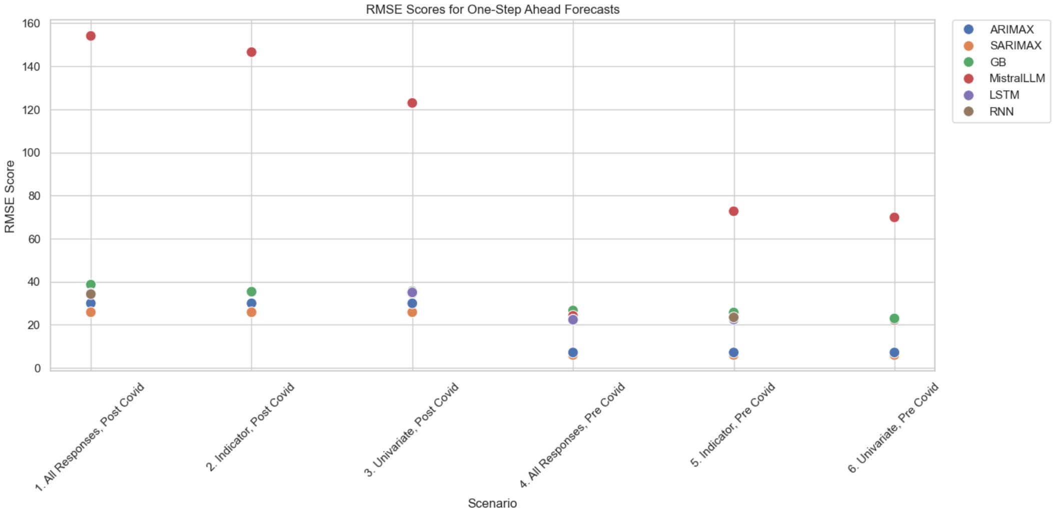
Figure 8: Multi-Step Ahead MAPE Metrics

In the multi-step ahead scenarios in Figure 8, the ARIMA models are best in a single scenario — scenario 5. The GB model is best in scenario 4, the RNN is best in scenario 1, and LSTM outperforms the rest in scenario 2 and 6.

Interestingly, the Mistral LLM forecast outperformed the other models in scenario 3. Further, the LLM predictions were clustered with the other models forecasts in scenario 4, 6, and 5 to some degree. This could suggest that the Mistral model is better suited at long-term forecast than short-term.

6.2 RMSE Results

To complement the above MAPE scores, the RMSE scores are presented in this section. As described in subsection 4.5, there might be slight differences in the results analysed by a MAPE metric and a RMSE metric. The following section will focus on these slight differences from the initial analysis.



Source: Author's calculations. Code available through Github.

Figure 9: One-Step Ahead RMSE Metrics

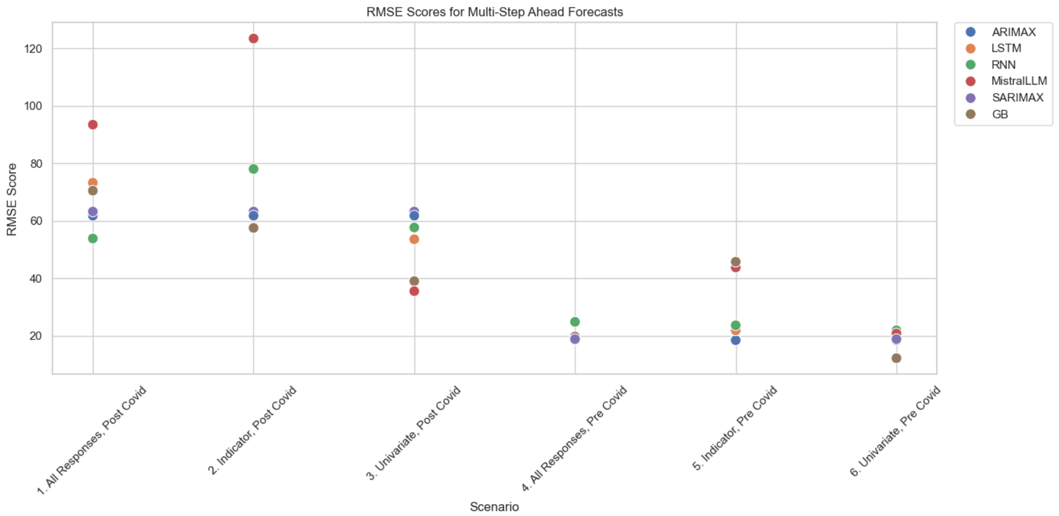
For the one-step ahead RMSE scores presented in Figure 9, the overall results are consistent with the MAPE scores in Figure 7.

