Kaunas University of Technology

Faculty of Mathematics and Natural Sciences

Lithuanian Consumer Price Index Forecasting Using Transformer Models

Masters's Final Degree Project

Laurynas Grušas

Project Author

Assoc. Prof. Dr. Kristina Šutienė

Supervisor

Assoc. Prof. Dr. Kristina Kundelienė

Supervisor

Kaunas, 2024

Kaunas University of Technology

Faculty of Mathematics and Natural Sciences

Lithuanian Consumer Price Index Forecasting Using Transformer Models

Masters's Final Degree Project

Business Big Data Analytics (6213AX001)

Laurynas Grušas
Project author

Assoc. Prof. Dr. Kristina Šutienė
Supervisor

Assoc. Prof. Dr. Kristina Kundelienė
Supervisor

Assoc. Prof. Dr. Mindaugas Kavaliauskas
Reviewer

Prof. Dr. Lina Dagilienė
Reviewer

Kaunas, 2024

Kaunas University of Technology

Faculty of Mathematics and Natural Sciences

Laurynas Grušas

Lithuanian Consumer Price Index Forecasting Using Transformer Models

Declaration of Academic Integrity

I confirm the following:

1. I have prepared the final degree project independently and honestly without any violations of the copyrights or other rights of others, following the provisions of the Law on Copyrights and Related Rights of the Republic of Lithuania, the Regulations on the Management and Transfer of Intellectual Property of Kaunas University of Technology (hereinafter – University) and the ethical requirements stipulated by the Code of Academic Ethics of the University;

2. All the data and research results provided in the final degree project are correct and obtained legally; none of the parts of this project are plagiarised from any printed or electronic sources; all the quotations and references provided in the text of the final degree project are indicated in the list of references;

3. I have not paid anyone any monetary funds for the final degree project or the parts thereof unless required by the law;

4. I understand that in the case of any discovery of the fact of dishonesty or violation of any rights of others, the academic penalties will be imposed on me under the procedure applied at the University; I will be expelled from the University and my final degree project can be submitted to the Office of the Ombudsperson for Academic Ethics and Procedures in the examination of a possible violation of academic ethics.

Laurynas Grušas

Confirmed electronically

Grušas Laurynas. Lithuanian Consumer Price Index Forecasting Using Transformer Models. Master's Final Degree Project/ supervisors Assoc. Prof. Dr. Kristina Šutienė and Assoc. Prof. Dr. Kristina Kundelienė; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Study field and area (study field group): Mathematics, Applied mathematics.

Keywords: consumer price index, transformers, time series forecasting, neural networks, harmonised index of consumer prices.

Kaunas, 2024. 59 pages.

Summary

This master thesis aims to improve Consumer Price Index forecasts. Forecasting accurate price levels is important for governments in formulating monetary policy, such as indexing wages, social benefits, or commercial contracts. While traditionally univariate ARIMA models are used to forecast price indices, we propose transformer models. Harmonised Index of Consumer Prices (HICP) for Lithuania is forecasted using monthly HICP rates from 10 European countries together with unemployment rates, Producer Price Index in Lithuania, two dummy variables, and feature engineered variables for the period 1998-2024. Different types of transformer models are tested against linear models (ARIMA, VAR and ARDL) both in univariate and multivariate settings, with forecasting horizons set to one, three and six months. The results show that linear models are more accurate in short-term (1-3 months) forecasts, whereas transformers are more effective in long-term forecasts. Moreover, multivariate models defeat univariate ones, suggesting a demand for further research on covariates impacting European price indices. In addition, it was discovered that multi-head attention layer, a crucial part in the transformer architecture, could be leveraged as an explanatory tool in multivariate forecasts.

Grušas Laurynas. Lietuvos vartotojų kainų indekso prognozavimas taikant transformerių modelius. Magistro studijų baigiamasis projektas / vadovė Doc. Dr. Kristina Šutienė / vadovė Doc. Dr. Kristina Kundelienė; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Matematikos mokslai, taikomoji matematika.

Reikšminiai žodžiai: vartotojų kainų indeksas, transformeriai, laiko eilučių prognozavimas, neuroniniai tinklai, suderintas vartotojų kainų indeksas.

Kaunas, 2024. 59 p.

Santrauka

Šis magistro studijų baigiamasis projektas siekia pagerinti Vartotojų Kainų Indekso (VKI) prognozes. Šis indeksas yra svarbus indikatorius, naudojamas monetarinėje politikoje, pavyzdžiui, indeksuojant atlyginimus, socialines pašalpas ar komercines sutartis. Nors tradiciškai vienmačiai ARIMA modeliai būdavo naudojami VKI prognozėms, šiame projekte siūlome prognozavimui pasitelkti transformerių modelius. Lietuvos Suderinto Vartotojų Kainų Indeksui (SVKI) prognozuoti pasitelkiami 1998-2024 metų laikotarpio SVKI duomenys iš 10 Europos valstybių, kartu su duomenimis apie nedarbą, Gamintojų parduotos pramonės produkcijos kainų indeksą (angl. *Producer Price Index*) Lietuvoje, dviem binariniais ir papildomais laiko kintamaisiais. Tyrimo metu vairių tipų transformerių modeliai lyginami su klasikiniais modeliais (ARIMA, VAR, ARDL) vienmačių ir daugiamačių modelių atvejais, kai prognozės vykdomos vieną, tris ir šešis mėnesius į priekį. Rezultatai parodė, jog klasikinių tiesinių modelių prognozės yra tikslesnės trumpo laikotarpio, t.y. vieno ir trijų mėnesių, prognozėms. Tuo tarpu transformerių modeliai tiksliau prognozuoja ilguoju laikotarpiu. Be to, daugiamačiai modeliai buvo tikslesni negu vienmačiai. Dėl šios priežasties kyla poreikis ištirti faktorius, veikiančius vartotojų kainų indeksus Lietuvoje ir Europoje. Galiausiai, transformerių modelių architektūroje naudojamas dėmesio mechanizmas (angl. *attention layer*) gali būti patogus įrankis, siekiant paaiškinti daugiamačių modelių prognozes ir naudotų kintamųjų svarbą.

Table of contents

List of figures

List of tables

List of abbreviations and terms

Abbreviations:

ADF - Augmented Dickey-Fuller test;
CNN - Convolutional Neural Networks;
CPI - Consumer Price Index;
HICP - Harmonised Index of Consumer Price;
PPI - Price Producer Index;
RNN - Recurrent Neural Network;
TFT - Temporal Fusion Transformer;

Introduction

In April, 2024, Ben Bernanke, the former chair of the US Federal Reserve, announced that The Bank of England "needs to open its ears, its mind and its wallet" [39]. His remark came in the conclusion of a 10-month review of the Bank of England COMPASS forecasting platform. The Bank's data tools became out of date after failing to predict 11% inflation surge in the UK in 2021-2022, following the Covid-19 pandemic and Russia's invasion in Ukraine. When inflation rates plummeted over 2023, the Bank failed once again to predict decreasing inflation, thereby keeping quantitative easing policies out of control. However, central banks around the world followed the same path. While the Bank of England will have to restructurize and make investments in updating the infrastructure and research capabilities, it is clear that there is a huge demand today to make use of new time series models to forecast economic indicators.

In search of new forecasting methods, we turn our attention to transformer, one of the state-of-the-art machine learning models. The transformer model was introduced in 2017 in the famous article "Attention is all you need" by Google researchers. The model is prominent in Natural Language Processing tasks, such as word prediction in a sequence of text (language modelling), summarization, and question answering. One of the best use cases of transformer models is the famous generative AI tool Chat GPT (generative pre-trained transformer). What distinguishes transformers from other deep learning models is the self-attention mechanism, where the model is able to weigh the importance of each word in the sequence to predict the next word in the sequence [18].

This master thesis is a link between two scientific fields: economics and machine learning. With the final aim to discover new methods to improve economic forecasting, this thesis will compare traditional time series models with transformer models in forecasting the Harmonised Consumer Price Index, an indicator used to calculate the inflation rate. Here, it is assumed that time series forecasting is similar to text prediction, where the model pays different attention to data points in the past. In the first chapter, the current literature on Consumer Price Index (CPI) forecasting will be reviewed. Then in the second chapter data, forecasting diagnostics, linear and machine learning models for forecasting used in the research will be described. Next, classical forecasting models will be compared with transformers and their alternatives. The thesis will be concluded with research findings and discussion.

1. Literature review

1.1. Economic forecasting

Economic forecasting originated in the aftermath of the 1930s economic shocks, caused by the Great Depression. According to [8], economic forecasting flourished under two circumstances. One is that economic relations are intricate and that the consequences of actions could take time to expose. Another dictates that the future is uncertain, but forecasts can help us deal with it better. Carnot, Koen, and Tissot explain the rise of forecasting methods. Firstly, as economic theories, such as Keynesian, were discussed and adapted in political agendas, researchers needed to find methods to validate them. Therefore, using statistical data and forecasts resembled conducting experiments in fundamental sciences to prove whether or not economic theories are reliable. Secondly, policymakers needed plausible and, preferably, quantitative tools to assess the risks and costs of political decisions in aggregate terms in addition to plotting them in scenarios. Thirdly, it turned out that economic forecasting could benefit not only the public, but also the private sector. For example, companies need to weigh the consequences before making investment decisions.

Forecasts impact agents (governments or businesses) in two ways. First, agents can adapt, i.e. take opportunity to reduce the consequences of events. For example, reacting to lower demand forecasts, a company can reduce production volumes and adjust its budgets accordingly. Second, agents can retroact or take action to influence the forecasted events. For instance, policymakers, such as central banks, can alter interest rates in response to inflation forecasts.

As such, the economic literature defines 4 types of forecasting horizons:

- very short term (up to two quarters)
- short term (6 months – 2 years)
- medium term (2 – 5 years)
- long term (> 5 years)

While certain ranges can differ, the very near-term forecasts are most widely used. This is because of short-term economic fluctuations; for example, the rise in unemployment can rapidly result in social problems. Similarly, short-term economic downturns can hinge future investment and economic growth due to "lost ground". This occurs when a country could lose attractiveness for investments due to financial crisis. Meanwhile, the said country will need more time to reinstate its attractiveness and catch up with other countries. Finally, short-term forecasts are necessary for regulating stable fiscal and monetary policy. Some examples of such regulations include budget preparation [8], material planning, or scheduling [24]. The government rarely uses medium- and long-term projections. The reason behind this is that economies might experience structural changes over the long term, which

are difficult to predict. Use cases usually involve evaluations of investment projects. Likewise, from the point of view of the time series method, although models can be used in multiple forecasting horizons, they accurately capture only short-term movements, since they "lack the economic spine needed when thinking about medium- and long-run developments"[8, pg. 85].

1.2. Consumer Price Index

Consumer Price Index (CPI) is, as defined by [13], "the change over time in the prices of consumer goods and services acquired, used, or paid for by households". The history of the index dates back to World War I, when the Bureau of Labor Statistics of the US government began collecting prices on a consumer level [48]. From 1917 to 1919 the Bureau collected data in 92 population centres in order to create weighted indexes of expenditures. During the Great Depression and rationing in World War II the structure of index altered so as to reflect changes in consumer expenses.

According to Eurostat, the CPI is calculated by collecting the prices of the goods and services purchased by the population in each country. The purpose of the CPI calculation, as explained by [47], is to assess the rate of change in cost of living for consumers. The cost of living defines preferences for goods and services (i.e., housing, food, healthcare, transportation) consumed by each individual at a certain point of time or in the near future. Since it is impossible to practically collect information about each consumer's consumption patterns, the cost-of-living index is realised by measuring price changes for a set of goods and services acquired on a market within a certain range of time. As per [48], the inflation rate is derived as a percentage change in CPI by the formula below:

$$R_{inf} = \frac{100 * (I_{CP} - I_{PP})}{I_{PP}}, \tag{1}$$

here R_{inf} is the inflation rate, I_{CP} is the current index value, and I_{PP} is the previous index value.
According to [48], there are three trends of price changes. **Inflation** refers to moderate increase, while **disinflation** means a slowing inflation rate. **Deflation** occurs when price levels decrease. Both high inflation and deflation are harmful to the economy. In case of the latter, consumers are more likely to delay purchases in anticipation of even lower prices in the future. Without consumer purchases, companies are forced to cut down on production levels or even lay off workers. Consequently, economic growth slows or even decreases. [12] grades a 3% increase in CPI as moderate inflation, whereas a price increase of 5% is regarded as strong inflation.

According to the economics literature ([42]), changes in price levels are essentially explained by the Phillips curve. In 1958 economist A. W. Phillips published a research paper in which he analysed the relationship between unemployment and the money wage rates in the UK over 1861-1957. His results proved an inverse relationship between the price dynamics and the unemployment rate. In fact, the Phillips curve states that as the level of unemployment increases, the rate of change in price growth decreases in general, and vice versa. As a result, policymakers are recommended to maintain the unemployment rate between 4% and 6%. However, as [36] describes, the data subsequent to the 1960s showed changes in understanding the Phillips curve. The modern approach states that the

public expectations of the would-be inflation need to be taken into consideration. According to the Accelerationist Hypothesis, predominantly shaped by Milton Friedman, the policymakers' efforts to sustain the unemployment rate at a certain level will lead to a continually increasing inflation.