It is important to note that the relative distance between the LLM models and the rest of the models is more pronounced in the RMSE figure than in the MAPE figure. This discrepancy can be attributed to the inherent properties of the MAPE and RMSE metrics.

The MAPE metric, due to its formulation of dividing the error by the actual value, tends to be more sensitive to errors occurring during periods with low realized values. As a result, the MAPE may understate the relative performance differences between models during periods with higher realized values.

On the other hand, the RMSE metric squares the errors before averaging, effectively giving more weight to larger errors. This property of the RMSE can amplify the relative performance differences between models, particularly when one model consistently produces larger errors across a range of realized values.

The larger distance between the models in figure Figure 9 is attributed to the RMSE inherent bias towards large error terms.



Source: Author's calculations. Code available through Github.

Figure 10: Multi-Step Ahead RMSE Metrics

For the multi-step ahead RMSE scores in Figure 10, some changes are observed. In scenario 2, the GB forecast now performs as well as the LSTM forecast. In scenario 4, the SARIMAX forecast outperforms the best performing MAPE model, the GB forecast. Lastly, in scenario 6, the GB forecast outperforms the LSTM forecast

These changes are small, but are enough to change the outcome of the best performing models in some cases. Though, it must be noted that these small changes do not change the outcome of which model type that performs best in most instances. It is only in scenario 4 that the Econometric model performs better than the simple ML model, depending on looking at MAPE or RMSE.

These changes do signify the need for incorporating multiple evaluation metrics into the analysis. By doing this, a forecaster is actively choosing what aspects of a forecast that are important, when selecting the optimal forecast strategy.

7 Discussion

The general results of the analysis, presented in section 6, reveal that the classical econometric models outperform the rest when short term forecasting (one-step ahead). But interestingly, in the long-term forecasts (multi-step ahead), the ML and LLM models are more accurate in their predictions, in most scenarios.

It can be argued that this difference can be attributed to the relative ease of the short-term forecasting, where the uncertainty of the future is smaller. This means that for a model to be better than the other models, it needs to be precise in its calculations. This precision can be found in the ARIMA modelling, where every forecast value has a strong and simple mathematical reasoning behind it.

In contrast, forecasting further into the future inevitably leads to a loss of accuracy due to increased uncertainty. An effective long-term forecasting model must therefore be able to capture the inherent non-linear dynamics of the time series more accurately than a simple ARIMA model.

When aggregating a long-term forecast from an ARIMA model, it often appears to primarily capture the general trend of the data trained upon, as the model's ability to capture seasonal or cyclical patterns will diminish in the aggregation over longer time horizons.

Machine learning models excel in this aspect, as their advanced learning capabilities allow them to grasp complex, long-term attributes of the data, surpassing the limitations of ARIMA's autoregressive and moving average processes. This ability to understand and model intricate patterns makes ML models more precise for long-term forecasts.

The use case of the individual forecast is therefore central for which model type to use. If the use case is to forecast a general trend of long term economic development, then ML models might be relevant to use.

Though, as the paper started by stating, GDP forecast is typically used in a short-term scenario. These forecasts are typically estimated one quarter ahead — or an end of year forecast. These forecast require precision, as a difference of 2% will be the difference between economic growth or stagnation. Such difference will result in different policies implemented by a sitting government.

As the ARIMA models did outperform the rest of the models in the one-step ahead scenarios, it must therefore be recommended to continue the use of econometric models in economic forecasting.

Given the importance of central economic variables like GDP, it is crucial to consider the accuracy of each individual forecast. The precision required in these forecasts because of their significant policy implications, makes econometric models a more reliable choice.

7.1 Compute Resources

While mentioned previously in the paper as a sidenote, an important consideration of using ML and AI models in forecasting is their need for bigger compute resources.

The resources needed to implement ML models like GB and RNNs is higher than implementing ARIMA models. Though, they are both pale compared to the rise of LLMs. The general use of LLMs comes with a rise in energy usage. As Alex De Vries, PhD at VU Amsterdam, has calculated;

> *"A single LLM interaction may consume as much power as leaving a low-brightness LED lightbulb on for one hour."*
>
> (Wells, 2023)

As further argued by Wells (2023), the current energy-mix of any country on earth is still not fully renewable, and therefore will any extra energy used typically come with a carbon-cost.