The aim of Consumer Price Index is to measure price level changes for consumer goods and services. [35] lists two methods to collect data. One way is to record all transactions of goods and services purchased by consumers within a time period, let's say a month. However, this methods is costly, time consuming, and impractical, due to a large number of goods and services purchased within the period by all consumers. An alternative and less complicated method is to take samples of goods and services with samples of transactions from samples of locations. Therefore, the collection of CPI data is organised by sampling the goods and services most frequently bought or consumed into consumer baskets, such as food and beverages, hygiene products, newspapers; housing, energy, and utility costs; expenses on health, transport, education, communications, restaurants, and hotels. The service and good categories selected for the calculation of the index meet the COICOP (Classification of Individual Consumption by Purpose) classification standards developed by the United Nations Statistics Division [28]. The composition of the consumer basket is constantly changing, as new goods and services are included and excluded from the basket to reflect changing consumer tastes. For example, in 2010, the UK's consumer basket removed hair dryers and added hair straighteners [35].

In Lithuania, data for CPIs are collected and calculated by the State Data Agency of Lithuania. The indices are used by different government bodies, such as the ministries of Finance, Economy and Innovation, Social Security and Labour, the Lithuanian Central Bank, as well as international organisations (Eurostat, OECD) [28]. Due to its importance in economic regulation, the CPI is subjected to a careful examination each time before being published in government statistics reports.

As [41] argues, the demand for accurate CPI forecasts comes from both public and private agents. For private agents, such as households, decisions on savings, investments, and purchases depend on the perception of purchasing power in the future, which depends on their expectations of inflationary trends. For the public sector, the CPI is one of the key indicators in formulating monetary policy. Central banks update interest rates based on inflationary trends and set inflation targets. For instance, a sudden increase in CPI in 1993 led to the US Federal Reserve raising interest rates. The inflationary surge flattened out quickly, so the FED soon reversed its decision [47]. In addition, CPI is adopted in accounting, as well as indexing commercial contracts, wages, social benefits, and other financial instruments. Likewise, CPI helps to monitor changes in household consumption patterns and perform cross-country economic comparisons.

The consumer baskets used to calculate the CPI differ from one country to another. In Europe, there have been attempts since mid-1970s to uniform CPI calculation approaches. However, they were unsuccessful, as countries were not willing or able by their own laws to adapt their own methodologies for such an important indicator with the aim of allowing questionable international comparison [14]. However, the introduction of the Maastricht Treaty in 1992, where the criteria to join the euro zone were laid out, made it crucial to establish an agreed CPI approach among the Member States. Therefore, in 1995 the European Union's Council of Ministers adopted a new regulation establishing

uniformed methodologies for harmonised indices of consumer prices. While interim indices were compiled in 1996, since 1997 the official Harmonised Index of Consumer Prices (HICP) have been compiled. Today, the European Central Bank uses it to track official inflation convergence in the euro zone [13] and as an official indicator under the Maastricht criteria for countries accessing the euro zone. HICP is adopted by national central banks in Europe as well as international organisations (OECD, IMF). Based on COICOP classification, the index consists of twelve components: all-items (or total); food and non-alcoholic beverages; alcoholic beverages and tobacco; clothing and footwear; housing, water, electricity, gas and other fuels; furnishings, household equipment and routine maintenance of the house; health; transport; communication; recreation and culture; education; restaurants and hotels; miscellaneous goods and services. However, when predicting price levels, in general, forecasts are made at the aggregate level (all items) [41].

In Lithuania, HICP has been calculated since 1996 [27] on a monthly basis. The index methodology was updated in 2005 and 2016 in order to comply with the EU regulations and the revised methodology by the Eurostat agency. Whereas data for the index calculation have been collected by price collectors from 28 retail outlets in Lithuania since the beginning, starting from 2023 data collection also relies on scanner data from major retail outlets in Lithuania. The indices for each COICOP classification category are then weighed based on the initial revised weights. These weights reflect the structure and trends in consumer expenses in Lithuania. The scope of HICP index includes not only resident households, but also nonpermanent residents in Lithuania and visitors from abroad, irrespective of their citizenship and nationality. On the other hand, HICP does not have data on expenses of Lithuanian residents abroad [27].

1.3. Modelling Consumer Price Index

1.3.1. Data preprocessing

When forecasting the CPI, the shape of the data becomes an important choice in the modelling part [1]. As common in financial forecasting, the first differences of the index logarithm (in our case, the CPI) are usually calculated following the formula:

$$y_t = \log(Y_t) - \log(Y_{t-1}), \tag{2}$$

here Y_t and $Y_t - 1$ denote the Consumer Price Index at times t and $t-1$ after calculating first differences and their logarithmic transformations.

According to the same authors, the rationale behind taking the first differences is to ensure stationary time series, a common requirement in forecasting models. What is interesting is that, by taking the first differences in monthly CPIs, the time series marks the change of growth in monthly CPI or, in other words, the inflation rate. Naturally, the forecasting of the CPI rate becomes the forecasting of inflation rate. The first differences are also used by [46] to obtain stationary data for the monthly Chinese CPI forecasts.

A similar approach to the use of stationary transformation (first differences) data was adopted by

[30] in the inflation forecasting model using neural networks. They used the monthly and seasonally adjusted CPI and Producer Price Index (PPI) together with indices for food, energy and service prices. Their forecasts focused not only on the US market (data provided by the Federal Reserve database), but included the euro zone (Germany, France, Italy, and Spain), for which data was employed from the European Central Bank and Japan (data used from the OECD database). However, unlike Álvarez-Díaz and Gupta, McNelis and McAdam propose scaled data for estimation. They argue that scaling helps avoid underflow or overflow problems related to computer memory. In essence, the neural network input or output could be unexpectedly a very small or high value (an outlier), which leads to computer returning zero value or stopping. Therefore, McNelis and McAdam describe and adopt three different scaling functions for their model estimation:

1. In linear scaling (also called min max normalization), all values are transformed into $[0,1]$ or $[-1,1]$ ranges by using minimum and maximum values of data series. For instance, the $[0,1]$ scaling is performed using the following function:

$$f(x) = \frac{x - min(x)}{max(x) - min(x)}, \tag{3}$$

   here x refers to a value to be scaled, $min(x)$, $max(x)$ – the minimum and maximum values in the series. This scaling method was also adopted by [49].

2. Based on the non-linear scaling method of Helge Petersohn, the series are also transformed into the range $[0,1]$ according to the formula:

$$z_{k,t} = \left(1 + \exp\left[\left(\frac{\ln\left(z_k^{-1} - 1\right) - \ln\left(z_k^{-1} - 1\right)}{max(x_k) - min(x_k)}\right)\left[x_{k,t} - min(x_k)\right] + \ln\left(z_k^{-1} - 1\right)^{-1}\right]\right)^{-1} \tag{4}$$

   here $z_{k,t}$ refers to the standardised value of the original data point at time t for variable k, $x_k t$ to the original data point at time t for the variable k; $min(x_k)$, $max(x_k)$ – the minimum and maximum values in the series k.

3. Dayhoff and De Leo transformation first standardises x to z and then takes the log-sigmoid transformation of z:

$$z = \frac{x - \bar{x}}{\sigma} \tag{5}$$

   here z refers to the standardised value, x to the original data point before transformation, $\bar{x}$ to the mean of the original time series, σ to the standard deviation of the original time series

$$x^* = \frac{1}{1 + \exp^{-z}} \tag{6}$$

   here x^* refers to the transformed value after applying the log-sigmoid function, z to the standardised value. [5] suggests that the chosen scaling method should approximate the activation function. For instance, if scaling $[-1;1]$ is used, then tanh activation function should be chosen. For $[0;1]$ scaling, sigmoid function should be used.

When it comes to the size of the data set, [1] used 408 time series observation points to train different forecasting models. The dataset was split into 3 periods: the first 288 observations (70%) for training, the next 79 ones (20%) for validation (both parts are referred to by the authors as the in-sample period) and the last 41 ones (10%) for testing (the so-called out-of-sample period). On the other hand, some authors, namely [49], used only training-test parts split into 70:30 proportions in their model training.

1.3.2. Consumer Price Index forecasts in the past

An interesting discovery in the scientific literature about CPI data used in modelling refers to the choice of univariate vs. multivariate models. As [1] argue, traditionally Consumer Price Indices are forecasted using univariate models. One of the reasons why researchers have been struggling with multivariate models is the reliance on inflationary dynamics in monetary policy. For instance, the inflation rate in the USA was highly volatile during the 1970s due to the expansionary monetary policy pursued by the Federal Reserve. However, even when using CPI time series data in forecasting, training data consider only periods characterised by stable monetary and financial policies. To avoid volatile periods when forecasting the CPI in the US, Álvarez-Díaz and Gupta use monthly data ranging from 1980 to 2013. This period marks the debut of Volcker stabilisation monetary policy, which is known for a lower inflation rate.

Another reason for choosing univariate models for CPI modelling is the forecast accuracy. According to [42], multivariate models with monetary variables (energy or housing prices, real GDP) had not exceeded a higher accuracy than the random walk model by Atkeson-Ohanian in 2001 (however, this one failed to capture the inflation dynamics of the 2000s in the US) or the time-varying unobserved components model by Stock and Watson in 2007. [42] propose a new measure, the unemployment recession gap, to forecast long-term inflation. Although the proposed variable seems to catch the inflationary dynamics, the authors suggest using it with caution.

So far, most of the reviewed authors [1, 30, 42, 46] have used univariate approaches to forecast the CPI (or inflation rate). However, there are instances in the scientific literature of using multivariate models to predict CPI. [49] adopted 34 staple food prices from 38 cities in East Java to estimate the Indonesian CPI. Their choice of multivariate model was argued by the high fluctuation of food prices in the province. Similarly, [33] used crude oil prices (OP), the world gold price (GP), and the effective rate of federal funds (FFER) to forecast the US CPI. Having used the US CPI data from 2017 to 2022, their Multivariate Adaptive Regression Splines (MARS) model achieved a 97. 6% R2 score and 0.004 MAPE for the test dataset and defeated other models used in the study: Multivariate Linear Regression (MLR), Support Vector Regression (SVR), and Autoregressive Distributed Lag (ARDL). The authors' choice of using these three explanatory variables was based on scientific literature on the factors affecting US CPI. The advantage of using a multivariate model is that the impact of each predictor on the US CPI could be analysed using the SHAP tool. However, there are very few studies that analyse the factors that influence CPI in European countries. Existing studies [23] on European CPI forecasts included univariate rather than multivariate approaches.

In CPI time-series modelling, ARIMA short-term forecast models have demonstrated remarkable

results so far. [12] model for monthly Chinese CPI predictions in the period from January 2000 to December 2008 has achieved a MAPE of 1.17%. The Multivariate Adaptive Regression Splines (MARS) model by [33] achieved 0.933 MAPE on the test dataset. However, in [11] study SARIMA model outperformed the Multivariate Linear Regression model; [38] came to the same conclusion when forecasting Columbian CPI with the ARIMA model against the Exponential Smoothing, Holt-Winters and Singular Spectrum Analysis models.

The literature revealed that machine learning algorithms have already been used in CPI forecasts. So far, [49] used Long Short Term Memory (LSTM) algorithm to predict monthly CPI based on 34 staple food prices in Indonesia. Out of 7 optimisation algorithms used in the LSTM model (Stochastic Gradient Descent, Root Mean Square Propagation, Adaptive Gradient, Adaptive moment (Adam), Adadelta, Nesterov Adam (Nadam) and Adamax), Nadam was found to return the lowest RMSE (4.088). [49] recommend different variations of epoch, hidden layer, batch size, and input variables to improve the accuracy of the forecast. In fact, LSTM algorithms derive from Recurrent Neural Netwok (RNN) and are defined by 3 types of gates (input, forget, and output) used in the architecture. According to [25], the LSTM algorithms solve the limitations of RNN models. The latter suffer from exploding and vanishing gradient problems occurring during the training phase due to capturing information across long temporal distance. Meanwhile, LSTM algorithms select which long-range information to store and which to forget through the gate functionality.

It is not uncommon in HICP forecasting models to employ information on tax changes. As [41] argues, the impact of changes in the tax system is modelled with the help of dummy variables.