The increased compute resources required for running LLMs directly translate into higher economic costs for the institutions deploying these models. Consequently, the economic barrier to entry will rise if LLMs become the standard for forecasting in the future. While most established institutions currently providing economic forecasts may overcome this barrier, it is crucial to research how this economic threshold will affect institutions with limited resources.

Assuming a relationship between the accuracy of a country's economic forecast and its ability to navigate economic situations, access to compute resources will become critical for a country's prosperity if LLMs become the forecasting standard. This could widen the gap between well-resourced institutions and those with fewer resources, potentially exacerbating economic inequalities. Developing countries or smaller institutions might struggle to afford the necessary infrastructure, limiting their ability to produce accurate forecasts and, by extension, effectively manage their economic policies.

Any use of the these more advanced models should therefore always consider whether the benefits of using these can balance the extra cost of using them — both in terms of economic cost and environmental cost.

7.2 The Sentiment Divergence

As mentioned in subsection 3.1, consumer and industry sentiment indicators have historically served as reliable proxies for gauging the overall economic conditions, often closely aligning with key macroeconomic indicators like GDP. However, the analysis presented in subsection 6.1 revealed a notable divergence between sentiment data and actual GDP growth, particularly in the post-Covid scenarios.

The beginning of the Covid-19 pandemic in early 2020 resulted in a temporary decline in GDP, which was soon followed by a rapid recovery and unprecedented growth rates not observed in the past decade. Paradoxically, consumer sentiment plummeted to an all-time low of -24.4 in October 2022, as illustrated in Figure 2 earlier in the paper, suggesting a significant disconnect between the public's perception of economic conditions and the underlying economic data.

The implications of this divergence are clearly reflected in the forecasting performance of the various models, particularly the advanced machine learning models like the LSTM. As shown in Figure 11, the multi-step ahead forecast from a LSTM model trained on sentiment data predicted a substantial decline in GDP, contrary to the actual economic trajectory.

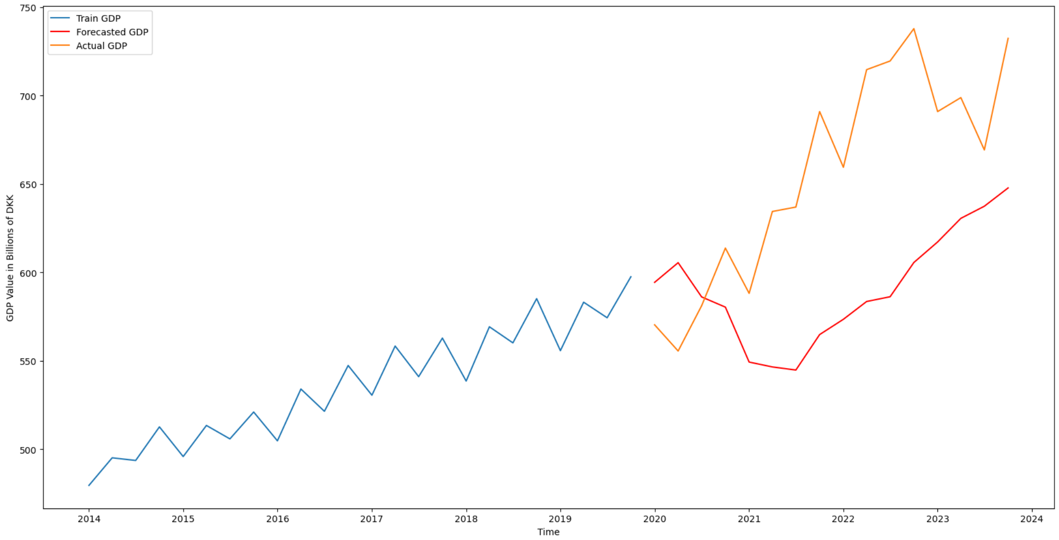


Figure 11: A Post-Covid LSTM Forecast

This phenomenon, dubbed the "Vibescession" by economic commentator Kyla Scanlon (2022), refers to a period where sentiment indicators decline temporarily, while economic data such as trade and industrial activity remain relatively stable or even positive.

As Paul Krugman (2024) further cemented the term, the Vibescession highlights the potential disconnect between public perception and actual economic conditions.

The divergence between the forecasted GDP based on sentiment data and the realized GDP values underscores a fundamental challenge in econometric modelling: the assumption of a causal relationship between a set of exogenous variables and the target economic variable.

In reality, economic metrics are influenced by a multitude of complex and often unquantifiable factors, making it difficult to fully capture the underlying dynamics through a limited set of variables.

This observation aligns with the famous quote by statistician George E. Box: "Every model is wrong, but some are useful." (G. Box, 1979, p. 2).

Traditional econometric models, while valuable tools, inevitably oversimplify the intricate relationships and interactions that shape economic outcomes. The introduction of LLMs presents an exciting opportunity to address these limitations. By leveraging their ability to process and integrate vast amounts of data from diverse sources, LLMs hold the potential to capture the complexities and non-linear relationships that traditional econometric models struggle with. Their capacity to incorporate contextual information and unstructured data could prove invaluable in modelling the multifaceted factors that influence economic variables.

However, it is important to acknowledge the challenges associated with integrating LLMs into econometric modelling. The inherent biases and limitations of the training data used for LLMs could introduce new sources of error or uncertainty in economic forecasting. Further, the discussion in subsection 7.1 highlights that the computational resources needed for these types of models must be considered.

Nonetheless, the observed divergence between sentiment indicators and economic growth during the post-Covid period highlights the need for more sophisticated modelling techniques that can account for the complex interplay of factors shaping economic outcomes.

The potential of LLMs to capture these complexities warrants further research and exploration, as they could pave the way for more accurate and robust econometric modelling and forecasting.

8 Implications

The above findings detail the evolution of time series analysis. While it could have been believed that this evolution must have improved the methodologies for the better over time, instead this thesis has depicted a more nuanced picture.

The thesis suggests that ML models and LLM models do have a role in future time series analysis projects, but each project needs to specify the objective of the project and select the optimal model type for achieving this objective.

Importantly, the research underscores that classical econometric models, such as ARIMA, remain relevant. These models offer explainable results grounded in previous observations, making them preferable when the characteristics of the variable are well-understood. The transparency and simplicity of these models allow for more straightforward interpretation and communication of results.

In contrast, when the attributes of a variable are less clear, ML methods may be more appropriate. While these models often sacrifice some degree of explainability, they can uncover complex patterns and relationships that are not immediately apparent to the model builder. This ability to leverage vast amounts of data and identify intricate correlations makes ML models valuable for long-term forecasting and scenarios where traditional assumptions may not hold.

8.1 Future Work

The idea of incorporating LLMs in time series analysis stems from Zhou et al. (2023) article, arguing that LLMs can output relevant numerical forecasts. While this paper adds to this research, further research papers may reach other results, either by prompting the LLMs differently, or by using better models, not available at the time of this paper's deadline.

While classical time series analysis requires a technical background to ensure that the model performs well, LLMs introduces a linguistical challenge of prompting correctly. The prompting process still requires the knowledge of time series terminology, but future works should consider incorporating linguistical fields of studies in their future process.

This paper has used the Consumer and Industry Sentiments in the form of a numeric indicator. This is needed when used in time series models, but future work could research the use of qualitative sentiment answers in LLM forecasts. Economic opinion pieces written in newspapers could, as an example, provide valuable insight for a LLM to provide a forecast based upon.

Further recommended research is also to investigate the LLMs capabilities in other econometric fields. This paper has focused on the models' ability to forecast a time series. While this is a central field in econometric studies, it is not the only field where LLMs might prove its relevancy.

The paper especially recommends future works to investigate whether LLMs could contribute positively to the econometrics field of Policy Events Studies. It could be imagined, that similar to the forecast field, a LLM will be able to predict the impact effectively of a policy change, based on the model's inherent understanding of a society.

9 Limitations

This project has to some degree been limited by the access to compute resources. ARIMA, GB, and the RNN models are relatively easy to run on low-end consumer compute units. LLMs are, on the other hand, virtually impossible to run effectively without a Graphical Processing Unit (GPU) with CUDA cores[11] and a high amount of compute memory. While AAU has such resources available, a lengthy application process meant that the prompt refinement process was limited in time.

It is also important to note that implementing this kind of backward looking forecast evaluations does require a footnote when testing the LLMs. The LLM's prompt has specifically noted that the model cannot use any information gained on or after the period it is trying to predict. Ideally, it means that the model uses its general knowledge acquired before the time of the prediction, though this is not possible to know.

Large amounts of compute are needed to generate a coherent text from such a model, alias, it is not viable to also analyse each step the model took to investigate how the model reached this result.

A relevant piece of information would be to know what training information the model has been trained upon. If the training data was publicised, it too would be known in what period it has been trained upon. This type of information is not public, as each company regards it as a company secret, even the open-source models.