When it comes to forecasting Lithuania's HICP, most forecasting attempts were conducted in a small number of bachelor or master theses. The remaining theses were published more than 10 years ago. Among them worth mentioning is the [21] investigation on factors with the greatest impact on Lithuania's HICP in 2011-2012. Using regression analysis, he discovered that three COICOP categories contributed the most to inflation for the researched period: housing, energy, and utility costs; food and beverages, as well as transport. He also found that the inflation from a previous impact present HICP prices for the mentioned categories at most. In other words, the inflation of food prices at the moment will impact food prices in the future, thus confirming the adaptive expectations theory. Likewise, he found out that the inflation of food prices in Lithuania relies on world food prices, such as meat, oil, grains, milk, beverages, fish, fertilisers, as well as industrial production volumes and consumer loan interest rates. Another outstanding study by [29] investigates the impact of the recent Russian invasion of Ukraine on Lithuania's HICP. The author trained two seasonal ARIMA models: one with data prior to the beginning of the invasion in February 2022, and another with data up to March 2023. The forecasts for both models were made up to December 2023. The first model resulted in the failure to capture the shock in Lithuania's HICP after the invasion over the entire forecasting period. In detail, the highest error between the real and predicted HICP was observed in September 2022, when the first model underestimated Lithuania's HICP even by 17%. This finding suggests including information about the invasion in future Lithuania's HICP forecasting models.

Other articles on Lithuania's HCIP forecasting have been published by academics or institutions. [41] used data from 1996 to 2013 to forecast disaggregated HCIPs for the following categories: un-

processed food, processed food, non-energy industrial goods, energy and services. As exogenous variables, he used Lithuania's import deflator, labor costs (average gross monthly wage, total compensation to employees), raw commodity prices (wheat and milk powder prices in the EU, indices of coffee, vegetable oil, meat, Brent oil prices as well as Australian coal price), LTL/USD exchange rate together with price indices of 92 items. There were several types of forecasting models used: 12-month Moving Average (MA), Exponential Smoothing, ARMA, State Space Model, Factor Augmented Autoregressive (FAAR) model, Vector Autoregression (VAR) model and the author's proposed Mark-up model, consisting of univariate equations. The forecasts were made for 3, 6, 9, 12 and 15-month horizons. The mark-up model proved to produce the most accurate forecasts. Furthermore, [31] used Vector Autoregression (VAR) models to forecast each of the 12 HCIP components of Lithuania. In addition to the components themselves, the authors employed world oil prices in LTL per barrel as exogenous variables. Data were derived for the period January, 2002, - December, 2007 and 4 lags were used in the model. However, the authors recommend including more variables in future models.

Table 1. Reviewed authors and models used to forecast Consumer Price Index.

Authors	**Forecasting models**
[1]	Random Walk (RW), Autoregressive (AR), Seasonal Autoregressive Integrated Moving Average (SARIMA), Artificial Neural Networks (ANN), Genetic Programming (GP)
[3]	Recurrent Neural Networks (RNNs), Autoregression (AR), Vector Autoregression (VAR), Random Walk (RW), Logistic Smooth Transition Autoregressive Model (LSTAR), Random Forests (RF), Gradient Boosted Trees (GBT), Artificial Neural network (ANN), Deep Neural Network (Deep-NN)
[23]	Autoregressive (AR), Artificial Neural Networks (ANN)
[30]	Artificial Neural Networks (ANN)
[42]	Unobserved Components-Stochastic Volatility (UC-SV), Autoregressive Distributed Lag (ARDL)
[46]	Autoregressive Integrated Moving Average (ARIMA)
[49]	Long Short Term Memory Neural Networks (LSTM)
[11]	SARIMA, Multiregression Model (MLR)
[38]	Autoregressive Integrated Moving Average (ARIMA), Holt-Winters, Exponential Smoothing, Singular Spectrum Analysis
[17]	Multilayered Perceptron
[32]	Grey Systems modelling, Autoregressive Integrated Moving Average (ARIMA)
[33]	Autoregressive, Random Walk, Autoregressive Distributed Lag (ARDL), VAR, k-NN, Artificial Neural Networks (ANN), Support Vector Regression (SVR)
[41]	12-month Moving Average model, Exponential Smoothing, ARMA, State Space model, Factor Augmented Autoregressive model (FAAR), Vector Autoregression (VAR), Mark-Up model
[31]	Vector Autoregression (VAR)
[29]	SARIMA

Table 2. Reviewed authors and variables used in forecasting models.

Authors	Variables
[23]	Finnish CPI from 1960 to 2009
[17]	CPI of G-7 countries (USA, UK, France, Germany, Italy, Japan, Canada) from January, 1996 to February 2013
[38]	54 monthly Colombian food components for CPI on food from January 1999 to October 2012
[30]	CPI, PPI, indices for food, energy and services for inflation rates in the USA, Japan and the euro area (Germany, France, Italy and Spain)
[1]	Monthly CPI of the USA from January 1980 to December 2013.
[11]	Albanian gross domestic product, unemployment rate, exchange rate, the number of Albanian people traveling abroad quarterly from January 1994 to December 2017
[3]	The US consumer price index from January 1994 until March 2019
[32]	Vietnamese CPI, Raw Materials Price (RMP), Gold Price (GP), Dollar Price (DP) from 2005 to 2013
[42]	Unemployment recession gap quarterly from 1959 to 2010
[33]	CPI, crude oil price, world gold price, federal fund effective rate (FFER) data from January 2017 to February 2022
[12]	Chinese CPI from 2000 to 2009
[49]	34 types of staple food prices from 2014 to 2018
[41]	Import deflator, variables describing labour costs, international commodity prices, EU food commodity prices world food prices
[31]	12 HICP groups and world oil price (in LTL per barrel) for January 2002 – December 2007
[46]	Chinese CPI from 2000 to 2009

1.4. Key findings

The table 1 summarises reviewed authors and the models they used to forecast the CPI. It shows that autoregressive linear models have been dominating research on CPI forecasting. In addition to classical linear models, neural networks have already been used in CPI forecasting.

The table 2 summarises the variables used in the univariate and multivariate CPI forecasting models by the reviewed authors. While national consumer prices indices are unsurprisingly the most often used variables, other price indices, such as the Price Producer Index (PPI) or commodities (food, energy, gold), or economic indicators (unemployment rate) have also been employed as covariates in multivariate models.

In summary, the literature review revealed that, firstly, in the European context researchers often use Harmonised Index of Consumer Prices (HICP) in forecasting models, since these indicators allow cross-country comparison owing to an agreed methodology. Secondly, univariate models are preferred to multivariate ones, since monetary variables often fail to capture variation in inflation. Similarly, the literature revealed a lack of studies on factors impacting European Consumer Price Indices. Thirdly,

there are few studies on Lithuania's HCIP and most works are published by undergraduate or graduate students. Fourthly, autoregressive integrated moving average (ARIMA) models have achieved the highest accuracy, compared to other linear and machine learning models. So far, among other deep learning models, transformers have not been used in CPI forecasting. This identifies an opportunity in this thesis to leverage transformer models. Nevertheless, due to the scope of this thesis, the focus here will be not on exploring indicators explaining variation in CPIs, but rather on comparing traditional forecasting models with transformer alternatives.

As such, the **research goal** is to investigate the effectiveness of transformer-based models to forecast Lithuania's Harmonised Index of Consumer Prices (HICP) compared to traditional time series forecasting models.

To accomplish this goal, the following **research tasks** will be carried out:

1. Conduct a review of classical time series forecasting models and the application of machine learning models in time series analysis.
2. Gain an understanding of transformer architecture and identify its benefits in time series forecasting.
3. Build a transformer-based model to forecast Lithuania's Harmonised Consumer Price Index model and compare its performance with established time series models.

2. Methodology

2.1. Diagnostics

Various statistics and statistical tests are used to assess CPI time series data or its transformations before the modelling part. [1] estimated partial autocorrelation coefficients (PACF) together with the Ljung-Box test on the US CPI monthly data. Both methods indicated a strong dependence on the 1st and 12th lags. In addition, the Jarque-Bera test was used to indicate whether the time series is normally distributed. The null hypothesis was rejected at 1% significance level. Similarly, non-parametric Dickey-Fuller (ADF) and parametric Phillips-Perron tests for CPI time series stationarity were conducted. It was found that monthly CPI series were nonstationary, while the differenced series (i.e. the inflation rate) were stationary. This evidence only proves the choice to apply the first differences for the original time series data. [1] also applied the Brock-Decker-Scheinkman (or BDS) test for nonlinear patterns in the data. The test was used on the residuals captured by the GARCH(1,1) model. While the results did not allow us to reject the null hypothesis about independent and identically distributed (i.i.d.) data, the researchers discovered some evidence in favour of nonlinearity in data. As a result, the authors recommended using nonlinear CPI forecasting methods, such as neural networks.

2.1.1. Stationarity

Having stationary data is not only a prerequisite for ARIMA, VAR and ARDL models, but also an effective training technique for time series forecasting neural network models. As [22] argues, variable transformation by differencing is a method to achieve stationary time series. Unit root tests test for the presence of unit roots, which indicate nonstationary time series. Augmented Dickey-Fuller (ADF) test is one of the most practically applied statistical tests for stationarity. It tests the null hypothesis:

H_0: There is a presence of a unit root;
against the alternative hypothesis:

H_A: There is no unit root present (the time series is stationary);

The null hypothesis is rejected if the t-test value is lower than the critical value and vice versa. In fact, [22] describe three test specifications: for stationarity (neither intercept nor trend included), level stationarity (only intercept included), or trend stationarity (both intercept and trend included).

2.1.2. Evaluation criteria

[1] discussed the evaluation criteria for different forecasting models. Together with [46], they recommend using the mean absolute percentage error (MAPE) to compare the models among themselves. The measure is calculated using the following formula:

$$\text{MAPE} = \frac{\sum_{t=1}^{T} |y_t - \hat{y}_t|}{T} \times 100, \tag{7}$$

here y_t stands for the actual CPI value, $\hat{y}_t$ for the forecasted value, T for the sample size.

That is to say, higher MAPE values indicate higher model forecasting error, and vice versa, lower MAPE values indicate lower forecasting capability. The authors point out that the errors must be predicted on the original CPI values rather than on the transformed (inflation rate) time series. The choice of MAPE to assess competing models has several arguments. First, it is more resilient to outliers than squared measure methods (MSE, RMSE). Second, MAPE is easy to calculate and interpret. Third, by calculating the percentage difference of actual and predicted values of a variable, the measure is irrespective of the variable scale.

On the other hand, this evaluation criterion is criticised by [16]. When the figures are small or close to zero, the absolute percentage error becomes large. In the given example by [16], the prediction of 6 units for the actual value equal to 2 units has a 200% MAPE error. Therefore, it is recommended to use SMAPE (symmetric mean absolute percentage error) criterion instead. It takes the average between the actual and predicted values. The error is calculated then as a per cent value of this average given the formula below:

$$\text{SMAPE} = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{2|A_i - F_i|}{|A_i| + |F_i|} \cdot 100\%, \tag{8}$$

where A_i refers to actual values, F_i - forecasted values, N - the number of observations.

Some other authors [23, 30, 49] used the popular RMSE score as an evaluation criterion. However, the authors did not motivate their choices. This error measure is calculated using the formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - Y)^2}{n}}, \tag{9}$$

here Y_i and Y refer to predicted and actual values, respectively. Squared errors are summed up and divided by the number of forecasts. Taking the squared root allows for an easier interpretation of the measure [16].

Lastly, R-squared has also been adopted as a precision measure by some researchers [33]. This statistical measure verifies how much of the variance in the outcome variable is explained by the predictor(s). Naturally, R-squared is measured on the scale $[0, 1]$.

2.2. Time series forecasting

Time series forecasting is defined as "taking [..] models, training them on historical time series data, and consuming them to forecast future predictions "[24, pg. 5]. When it comes to forecasting, the scientific literature [see 8, 24] defines two types of models:

- **Univariate models**, where a sequentially recorded single feature is observed through time and the forecast depends only on its past values. Examples of such models include simple naïve methods, for instance, the next value is equal to the past one, or more complex methods, such as ARIMA models. These types of models, as a result, do not explain relationships or causes of processes and only descriptive properties about the feature (mean, mode, median, dispersion, frequency) can be derived.
- **Multivariate models**, where a single feature is observed over time using multiple features and forecasts are expressed following their previous observations. An example of such model is the Vector auto-regressive (VAR) model. Although multivariate models bring more complexity, they are able to capture correlations among features.

Depending on the forecast horizon (the number of data points to be forecasted), there are two formats of forecasts:

- **One-step forecast**, where forecasts are made for $t+1$ data point.
- **Multi-step forecasts**, where forecasts are made for $t+n$ data points, that is, more than one step ahead. This forecasting format could be divided into two additional categories. In multiple output forecasts, the model predicts t+n data points at once. Meanwhile, in recursive multi-step forecast for $t+n$ data points, the model outputs the value for the $t+1$ step, which is used by the model to predict $t+2$ data points, and so on.