[11] A computing architecture that allows parallel compute on a GPU.

10 Conclusion

This paper has investigated the forecasting abilities of various time series models, spanning from classical statistical models in terms of ARIMA models, to simple and advanced Machine Learning models, and Large Language Model forecasting.

The literature review in section 2 highlighted the evolution of forecasting methods and indicated a gap: a comprehensive comparison of different models using consistent input data on a real-world problem. This gap motivated our comparative study, particularly focusing on forecasting Danish Gross Domestic Product (GDP).

For section 4, the theoretical foundation of each model type is described. The classical ARIMA models are clearly structured based on rather simple concepts. Specifically, the idea of a time series being able to be estimated on its inherent autoregressive features and some degree of error correction in terms of the moving average of past error terms.

The machine learning models introduce the concept of iterative learning, where simple ML models like the Gradient Boosting uses learning trees to iteratively estimate an optimal model. More advanced ML models like, especially, the Long Short-Term Memory (LSTM) model introduces a memory cell, causing these models to further estimate a model based on past estimations.

Lastly, the concept of Large Language Models (LLM) is introduced. The relevancy for time series estimation comes from the hypothesis that a trained LLM inherently has a general knowledge of economic time series attributes in their embedding. This knowledge can be harnessed to provide accurate time series forecasts that not only rely on the attributes of the time series, but the attributes of the societal concept that the economic time series is representing.

The framework outlined in section 5 sets up multiple scenarios to robustly test model performance, including pre- and post-Covid periods, varying numbers of exogenous variables, and both one-step and multi-step ahead forecasts. These scenarios allowed us to rigorously analyse model robustness and adaptability under different conditions.

Section 6 presents the models forecast' evaluation metrics of each model type across the 12 scenarios. It was found that the ARIMA models performed best in the one-step ahead scenarios. The models performed significantly worse when forecasting the post-Covid period, and it seemed that the exogenous variables generally introduced noise to the models, resulting in worse forecasts.

The ML models performed best in the multi-step ahead forecasts, but the above observations seem to generally apply to the multi-step ahead results as well. Though it seemed that some ML models were able to extract some relevant information from some sentiment variables, and these models were able to return a more precise forecast than the same model without the exogenous variables. This ability shows an advantage of the ML models, which are better at learning the intricate details of a relation between variables.

Section 7 delves into the implications of the above results. It is argued that the results suggest that the econometrical models, represented by the ARIMA model, is better at forecasting short-term, as the margin of error is smaller. The straightforward statistical approach of the ARIMA process provides this accuracy, as shown in the results.

On the contrary, long-term forecasts seem to require inherent knowledge of the time series, which must be better learnt in the ML models.

Subsection 7.2 further discusses the inherent problem of forecasting. By using exogenous variables in a forecast, the model typically assumes a constant correlation across time. The paper commends the new LLM approach as a possible solution to this problem. By using a model with a general intelligence, the future of forecasting will be exposed to a paradigm shift in the coming years.

In summary, this thesis demonstrates that while traditional models remain valuable for certain applications, advanced AI methodologies, particularly ML and LLMs, offer significant advantages in capturing complex patterns and improving long-term forecasts. The integration of these advanced models could enhance the accuracy and reliability of economic forecasting, paving the way for more informed decision-making in economic policymaking.

References

Artley, B. (2022, April). Time Series Forecasting with ARIMA , SARIMA and SARIMAX. Retrieved May 28, 2024, from https://towardsdatascience.com/time-series-forecasting-with-arima-sarima-and-sarimax-ee61099e78f6

Biau, O., & D'Elia, A. (2012). Euro area GDP forecasting using large survey datasets: A random forest approach. *6th Eurostat Colloquium on Modern Tools for Business Cycle Analysis: the lessons from global economic crisis, held in Luxembourg, 26th - 29th September 2010*, (6), 1–24. https://doi.org/10.2901/1977-3331.2011.002

Bosanac, Z., Paludan-Müller, G., & Rose-Nielsen, K. M. (2022, March). *Kan forbrugertillidsindikatoren give en indikation om udviklingen i husholdningernes forbrug?* (Tech. rep.). Danmarks Statistik. København. Retrieved May 17, 2024, from https://www.dst.dk/Site/Dst/Udgivelser/nyt/GetAnalyse.aspx?cid=47869

Box, G. E. P., & Jenkins, G. M. (1979). *Time series analysis: Forecasting and control.* Holden-Day.