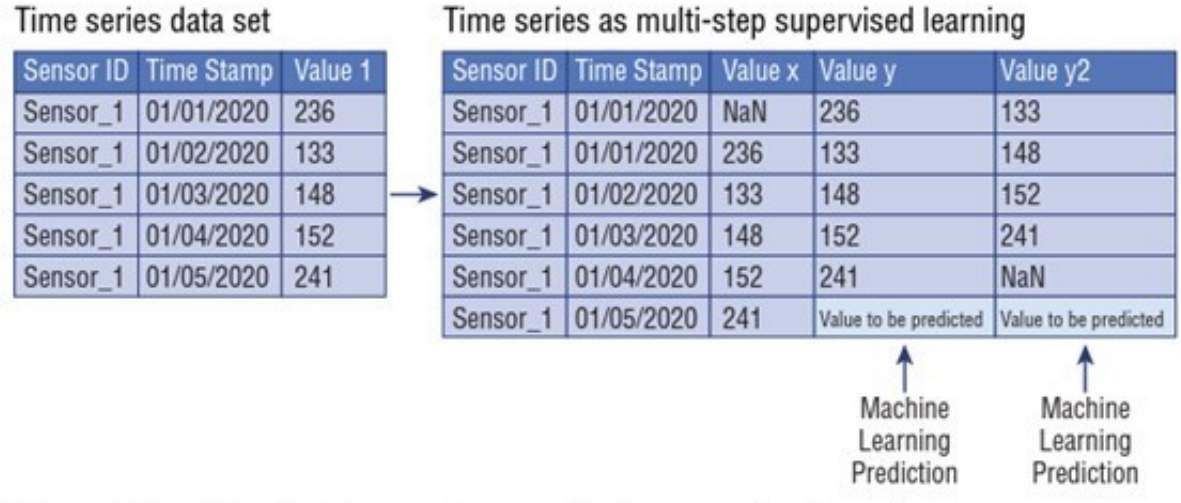


Fig. 1. An Example of Univariate Multi-step Forecast [24].

2.2.1. Classical linear methods

Classical linear methods assume that past and future data points are related to each other through a linear relationship, or autoregression. In other words, the future values are expressed as a linear function of the previous values in the series. An important statistical property in autoregressive models is autocorrelation, which measures how a time series is related to its past lags. The stronger the correlation between a value and its lag, the more weight it carries over for the future forecast.

In business, economics, and finance settings, one of the most widely used autoregressive models is **ARIMA (Autoregressive Integrated Moving Average)** [24]. The model consists of three components:

- Autoregressive (AR) part defines an autoregressive relationship to the time series. It is measured by order of p, which shows the number of lag values included in the model.
- The integration (I) part determines the number of times the series is differenced. One of the model assumptions is stationarity, which means that statistical properties of the time series (mean, variance, covariance) do not change over time. Therefore, differences are needed to make the time series process stationary. The component is measured by d, or the degree of differencing.
- Moving Average (MA) part marks the time series dependency on the previous forecast errors. The idea behind this is that future values can be predicted as a weighted moving average of past forecast errors. The component is measured by q, the order of the moving average.

The expression for the ARIMA model in lag operator form is given below:

$$(1-\phi_1 L-\phi_2 L^2-\ldots-\phi_p L^p)(1-L)^d y_t = (1+\theta_1 L+\theta_2 L^2+\ldots+\theta_q L^q)\varepsilon_t, \tag{10}$$

here ϕ refers to autoregressive coefficients up to the order p, L - the lag operator, d - the degree of differencing, θ - moving average coefficients up to the order q, ε_t - white noise with mean zero and constant variance.

ARIMA models are estimated following the Box-Jenkins methodology. According to [22], the simplified process can be defined in three stages:

1. Data generations proccess is **identified**. Here, with the help of the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots or information criteria, the orders of p and q for stationary data are determined.
2. AR(I)MA model is **estimated** using p and q lags, and the residuals are derived.
3. **Diagnostics for residuals**. A correct model should return residuals that are white noise, in other words, it does not possess information useful with the estimation. In practice, two types of tests are used for residual diagnostics: Box-Pierce and Ljung-Box test. The former test is used for large datasets, whereas the latter performs better in small samples.

On a separate note, the Ljung-Box test is a more often used white noise test. According to [22], the test is one of the tools to verify that the number of p and q lags is suitable for the model. Its null hypothesis states that:

H_0 The residuals up to lag k are distributed independently. In other words, all residual autocorrelations up to lag k are zero.
against the alternative hypothesis:

H_A The residuals up to lag k are not distributed independently. This means that the residuals are autocorrelated.

[22] recommends the number of k lags from $k = p+q+2$ up to a quarter of the sample size $k = T/4$.

In addition, there are alternatives of ARIMA models. For instance, ARIMAX model includes exogenous variables into the equation, AR model considers only autoregressive lags, whereas SARIMA includes seasonality, if the time series has an expressed seasonal component.

Among multivariate models, **vector autoregressive models (VAR)** are some of the most widely used in economic and time series analysis and forecasting [7]. The reason behind their popularity is their relative similarity to ordinary regression models, as VAR models can be estimated using least-squares methods. The model can be expressed in the form below:

$$Y_t = \mu + \sum_{i=1}^{p} A_i Y_{t-1} + \varepsilon_t, \tag{11}$$

here Y_t is the vector of the time series at time t, μ is the vector of constant, A_i is the matrix of coefficients up to the lag p, ε_t is the vector of white noises.

We can see that the future variables for each feature are defined as a linear expression of its past values and the past values of other predictors. The model previews the bidirectional relationship among pairs of features. In other words, the features have an impact on each other. Therefore, there are as many systems of equations as there are variables in the model. It is noted that, likewise for ARIMA, the VAR process assumes stationary time series features.

2.2.2. Machine Learning in time series forecasting

The increasing availability of data as well as computing power allowed machine learning to be used in time series forecasting [25]. The area has been recently dominated by Recurrent Neural Network (RNN) algorithms. According to [6], the efficiency of RNNs is explained by its ability to take into account all the values of the history of the time series in predicting future ones. [25] explain the technical side of the RNN working mechanism. RNN layers contain memory state, which summarises past information. These layers are recursively updated with new information. As a result, RNNs naturally serve as a great tool for time series analysis. Moreover, unlike convolutional neural networks,

to be soon discussed, RNNs do not require a lookback window of a certain size.

[6] proposed using Convolutional Neural Networks (CNN) for financial time series forecasting. CNN is another deep learning algorithm, widely used in image classification. CNNs consist of convolutional layers, where a filter (weight matrix) slides over the inputs producing dot products. This way the model learns specific features from the data. [6] employ a Wavenet architecture for their CNN model to forecast various financial time series (S&P500, the volatility index, the CBOE interest rate as well as a couple of exchange rates). Compared to autoregressive linear models, the CNN model managed to capture noisy dependencies. Similarly, compared to LSTMs, CNNs have fewer weights, which allows speedier training and more accurate predictions. The authors used WaveNet architecture, initially employed for audio forecasting. The model resulted in lower MASE (mean standard deviation) than its autoregressive and LSTM equivalents. For better forecasts, the authors recommended to use a greater number of filters and layers.

2.2.3. Transformer models

A potential major improvement to time series modelling could be transformer models, introduced by Vaswani et al. Major advancements in speech recognition, computer vision, and particularly in Natural Language Processing (NLP), have been achieved thanks to transformer architecture [50]. Transformers are defined by self-attention mechanisms (usually in the NLP context), where meaningful connections among elements (for instance, words in a sentence) are extracted. As a result, the transformer architecture has received increasing interest in time series modelling. [45] name transformer applications in various time series tasks, such as anomaly detection and classification (for example, classifying raw optical satellite time series). As [25] point out, the attention mechanism enriches time series modelling in two ways. Firstly, transformer networks capture significant events, for instance, holiday or promotional periods in a retail context. As a result, by placing higher weights on holiday or promotional periods, neural networks with attention mechanism can improve sales forecasts. Secondly, models with an attention mechanism are capable of adjusting different weight patterns to different temporal dynamics. For instance, in stock markets, they could adjust different weights depending on whether the market is stable or on the crash.

On the other hand, the self-attention mechanism suffers from lack of sensitivity to the order of elements, which could be harmful in time series modelling, where the order of elements is very important when forecasting future values [45].

Due to parallelisation, transformers outweigh other complex architectures, such as convolutional layers or recurrent neural networks (RNN). This feature allows transformer models to be trained on large datasets without having to process tokens one by one, unlike RNN layers [15].

So far, there have been no attempts to forecast Consumer Price Index using transformer models. For the above mentioned benefits of transformer models, it would be an innovation to study how accurately transformer models could predict CPIs.

The figure 2 outlines the complete architecture of the transformer layer. A full block consists of an

encoder (left) and a decoder (right) structure. The encoder maps an input structure to a continuous representation and then passes it over to the decoder to generate the output sequence. The layer is autoregressive, meaning that the model takes previously generated outputs as input to the decoder. In the original Vaswani et al. paper, the transformer layer consists of 6 identical encoder-decoder blocks.

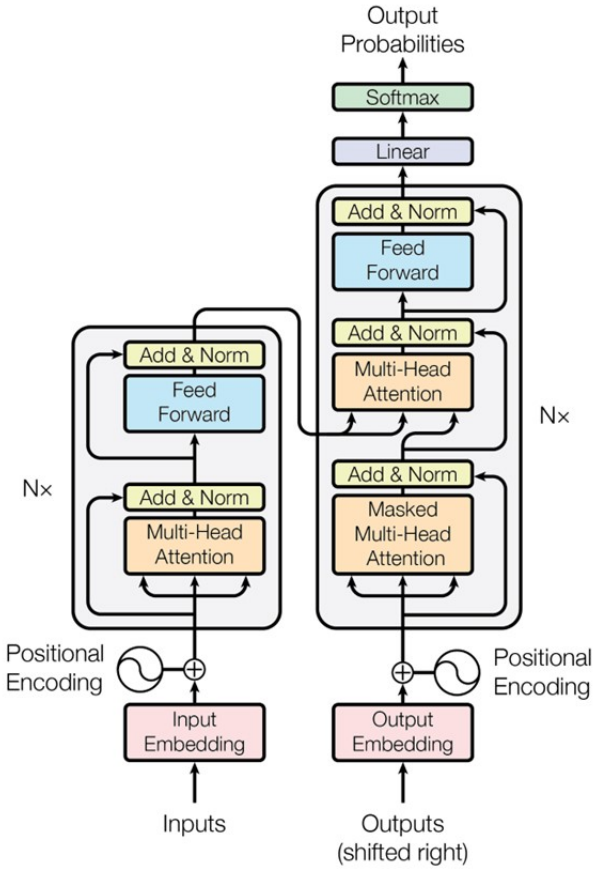


Fig. 2. The structure of a transformer architecture [44].

[15] describes each block in detail:

- The **Multihead attention** layer in **encoder** block consists of several attention heads. Each attention head learns a unique attention mechanism, thus, allowing the model to capture complex relations in the data. In other words, an attention mechanism decides where in the input it will extract the information. It does so by multiplying query (Q), key (K) and value (V) matrices, which are representations of inputs embeddings. The dot product of Q and K vectors is scaled down by $\sqrt{d_k}$ (the dimension of the K matrix) and passed through the softmax layer to sum up to 1. See figure 3 in for graphical representation of calculations in the attention layer.

$$\text{Attention}(Q,K,V) = \ softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{12}$$

  Next, we have a **fully connected layer** with Rectified Linear Unit (ReLU) activation function. Both sublayers follow the **normalization layer**, which provides stability for the model during

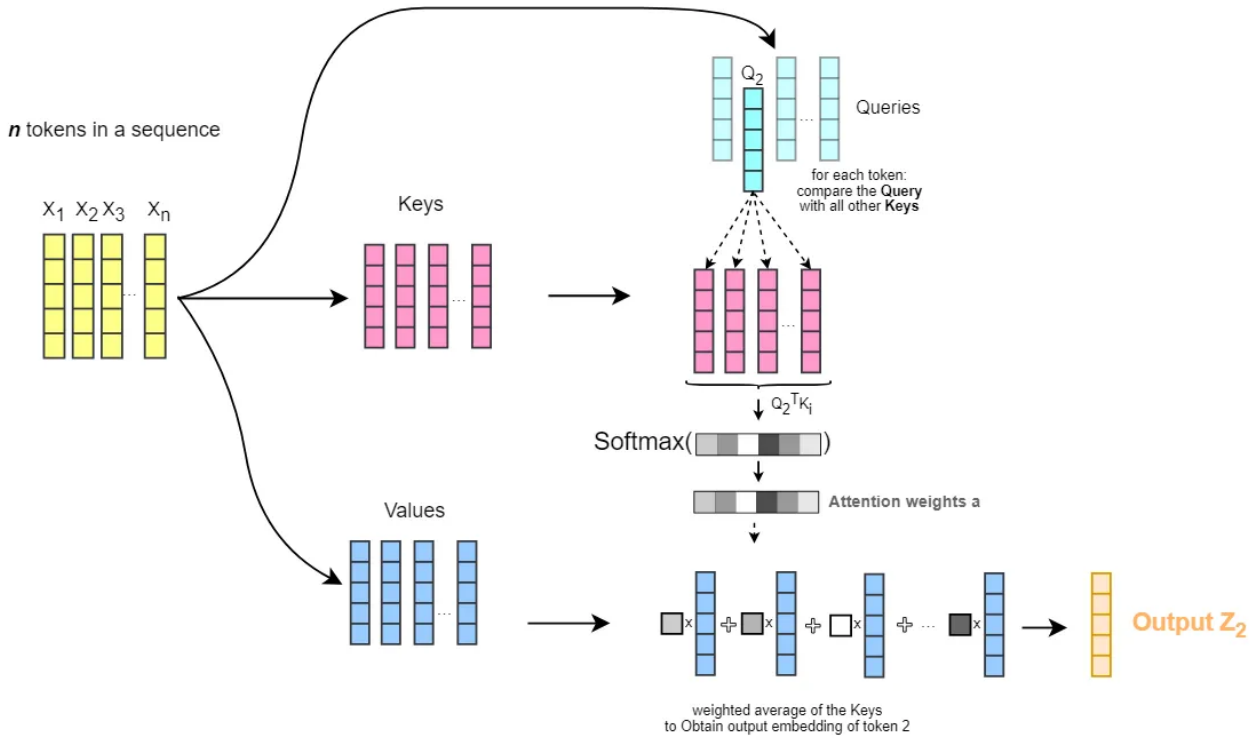


Fig. 3. Calculation of attention mechanism. Illustration by [34].

the training process.