Box, G. (1979). Robustness in the Strategy of Scientific Model Building. In *Robustness in Statistics* (pp. 201–236). Elsevier. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Christensen, T., Kling-Petersen, A. R., & Nørgaard, C. (2024). Kunstig intelligens barberer en time af ventetiden på hospital. *Danmarks Radio*. Retrieved May 9, 2024, from https://www.dr.dk/nyheder/indland/kunstig-intelligens-barberer-en-time-af-ventetiden-paa-hospital

European Commission. (2023). *2023 country report: Denmark* [OCLC: 1400079528]. Publications Office of the European Union.

European Commission. (2024a, February). Consumer Survey. Retrieved March 3, 2024, from https://economy-finance.ec.europa.eu/economic-forecast-and-surveys/business-and-consumer-surveys/download-business-and-consumer-survey-data/time-series_en#all-surveys

European Commission. (2024b, February). Industry Survey. Retrieved March 3, 2024, from https://economy-finance.ec.europa.eu/economic-forecast-and-surveys/business-and-consumer-surveys/download-business-and-consumer-survey-data/time-series_en#all-surveys

Ganesh, P. (2019, December). Pre-Trained Language Models: Simplified. Retrieved May 3, 2024, from https://towardsdatascience.com/pre-trained-language-models-simplified-b8ec80c62217

Hopp, D. (2022). Economic Nowcasting with Long Short-Term Memory Artificial Neural Networks (LSTM). *Journal of Official Statistics*, *38*(3), 847–873. https://doi.org/10.2478/jos-2022-0037

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023, October). Mistral 7B. https://doi.org/10.48550/ARXIV.2310.06825

Krugman, P. (2024). Is the Vibecession Finally Coming to an End? *New York Times*. Retrieved May 14, 2024, from https://www.nytimes.com/2024/01/22/opinion/biden-trump-vibecession-economy.html

Maccarrone, G., Morelli, G., & Spadaccini, S. (2021). GDP Forecasting: Machine Learning, Linear or Autoregression? *Frontiers in Artificial Intelligence*, *4*, 757864. https://doi.org/10.3389/frai.2021.757864

Madhumita, M. (2023). Generative AI exists because of the transformer. *Financial Times*. Retrieved May 4, 2024, from https://ig.ft.com/generative-ai/

Manikantan, A. (2021, October). Akaike Information Criterion: Model Selection. Retrieved June 2, 2024, from https://medium.com/geekculture/akaike-information-criterion-model-selection-c47df96ee9a8

Masui, T. (2022, January). All You Need to Know about Gradient Boosting Algorithm Part 1. Regression. Retrieved May 26, 2024, from https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502

Newbold, P., Carlson, W. L., & Thorne, B. M. (2020). *Statistics for business and economics* (Ninth edition, global edition). Pearson.

Olah, C. (2015, August). Understanding LSTM Networks. Retrieved April 30, 2024, from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Olsen, L., Ilvonen, A., Hansen, L. A., Mehren, A. v., Sillemann, B. T., Kuusisto, M., Grahn, M., Kuoppamäki, P., Sundén, G., Jullum, F., & Johansen, R. T. (2024, May). *Nordic Outlook* (tech. rep.). Danske Bank Research. Copenhagen. Retrieved May 13, 2024, from https://research.danskebank.com/link/NordicOutlook050324/$file/Nordic%20Outlook_050324.pdf

Pedersen, R. G. (2021, January). *Prognosestyret Elopvarmning baseret på Kunstig Intelligens og Variable Elpriser* (Case No. 351-060). Elforsk. Retrieved May 9, 2024, from https://elforsk.dk/files/media/dokumenter/2021-07/351-060%20PEKIVE%20Slutrapport_v11.pdf

Sa'adah, S., & Wibowo, M. S. (2020). Prediction of Gross Domestic Product (GDP) in Indonesia Using Deep Learning Algorithm. *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 32–36. https://doi.org/10.1109/ISRITI51436.2020.9315519

Scanlon, K. (2022, June). The Vibecession: The Self-Fulfilling Prophecy. Retrieved May 14, 2024, from https://kyla.substack.com/p/the-vibecession-the-self-fulfilling

StatBank Denmark. (2024, March). Demand and supply by transaction, price unit and seasonal adjustment. Retrieved April 5, 2024, from https://www.statistikbanken.dk/20115

Vandeput, N. (2019, May). Forecast KPIs: RMSE, MAE, MAPE & Bias. Retrieved April 30, 2024, from https://medium.com/towards-data-science/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, August). Attention Is All You Need. Retrieved May 3, 2024, from http://arxiv.org/abs/1706.03762

Wells, S. (2023). Generative AI's Energy Problem Today Is Foundational. *IEEE Spectrum*. https://spectrum.ieee.org/ai-energy-consumption

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (Sixth edition, student edition). Cengage Learning.

Yoon, J. (2021). Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach. *Computational Economics*, *57*(1), 247–265. https://doi.org/10.1007/s10614-020-10054-w

Zhou, T., Niu, P., wang xue, x., Sun, L., & Jin, R. (2023). One Fits All: Power General Time Series Analysis by Pretrained LM. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/86c17de05579cde52025f9984e6e2ebb-Paper-Conference.pdf

11 Appendix

A Input analysis

To get a better sense of what data the models are working upon, a presentation of the input data is presented. This presentation helps communicate the attributes of the underlying variables used in the paper, which are relevant when deciding how to model the variables.

A.1 GDP

Figure 1 visualizes the paper's endogenous variable. The variable displays the GDP of Denmark in current prices, in DKK (billions). The variable is not seasonally adjusted.

The GDP contains a significant seasonal aspect, and a clear trend through time. When testing the variable for stationarity in the model building, it became clear, that the variable has an order of integration of 1.

The variable seems to have a structural break in 2008, following the Global Financial Crisis. The variable also seems to become less predictable following year 2020 and the Covid-19 pandemic. As mentioned in subsection 5.2, the growth in nominal GDP is at least partially explained trough rising inflation. The consequence of this is that while the variable might suggest economic growth, the volume of goods and services available for the consumer has not increased proportionally.

A.2 Industry and Consumer Sentiments

The consumer and industry sentiments are monthly questionaries sent to a representative subset of consumers and industry representatives. The questions are presented in Appendix B, and are aggregated as the percentage of positive responses, minus the percentage of negative responses. The possible outcome for each response is therefore between -100 (100% of all respondents respond negatively) and 100 (100% of all respondents respond positively).

The paper uses two scenarios when applying the industry and consumer sentiments. The first scenario applies industry and consumer sentiments as two indicators. This is a single variable each for industry and consumer, which consists of an average questionary response from selected questions. The indicator is a standard noted in European Commission (2024b) and European Commission (2024b).

The indicator for industry is calculated as: $(Q2-Q4+Q5)/3$.

The indicator for consumers is calculated as: $(Q1+Q2+Q4+Q9)/4$

The second scenario, that uses the sentiments, includes the full range of responses from the consumers and industry as separate exogenous variables. This solution means that the models are fed many, and granular data from the sentiments surveys. This is an advantage for any models able to isolate relevant exogenous variables and ignore irrelevant variables. If the model is incapable of this, the granularity of each response as a variable could become simple noise for a model.

B Sentiment Survey Questions

B.1 Industry Survey

How has your production developed over the past 3 months? It has...

+ increased,

= remained unchanged

- decreased

Do you consider your current overall order books to be...?

+ more than sufficient (above normal)

= sufficient (normal for the season)

- not sufficient (below normal)

Do you consider your current export order books to be...?

+ more than sufficient (above normal)

= sufficient (normal for the season)

- not sufficient (below normal)

Do you consider your current stock of finished products to be...?

+ too large (above normal)

= adequate (normal for the season)

- too small (below normal)

How do you expect your production to develop over the next 3 months? It will...

+ increase

= remain unchanged

- decrease

How do you expect your selling prices to change over the next 3 months? They will...

+ increase

= remain unchanged

- decrease

How do you expect your firm's total employment to change over the next 3 months? It will...

+ increase

= remain unchanged

- decrease

B.2 Consumer Survey

How has the financial situation of your household changed over the last 12 months? It has...

+ + got a lot better

+ got a little better

= stayed the same

- got a little worse

- - got a lot worse

N don't know.

How do you expect the financial position of your household to change over the next 12 months? It will...

+ + get a lot better

+ get a little better

= stay the same

- get a little worse

- - get a lot worse

N don't know

How do you think the general economic situation in the country has changed over the past 12 months? It has...

+ + got a lot better

+ got a little better

= stayed the same

- got a little worse

- - got a lot worse

N don't know

How do you expect the general economic situation in this country to develop over the next 12 months? It will...

+ + got a lot better

+ got a little better

= stayed the same

- got a little worse

- - got a lot worse

N don't know

How do you think that consumer prices have developed over the last 12 months? They have...