- The first multi-head attention sublayer in the **decoder** block, as mentioned previously, receives the output from the previous output from the decoder. However, while the encoder does not regard the order of sequence in the original embeddings, the decoder is designed to use **positional encoding**. It encodes the position of each token in the token embedding, enabling to save the order of original sequence, so that decoder attends only preceding tokens. The future keys are hidden using the **causal mask**. The second multi-head attention sublayer receives outputs from the encoder and the previous multi-head attention layer in the decoder. For this multi-head attention mechanism the key (K) and value (V) matrices come from the encoder part, whereas the query (Q) part derives from the previous decoder output. As figure 2 highlights, the output is followed by the fully-connected layer and again inserted by normalization layers.

Since the introduction of transformers, new versions of transformers have been released to address weaknesses in time series forecasting. One of them concerns transformers failure to address multi-horizon forecasts accurately. Another weak assumption regards exogenous features, a common one in many autoregressive models. It says that models take external factors in forecasts as granted, despite uncertainty surrounding them. Similarly, models do not take advantage of the static features used in forecasting models. To address these issues, [25] have introduced **Temporal Fusion Transformers (TFT)**. As their architecture in figure 4 suggests, Temporal Fusion Transformers use additional features compared with traditional transformers:

1. **Gating mechanisms** are used for efficiency and adaptability by removing unnecessary features.

Relevant features are filtered in **variable selection networks.**

2. **Static features** are incorporated in the neural networks.

3. Both short- and long-term temporal features are extracted in the Temporal Fusion Decoder part, using upgraded multi-head attention block.

4. Predictions are generated as **quantile forecasts** in various percentiles (10th, 50th, 90th, etc.).

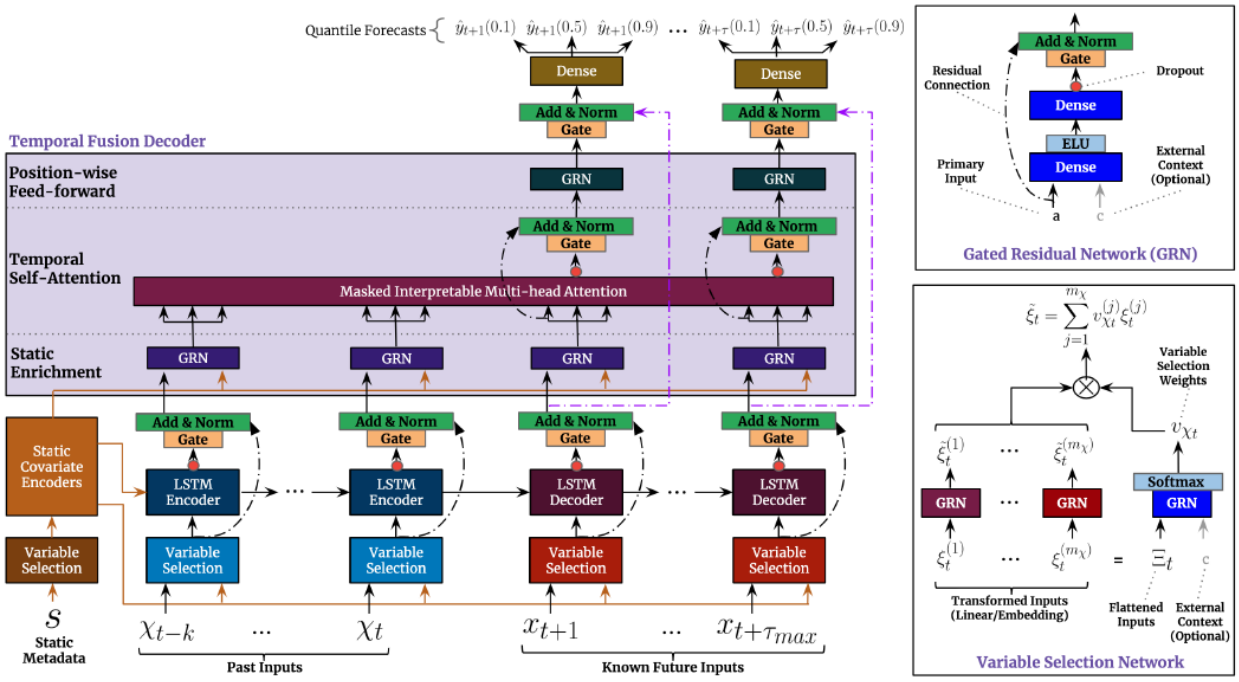


Fig. 4. Temporal Fusion Transformer architecture by [26].

Another advantage, compared to other types of neural networks, is that Temporal Fusion transformers provide interpretation of forecasts thanks to the magnitude of attention mechanism weights. As a result, all the features, including the static ones, at every time step can be displayed by their importance to the prediction output.

2.2.4. Interpretation of neural network time series forecasts

When it comes to making predictions, neural networks are notorious for the lack of interpretation when making predictions. It is not uncommon to consider these algorithms as "black boxes". [25] define two approaches to interpret deep learning models.

According to the first approach (techniques for post hoc interpretability), a simple approximated model is created for the inputs and outputs of the neural network model. Explanations are provided based on the interpretation of the surrogate model. There are two tools proposed for the post hoc interpretability. LIME (local interpretable model-agnostic explanation) fits a linear model for the input and output data. Feature importance is then interpreted as the model's linear coefficients. Another

tool, SHAP (Shapley additive explanations), identifies important features based on Shapley values from cooperative game theory. Nevertheless, the post hoc interpretability methods are criticised for focusing solely on feature importance rather than sequential dependency, in other words, that values in time series could be also influenced by past values, possibly trend or seasonality.

The second approach involves using attention weights. When designing neural network architecture, attention layers could be included at each time step with weights for features and time steps. With attention weights being produced as outputs from softmax function and therefore summing up to 1, it is possible to derive comparative feature importance at different time periods. [25] add up that analysing attention vectors over time could even help indicate seasonal patterns in time series data.

2.3. Data

The data used in the forecasting models consists of monthly Harmonised Consumer Price Indices from 29 European countries derived from Eurostat database. The dataset covers the period from January, 1998, to January, 2024, thus, comprising 313 observations. Harmonised Consumer Price Indices are indexed to the reference year 2015 (2015 = 100) and an all-item category is used. As a result, the index reflects weighted price changes according to their respective weights across the 12 COICOP components for each country. The target variable to be used in the forecasting models is Lithuania's HICP. In addition, the dataset contains two more features: the unemployment rate (UNEMP) and Price Producer Index (PPI) in Lithuania. The idea of adding the former stems from the Phillips curve, as was explained in the previous chapter. Meanwhile, PPI was adopted in the [30] CPI forecasting model. According to the State Data Agency of [28], similarly to CPI, it denotes price changes of producer output within a certain time range (in this case - a monthly one) at different stages of manufacturing. The motivation to include this variable is explained by Yamarone and Stakėnas: PPI tracks price changes from the producer side before they reach the retail level. Because it covers the prices of raw materials and intermediate goods, economists use it to predict changes in the consumer price index. Finally, both PPI and CPI are said to be highly correlated in the long term. As a result, together with the HICP of various European countries, Lithuania's PPI and unemployment rate will also be used to forecast Lithuania's HICP. All variables used in models are described in table 3 together with engineered features in table 4 (see section below on feature engineering).

The HICP values on a country basis are presented in figure 5. In the graphs, we can identify three trends. Firstly, what stands out in is the convergence of the new European Union Member States in Central and Eastern Europe (Bulgaria, Czechia, Estonia, Croatia, Lithuania, Latvia, Hungary, Poland, Romania, Slovenia, Slovakia) to their old counterparts in terms of their HICP values. All new Member States start off at HICP values below 60 in 1998 and witness sharp growth of inflation rates until they level off around 2010. Secondly, the next period from 2010 to 2020 is marked by a stable growth both in Eastern and Western Europe. The beginning of the third stage coincides with the outbreak of the COVID-19 pandemic, as the virus rapidly spread throughout the world and lockdowns were introduced in most countries. During this period, indices soar at all European countries and are even more fuelled by Russia's invasion in Ukraine in February, 2022. High inflation once again seems to

Table 3. Variables used for Lithuania's HICP modelling

Title	Description	Title	Description
BE	Belgium's monthly HICP rate	**LU**	Luxembourg's monthly HICP rate
BG	Bulgaria's monthly HICP rate	**HU**	Hungary's monthly HICP rate
CZ	Czechia's monthly HICP rate	**MT**	Malta's monthly HICP rate
DK	Denmark's monthly HICP rate	**NL**	Netherlands' monthly HICP rate
DE	Germany's monthly HICP rate	**AT**	Austria's monthly HICP rate
EE	Estonia's monthly HICP rate	**PL**	Poland's monthly HICP rate
IE	Irelad's monthly HICP rate	**PT**	Portugal's monthly HICP rate
EL	Greece's monthly HICP rate	**RO**	Romania's monthly HICP rate
ES	Spain's monthly HICP rate	**SI**	Slovenia's monthly HICP rate
FR	France's monthly HICP rate	**SK**	Slovakia's monthly HICP rate
HR	Croatia's monthly HICP rate	**FI**	Finland's monthly HICP rate
IT	Italy's monthly HICP rate	**SE**	Sweden's monthly HICP rate
CY	Cyprus' monthly HICP rate	**IS**	Iceland's monthly HICP rate
LV	Latvia's monthly HICP rate	**NO**	Norway's monthly HICP rate
LT	Lithuania's monthly HICP rate	**UNEMP**	Monthly unemployment rate in Lithuania
PPI	Monthly Producer Price Index in Lithuania		

have severely affected the economies in Central and Eastern Europe. However, the price level growth reached its highest point at the end of 2023.