+ + risen a lot

+ risen moderately

= risen slightly

- stayed about the same

- - fallen

N don't know.

By comparison with the past 12 months, how do you expect that consumer prices will develop in the next 12 months? They will...

+ + increase more rapidly

+ increase at the same rate

= increase at a slower rate

- stay about the same

- - fall

N don't know.

How do you expect the number of people unemployed in this country to change over the next 12 months? The number will...

+ + increase sharply

+ increase slightly

= remain the same

- fall slightly

- - fall sharply

N don't know.

In view of the general economic situation, do you think that now it is the right moment for people to make major purchases such as furniture, electrical/-electronic devices, etc.?

+ + yes, it is the right moment now

= it is neither the right moment nor the wrong moment

- - no, it is not the right moment now

N don't know.

Compared to the past 12 months, do you expect to spend more or less money on major purchases (furniture, electrical/electronic devices, etc.) over the next 12 months? I will spend. . .

+ + much more

+ a little more

= about the same

- a little less

- - much less

N don't know.

In view of the general economic situation, do you think that now is...?

+ + a very good moment to save

+ a fairly good moment to save

- not a good moment to save

- - a very bad moment to save

N don't know.

Over the next 12 months, how likely is it that you save any money?

+ + very likely

+ fairly likely

- not likely

- - not at all likely

N don't know.

Which of these statements best describes the current financial situation of your household?

+ + very likely

+ fairly likely

- not likely

- - not at all likely

N don't know.

C Evaluation Metric Tables

C.1 MAPE Metrics

Forecast Period	Post-Covid	Post-Covid	Post-Covid	Pre-Covid	Pre-Covid	Pre-Covid
Exogenous Variables	All Responses	Indicator	Univariate	All Responses	Indicator	Univariate
SARIMAX	3,40	3,40	3,40	0,95	0,95	0,95
ARIMAX	3,92	3,92	3,92	1,12	1,12	1,12
RNN	4,5	4,51	4,47	3,95	3,86	3,85
LSTM	4,55	4,58	4,57	3,95	3,87	3,88
GB	5,09	4,89	4,65	4,63	4,24	4,03
MistralLLM	18,71	17,19	14,65	3,59	11,19	10,59

Table 2: MAPE Values One-Step Forecast

Forecast Period	Post-Covid	Post-Covid	Post-Covid	Pre-Covid	Pre-Covid	Pre-Covid
Exogenous Variables	All Responses	Indicator	Univariate	All Responses	Indicator	Univariate
LSTM	8,93	7,05	6,55	2,80	3,42	1,67
ARIMAX	7,39	7,39	7,39	2,91	2,91	2,91
SARIMAX	7,58	7,58	7,58	3,04	3,04	3,04
GB	9,09	7,56	4,89	2,64	7,48	2,04
RNN	6,73	9,53	7,04	3,81	3,57	3,48
MistralLLM	11,46	14,48	4,38	2,84	6,52	3,15

Table 3: MAPE Values Multi-Step Forecast

C.2 RMSE Metrics

Forecast Period	Post-Covid	Post-Covid	Post-Covid	Pre-Covid	Pre-Covid	Pre-Covid
Exogenous Variables	All Responses	Indicator	Univariate	All Responses	Indicator	Univariate
SARIMAX	25,85	25,85	25,85	5,98	5,98	5,98
ARIMAX	29,99	29,99	29,99	7,18	7,18	7,18
RNN	34,27	35,13	34,58	22,43	23,48	22,24
LSTM	34,65	35,14	35,05	22,39	22,50	22,52
GB	38,68	35,38	35,68	26,67	25,76	22,98
MistralLLM	154,09	146,58	122,98	24,19	72,72	69,86

Table 4: RMSE Values One-Step Forecast

Forecast Period	Post-Covid	Post-Covid	Post-Covid	Pre-Covid	Pre-Covid	Pre-Covid
Exogenous Variables	All Responses	Indicator	Univariate	All Responses	Indicator	Univariate
LSTM	73,27	57,85	53,60	19,09	21,87	12,19
ARIMAX	61,80	61,80	61,80	18,47	18,47	18,47
GB	70,52	57,54	39,03	19,71	45,77	12,22
SARIMAX	63,24	63,24	63,24	18,85	18,85	18,85
RNN	53,87	78,06	57,69	24,86	23,66	21,90
MistralLLM	93,49	123,46	35,54	18,62	43,78	20,80

Table 5: RMSE Values Multi-Step Forecast