Table 4. Engineered features for the dataset

Title	Description
YEAR	Reported year
MONTH_1 - 11	Dummy variable for reported month
LT_EURO	Dummy variable denoting, whether the reported period is before Lithuania entered the euro zone (0) in January, 2015, or after (1).
LT_COVID	Dummy variable denoting, if the reported period prior (0) the Covid pandemic in April, 2020, or afterwards (1)
LT_MEAN_3	Three month rolling average of Lithuania's HICP
LT_PPI_MEAN_3	Three month rolling average of Lithuania's Producer Price Index
LT_UNEMP_MEAN_3	Three month rolling average of the unemployment rate in Lithuania

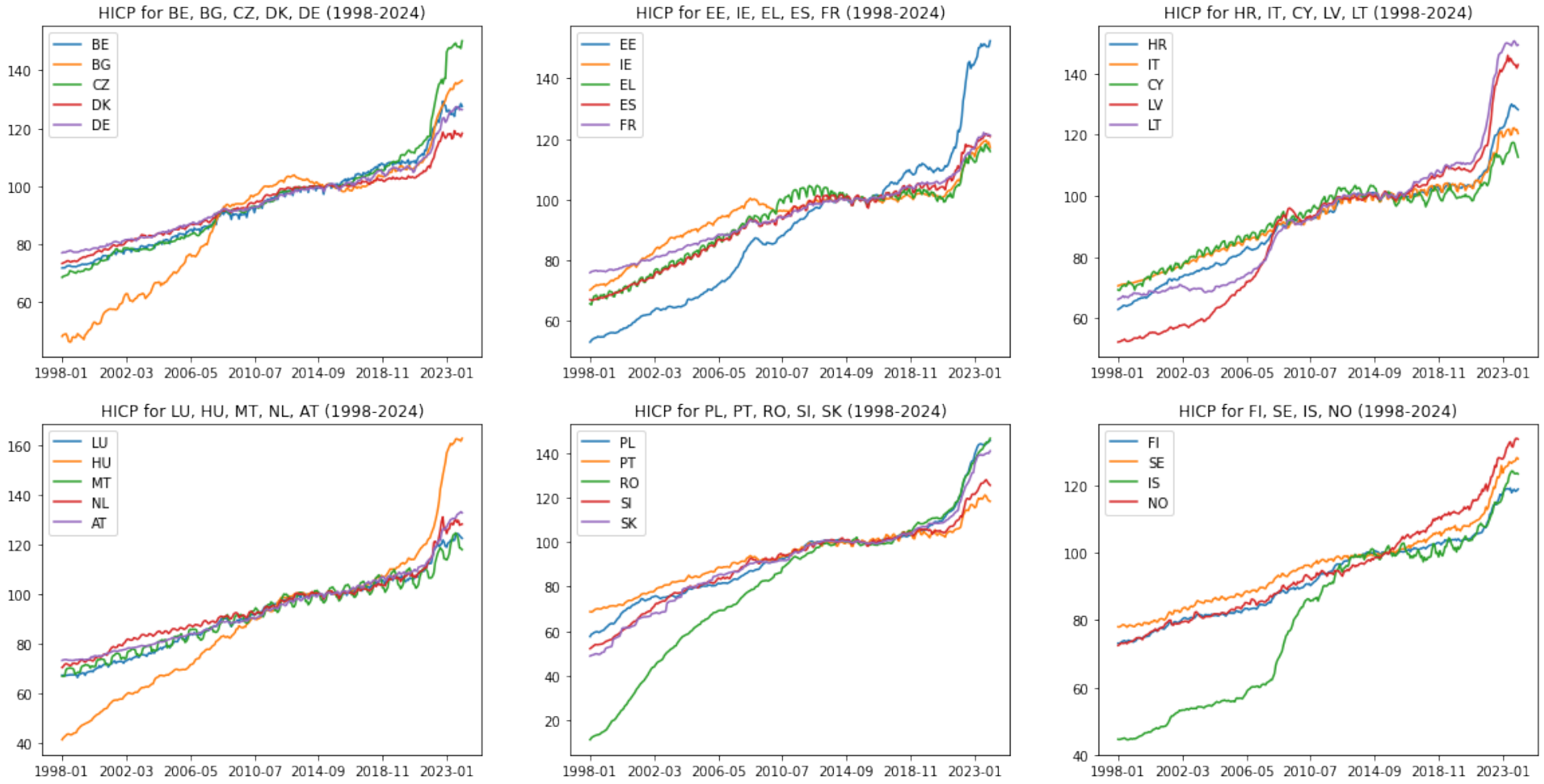


Fig. 5. Harmonised Index of Consumer Prices Monthly Values for 29 European Countries from 1998 to 2024.

To dig deeper into Lithuania's HICP, the indices are illustrated in figure 6. Although Lithuania clearly follows the trends described above and was affected by inflation shocks following the Covid-19 pandemic and Russia's invasion in Ukraine, it is noticeable that the country's price indices had previously gained another momentum in 2015, as the country joined the euro zone. Consequently, when forecasting Lithuania's consumer price index, it is crucial to consider both its accession to the euro zone and the pandemic together with Russia's invasion in Ukraine periods. Therefore, two dummy variables *LT_EURO* and *LT_COVID* will be added to the original dataset.

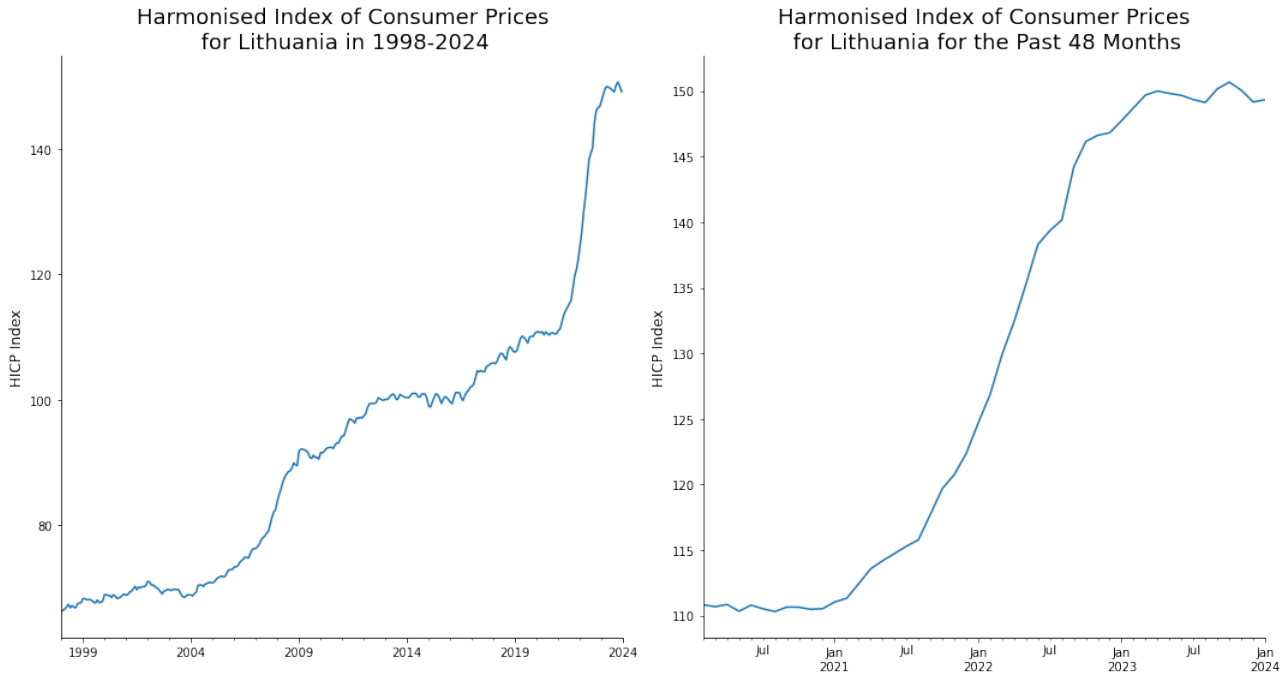


Fig. 6. Harmonised Index of Consumer Prices for Lithuania from 1998 to 2024.

Lithuanian HICP has been analysed using seasonal trend decomposition using LOESS (STL). The decomposition results are shown in figure 7. The plot reveals that the time series has an expressed trend component. In contrast, the seasonal component is low in amplitude, indicating that seasonality has a limited influence on Lithuania's HICP.

Next, two methods will be used for time series data analysis. Cross-correlation analysis is a tool to explore bivariate relationships between pairs of features. As [37] defines it, the conventional Pearson correlation compares the relationship between two time series variables at the same time point by a single coefficient, thereby missing any lead-lag relationship. In cross-correlation analysis, on the other hand, one variable has leads and lags against another variable, for which correlation coefficients are calculated at different time lags.

Granger causality was first described by Granger in 1969 [22]. It allows to test for causality between pairs of time series variables without inferring causal links between them. The test is popular in economic research when studying relationships between short- and long-term interest rates, returns on different stock markets, and so on. The null hypothesis states that x_t does not Granger cause y_t. This means that the lagged values of x_t do not explain the current ones of y_t. It should be noted that the test could be conducted in pairs both ways: that x_t Granger causes y_t and that y_t Granger causes

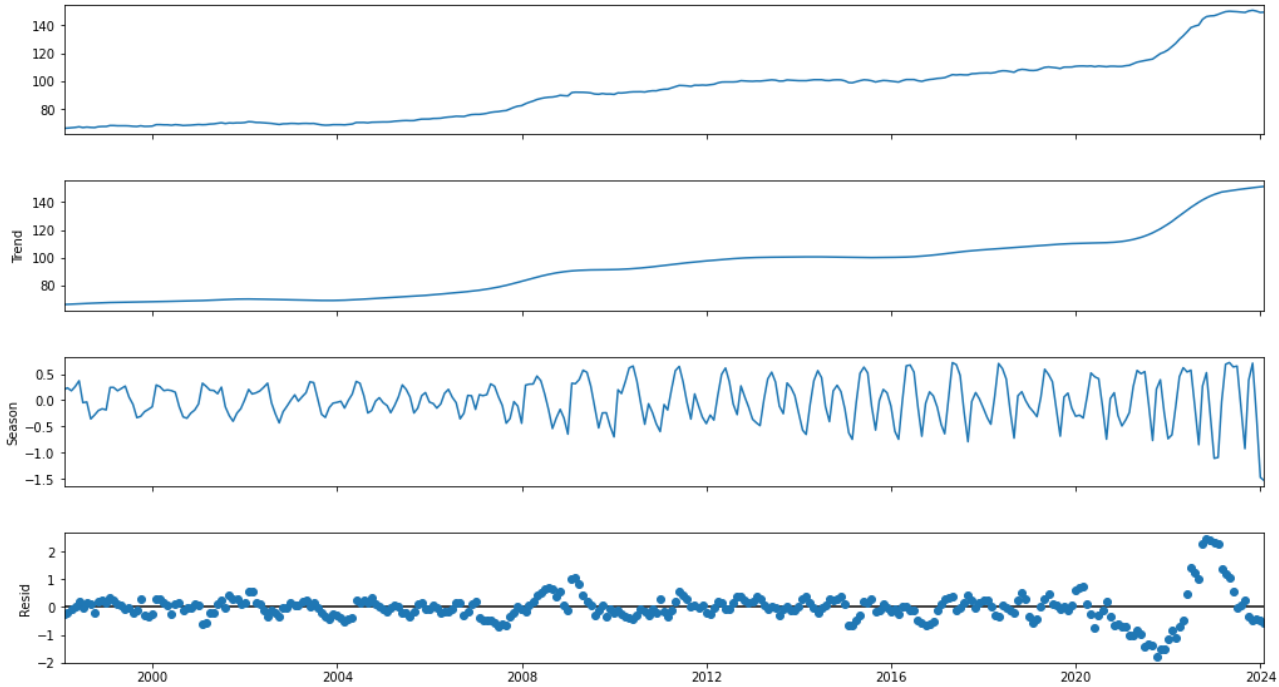


Fig. 7. Season-Trend Decomposition Using LOESS of Lithuania's HICP (1998-2024).

x_t. The test is calculated using F-statistics, using the formula below:

$$F = \frac{(SSR_R - SSR_U)/p}{SSR_U/(T-2p-1)} \tag{13}$$

here SSR_R and SSR_U refer to the sum of the squared residuals for restricted and unrestricted regressions, T - the number of observations, p - the number of lags.

2.4. Feature selection

[10] defines three ways to select features for the model. Firstly, features could be selected by using a linear model and filtering the features based on their *pvalues*. Secondly, features can be selected using stepwise selection procedure, where the model is constructed by adding features one by one to the model until it reaches a satisfying performance. The last recommended way is to build a random forest model and select N features based on their importance. The latter method will be used to select European countries whose HICP will be included to the model. 4 methods of feature selection techniques have been identified in the literature and will be used here:

- **Variable importance** in linear models, such as linear or logistic regression, is measured by p-value and t-statistics of each feature [43]. Meanwhile, in ensemble models, for instance, a decision tree, splits are caused by variables used in the model. A certain variable could be used as a split criterion for a couple of times in the tree and contribute more to the performance, while another variable may never appear in the splits. Likewise, the decision tree's performance is measured as a decrease in impurity for each split, where the impurity is set to be minimized.

Tattar gives an example of feature importance in a decision tree using the *kyphosis* dataset. The dataset consists of 4 variables: *x, y, z* and the target *Kyphosis*, which indicates the existence of the kyphosis following a surgery. The first two splits are made based on the variable *z*, whereas the two next ones are made based on the *x* variable. The structure of splits suggest that the variable *y* carries no importance for the overall model.

- According to the **Recursive feature elimination** method, a model is built by recursively removing different features from it and evaluating the model's performance. The goal of this technique is to construct a set of features to maximize the objective. The features are finally ranked by when they were eliminated. However, this method returns the ranking of features with no available numeric comparison among the features [10].

- Another method to evaluate feature importance is via **permutation** [40]. Following this method, values of a certain variable in different rows are swapped and passed to the model. Once the accuracy score for the permuted row is calculated, it is possible to compare it with the original row. The importance of a variable is calculated by calculating the change in model accuracy.

- Previously mentioned **SHAP tool** is somewhat similar to the permutation method. This method was proposed by Lundberg and Lee (as described in [19]). The idea behind the method is calculating Shapley values for each prediction both including and excluding a variable in the model. In detail, Shapley value describes an average marginal contribution of a feature to the prediction together with different sets of other features. As a result, Shapley values for features can be calculated for either an individual prediction or the model overall. The values can be positive or negative, indicating the respective feature's impact to the model.

2.5. Feature engineering

Lazzeri defines feature engineering as "the process of using raw data to create additional features or variables to augment your data set" (pg. 62) with the goal to improve the model performance. In time series forecasting, it is achieved in two steps. First, correct input features from historical are prepared for supervised learning. Second, additional features from the data are created, which create relationships between inputs and the target. Lazzeri recommends 4 methods to generate new features:

1. **Date features**, where hour, month, weekday, public holiday or other information is derived from the time stamp at each row.

2. **Lag and window features** assume that events in the past can provide information about the events in the future. For instance, a feature about sales on a previous day could be useful predicting the sales on the next day. Examples of such features include lagged values on various frequencies (calendar day, weekly, monthly, quarterly, etc.)

3. **Rolling window statistics** allow to create features on a fixed range of time before and/or after the sample time. Instances include rolling mean, maximum or minimum values.

4. **Expanding window statistics** are similar to rolling window statistics, except that features include not a fixed range, but all previous values up to the observation.

For this dataset the following features were engineered: rolling 3 month mean for Lithuanian Consumer Price Index, Producer Price Index and Unemployment rate variables. In addition, year and month features were extracted. Month features have been converted into dummy variables.

2.6. Training

Time series forecasting models are trained using **rolling-origin-recalibration evaluation**, also known as nested cross-validation, training method. As [4, pg. 194] defines it, "forecasts for a fixed horizon are performed by sequentially moving values from the test set to the training set, and changing the forecast origin accordingly. For each forecast, the model is recalibrated using all available data in the training set, which often means a complete retraining of the model." In this case, it means that each model will be retrained by taking sequentially row by row from the test dataset and making forecasts $t+h$ horizons ahead. As a baseline, three linear models will be used: Vector Autoregression (VAR), Autoregressive Integrated Moving Average with Explanatory Variables (ARIMAX) as well as Autoregressive Distributed Lag (ARDL).

The benchmark models will be compared with four different types of transformer models: (1) the one with only multihead attention layer, (2) the one with Encoder layer, (3) the one full transformer layer, consisting of Encoder and Decoder, and (4) the one with Temporal Fusion Transformer model. All models were trained in Python environment. For linear models, *pmdarima* and *statsmodels* libraries used, whereas the first three types of transformer models were trained in *Tensorflow* framework. Lastly, Temporal Fusion Transformer model was trained in *PyTorch Lightning* framework using *PyTorch Forecasting* library. Parameters for the latter model were taken from the original [26] paper. Neural network model architectures are displayed in the Appendix section.

The experiments will be carried out in two stages. At the beginning, univariate models will be trained, where Lithuania's HICP is forecasted using only the same variable's data. Then, multivariate models will be trained with previously selected country HICP variables, unemployment rate, PPI and engineered features. Each model will have three forecasting horizons: $h = 1$, $h = 3$ and $h = 6$. The choice of horizons is motivated by economic theory in the first chapter. On the other hand, it is important to admit that some reviewed CPI models had short term forecasts [33, 42], whereas other focused on medium-long term forecasts [1, 30, 32] or both [3, 38].

The number of look back periods, or lags, for VAR and ARIMA models was determined by Akaike Information Criterion: 3 lags for VAR and ARDL, (2,1,2)x(1,0,1,12) for SARIMA univariate model, (2,1,5) for multivariate ARIMAX model for forecasting horizons $h = 1$ and $h = 3$, and finally (4,1,5) for ARIMAX for forecasting horizon of six months. For the remaining models, the number of look back periods was also chosen 3. While the majority of reviewed papers on CPI forecasting do not detail the number of lags included in various models, [3] included up to 4 lags, [23] up to 6 lags. In addition, using 3 lags for all models in forecasting allows fair comparison in their performance.

3. Research results

3.1. Data analysis

The figure 8 in pg. 39 plots cross-correlations between Lithuania and other European country stationary Harmonized Consumer Price Indices (see the following section on stationarity) across the time. The plot reveals a trend of positive correlations for Lithuania with other European countries ranging a couple of lags preceding and following zero. However, in general, the maximum values for each pair do not exceed 0.5, indicating relatively weak relationships. There are notable exceptions for Lithuania's neighbors Latvia and Poland, where correlations reach 0.71 and 0.64 at lags zero for both, respectively. High correlations signify a strong positive relationship at the same time between Lithuania-Latvia and Lithuania-Poland HICPs.

Furthermore, Granger test was conducted with the aim to discover, whether any European country's HICP has Granger causality towards Lithuania's HICP. The test's p-values for each country-Lithuania pair are displayed in the Table 5. We can see that for all pairs the p-values are <0.01. It means that the null hypotheses are rejected in all cases. This suggests that all variables used in the test have strong Granger causality towards Lithuania's HICP. The analysis of HICP is concluded that conventional tools of cross-correlation analysis and Granger causality are not sufficient to select features for time series forecasting models. As result, more elaborated feature selection techniques using neural network models will be employed.

Table 5. Granger test p-values for country (x) - Lithuania (y) pairs.

BE ->LT	**BG ->LT**	**CZ ->LT**	**DK ->LT**	**DE ->LT**	**EE ->LT**
0.0000	0.0000	0.0000	0.0001	0.0002	0.0000
IE ->LT	**EL ->LT**	**ES ->LT**	**FR ->LT**	**HR ->LT**	**IT ->LT**
0.0000	0.0000	0.0000	0.0000	0.0035	0.0000
CY ->LT	**LV ->LT**	**LT_PPI ->LT**	**LT_UNEMP ->LT**	**LU ->LT**	**HU ->LT**
0.0003	0.0004	0.0000	0.0007	0.0000	0.0000
MT ->LT	**NL ->LT**	**AT ->LT**	**PL ->LT**	**PT ->LT**	**RO ->LT**
0.0000	0.0000	0.0000	0.0001	0.0000	0.0000
SI ->LT	**SK ->LT**	**FI ->LT**	**SE ->LT**	**IS ->LT**	**NO ->LT**
0.0000	0.0008	0.0011	0.0000	0.0000	0.0000

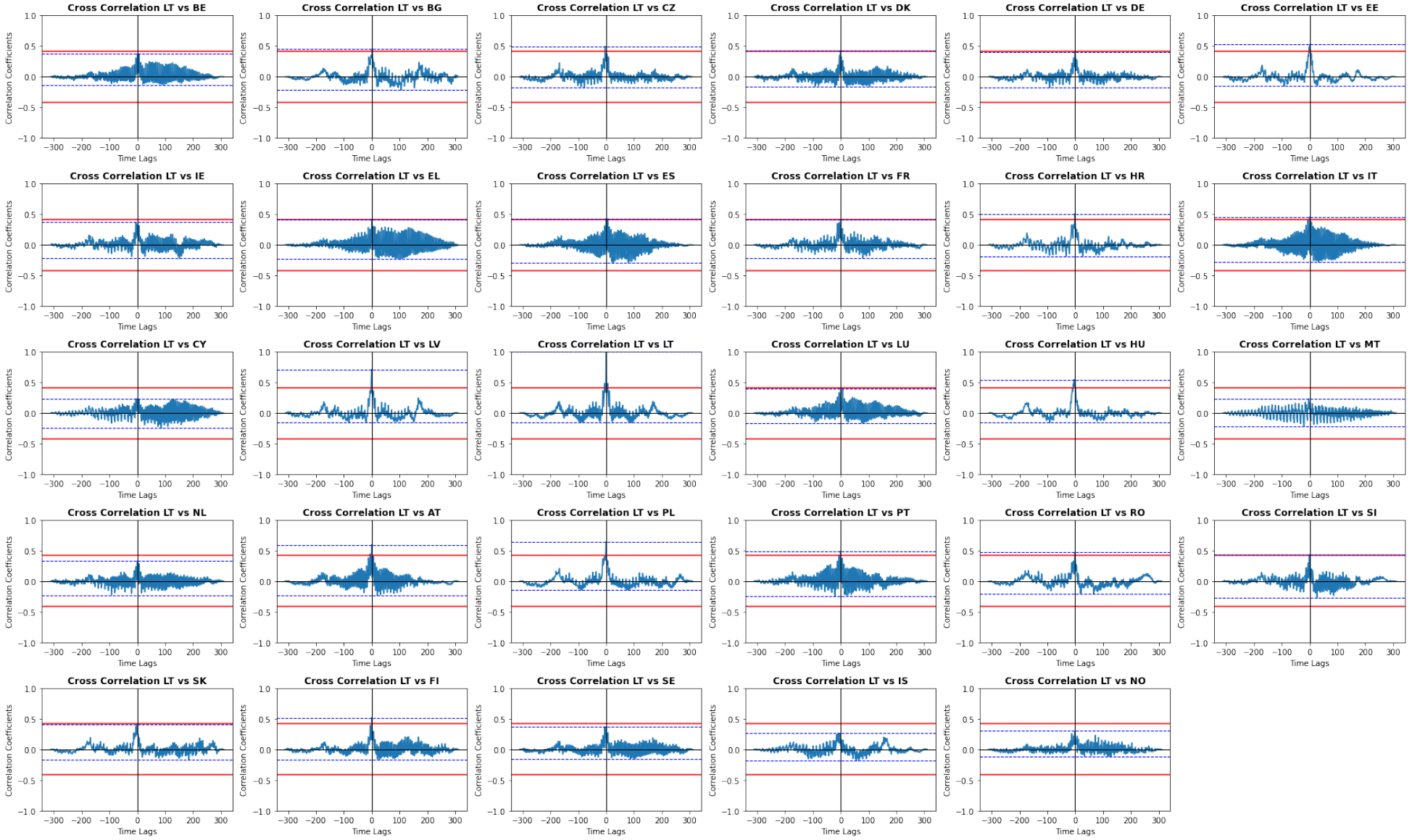


Fig. 8. Cross-correlations between Lithuania's Harmonized Consumer Price Index and other European countries

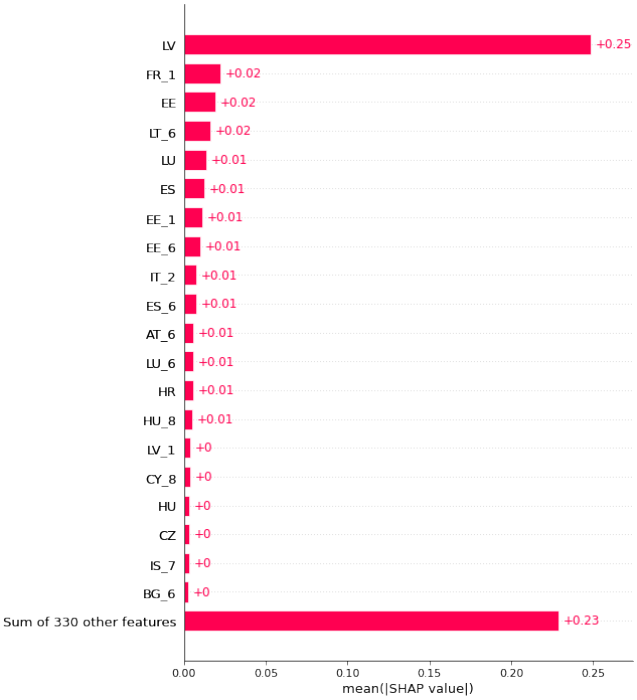
3.2. Feature Selection

Three different types of ensemble models (Random Forest Regressor, Gradient Boosting Regressor, and XGBoost) were trained together with a Linear Regression model with the aim to select country HICP features. In order to identify not only the most important countries, but also time lags, the dataset were transformed so as to include lags from t up to $t-12$ for each country. In total, 350 country-lag features passed to the feature selection models.

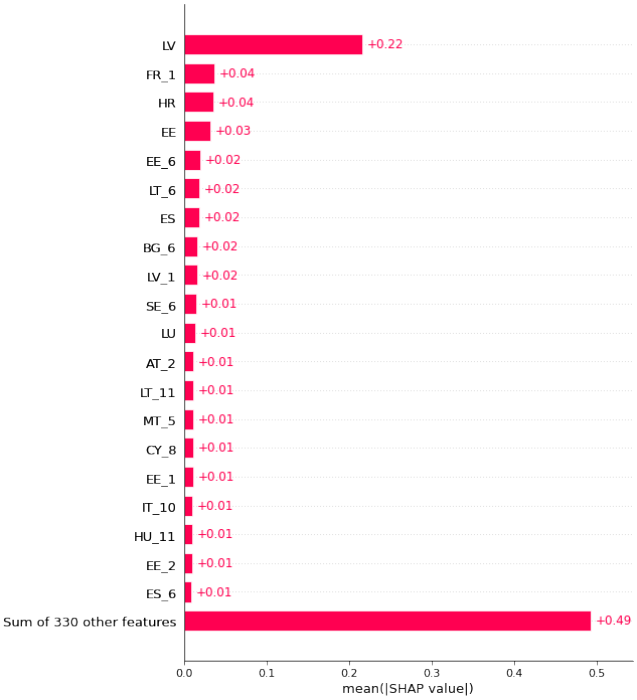
In the figure below 9 will be discussed the top 20 selected features by the SHAP tool. Firstly, what stands out from all three ensemble models is that Latvia's HICP values at the current month on average have the highest SHAP scores. In other words, Latvia's current month index values have the highest impact on Lithuania's ones. In fact, as we recall in the second chapter, the Lithuania-Latvia cross-correlation scored 0.71 at lag zero, so apparently the ensemble models took advantage of highly correlated values. Secondly, all three models place high importance on the past month (lag 1) values of the French HICP when determining Lithuania's present HICP values. All ensemble models highly evaluate Hungarian (lag 0), Estonian (lags 0, 6), Lithuania's itself (lag 6), Spanish (lag 0), and Latvian (lag 1) index scores. In summary, the other two Baltic States are seen by ensemble models as important features for Lithuania's harmonised consumer price index. As for other countries, on the other hand, we cannot single out one European region to have a high impact. We can see that among the most important country features, there are countries from various regions: Western Europe (France, the Netherlands, Luxembourg), Southern Europe (Spain, Greece, Italy, Malta, Cyprus, Portugal), Central Europe (Hungary, Austria, Czechia), Northern Europe (Iceland, Sweden, Finland), and Eastern Europe (Bulgaria). When it comes to the number of lags, current (lag 0) and past (lag 1) months; two (lag 2) and six (lag 6) months in the past are the most common lags. It would suggest that the SHAP tool does not consider HICP values older than 6 months as important to forecast Lithuania's HICP.

It is noted that other feature selection techniques (Feature Importance, Permutation Feature Importance, Recursive Feature Elimination) reach similar conclusions; they will not be discussed here. Detail graphs with selected features are provided in the Appendix. The Table **??** summarises the features selected by each model using the above-described methods. It is noted that the SHAP tool is not available for Linear Regression models.

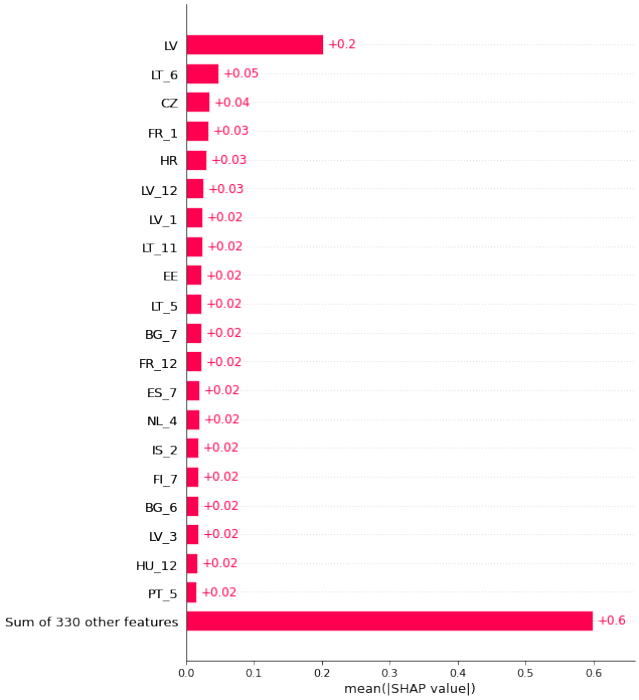
The features to the models were selected by the number of times that the country lag features appeared as the most important in each model. The selection criterion was at least 11 country occurrences in the table above, disregarding country lags. As such, 10 countries were selected for the modelling: **Latvia** (36 occurrences), **Estonia** (31), **Spain** (23), **France** (21), **Hungary** (21), **Luxembourg** (20), **Austria** (17), **Croatia** (15), **Italy** (15), **Bulgaria** (13), **Czechia** (11).



(a) Random Forest



(b) Gradient Boosting



(c) XGBoost

Fig. 9. Country-lag feature selection by SHAP tool

Table 6. Summary of 20 country-lag features selected using four feature selection methods: feature importance, Recursive Feature Elimination, Permutation Importance and SHAP techniques by Linear Regression, Random Forest, Gradient Boosting and XGBoost models.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
gbr_feat	LV	FR_1	LT_6	HR	LU	EE	CY_8	EE_6	ES	BG_6	LU_6	ES_6	IT_2	LT_2	EE_1	HU_1	EL_8	LV_1	AT_2	ES_2
gbr_perm	LV	FR_1	HR	EE	BG_6	EE_1	ES_6	HR_7	EE_6	LT_12	FI	HU	LU_6	HU_11	LV_1	LU	PL_8	CZ	AT	BE_4
gbr_shap	LV	FR_1	HR	EE	EE_6	LT_6	ES	BG_6	LV_1	SE_6	LU	AT_2	LT_11	MT_5	CY_8	EE_1	IT_10	HU_11	EE_2	ES_6
gbr_rfe	EE	ES	HR	LV	LU	EE_1	FR_1	LV_1	HU_1	LT_2	LT_3	BG_6	EE_6	ES_6	IT_6	LT_6	LU_6	CY_8	LT_11	LV_12
xgb_feat	LV	LU	HR	FR_1	LV_1	LT_6	IT_6	LV_12	LU_6	SI_9	EE	NL_4	ES_7	HU	FI_7	BG_6	ES_6	FR_2	AT_4	CZ
xgb_perm	LV	LT_6	NL_4	CZ	BG	FR_1	ES_6	HU_1	SI_1	DE	HU	PT_7	ES_7	LV_1	HR	FI	ES	FR_9	HU_11	SI_2
xgb_shap	LV	LT_6	CZ	FR_1	HR	LV_12	LV_1	LT_11	EE	LT_5	BG_7	FR_12	ES_7	NL_4	IS_2	FI_7	BG_6	LV_3	HU_12	PT_5
xgb_rfe	CZ	EE	HR	LV	LU	FR_1	LV_1	AT_1	FR_2	AT_2	HR_4	NL_4	BG_6	IT_6	LT_6	LU_6	FI_7	IT_10	IE_12	LV_12
rfr_feat	LV	FR_1	LU	EE	LT_6	ES	ES_6	EE_1	EE_6	LU_6	AT_6	LV_1	CY_8	IT_2	HR	HU_8	BE_6	SK_1	SI	IT_6
rfr_perm	LV	LT_6	FR_1	EE	ES_6	EE_1	ES	IT_2	HR	EE_6	BG_6	SK	CZ	HU	AT_6	LV_1	CY_8	DE_8	LU_7	BG_1
rfr_shap	LV	FR_1	EE	LT_6	LU	ES	EE_6	EE_1	ES_6	IT_2	LU_6	HR	AT_6	HU_8	CY_8	LV_1	HU	IS	CZ	BG_6
rfr_rfe	EE	ES	HR	LV	LU	EE_1	FR_1	LV_1	IT_2	EE_6	ES_6	LT_6	LU_6	AT_6	LU_7	CY_8	HU_8	SI_8	FI_10	LT_12
lr_feat	LV	CZ_12	CY_8	FI_7	DE	HU_2	HU	LV_3	DK_12	EE_6	DE_8	EE_11	CZ_1	CZ_9	EE_3	AT_4	SK_9	LT_11	PL_8	IT_5
lr_perm	IT_4	IS_1	IS_4	MT_10	LV	CY_8	CY_12	PL_12	ES_11	HU_1	LV_8	HR_12	BE_3	IT_10	SI_8	BG_11	AT_11	HU_7	BE_5	DE_12
lr_rfe	LV	FI	FI_1	IT_2	HU_2	LV_3	LT_3	IT_4	AT_4	IE_6	LU_6	BE_8	DE_8	HU_8	SI_8	BE_9	CZ_12	FR_12	LV_12	PL_12

3.3. Diagnostics

3.3.1. Stationarity tests

To recall the HICP characteristics of the country used in the models, the figure 5 showed a clear trend component present in the time series. As a result, the ADF test will be conducted for trend stationarity. The test results are shown in Table 7

Table 7. Augmented Dickey-Fuller test p-values for country features after first differences.

LV	**E**E	**LT**	**ES**	**FR**	**HU**	**LU**
0.00178	0.00396	0.00061	0.01883	0.10052	0.00879	0.02561
AT	**HR**	**IT**	**BG**	**CZ**	**LT_PPI**	**LT_UNEMP**
0.02671	0.08334	0.00072	0.15287	0.00650	0.02222	0.03000

As we can see, the p-values for all variables are significant, indicating that the variables became stationary after the first order of integration. This means that first differences for all features will be used in both linear and transformer models.

3.3.2. Cointegration tests

Because all variables have equal orders of integration, [22] warns of the risk that they are cointegrated. In other words, cointegration occurs when two time series, like x_t and y_t, are integrated of order 1 (I (1)) and have a linear combination $z_t = \beta_1 x_t + \beta_2 y_t$ that is stationary at the integration of order 0 (I (0)). In such a scenario, the time series VAR model is not suitable and *error correction models*, such as VECM (Vector Error Correction Model) should be estimated. In economics, examples of cointegrated variables include consumption and income or short- and long-term interest rates. As a result, on a separate note, cointegration will be tested using Johansen approach. As [22] explains, this methodology tests the possibility of cointegrating vectors, or cointegrating ranks, among $N-1$ variables. The null hypothesis H_0 states that there is no cointegrating relationship and there is no linear combination that is stationary I(0). The alternative H_A states that there is at least one cointegrating relationship between N pairs. The Johansen cointegration test was conducted and its trace and critical values at $90\%, 95\%, 99\%$ significance levels are summed up in the table 8 below on page 44.

According to the results, the test values exceed the critical values at various significance levels. It leads to failing to reject the null hypothesis and stating that time series features are not cointegrated. As a result, there is no need to use the error correction model instead of VAR. The first differences of all features will be used in both linear and transformer models.

3.3.3. Autocorrelation tests

The Ljung-Box test was performed to test for autocorrelation in residuals for ARIMA models. This test was selected in favour of the Box-Pierce test due to the size of the data set. The test results are

Table 8. Johansen cointegration test trace values and critical values for each rank up to N-1

Rank	Test value	Cointegrated	Significance level		
			90%	95%	99%
0	856.75	TRUE			
1	690.17	TRUE			
2	532.73	TRUE	302.91	311.13	326.97
3	428.75	TRUE	255.67	263.26	278.00
4	341.03	TRUE	212.47	219.41	232.83
5	263.07	TRUE	173.23	179.52	191.81
6	206.31	TRUE	138.00	143.67	154.80
7	158.82	TRUE	106.74	111.78	121.74
8	117.53	TRUE	79.53	83.94	92.71
9	83.01	TRUE	56.28	60.06	67.64
10	55.97	TRUE	37.03	40.17	46.57
11	30.81	TRUE	21.78	24.28	29.51
12	9.08	FALSE	10.47	12.32	16.36
13	1.94	FALSE	2.98	4.13	6.94

shown in the figures 10, 11, and 12 below. It appears that the test p-values in all cases are above 0.05. Therefore, it is concluded that the residuals in the ARIMA models are independently distributed. Moreover, the ACF and PACF plots in no picture show any significant spikes at various lags, and thus confirm the test results.

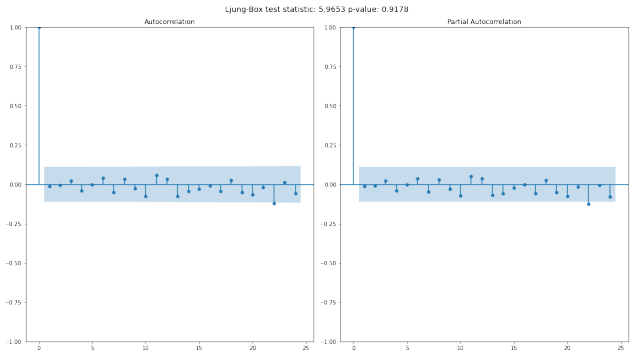


Fig. 10. Ljung-Box test for univariate SARIMA(2,1,2)x(1, 0, 1, 12) for h=1

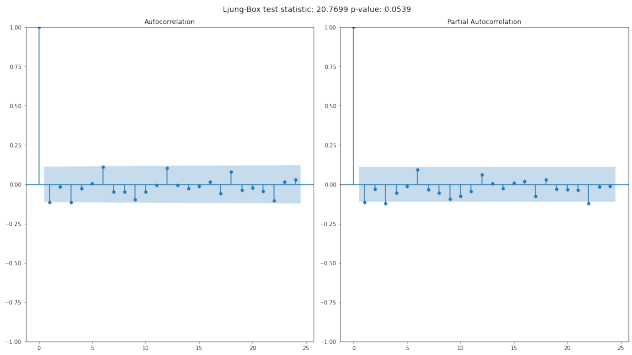
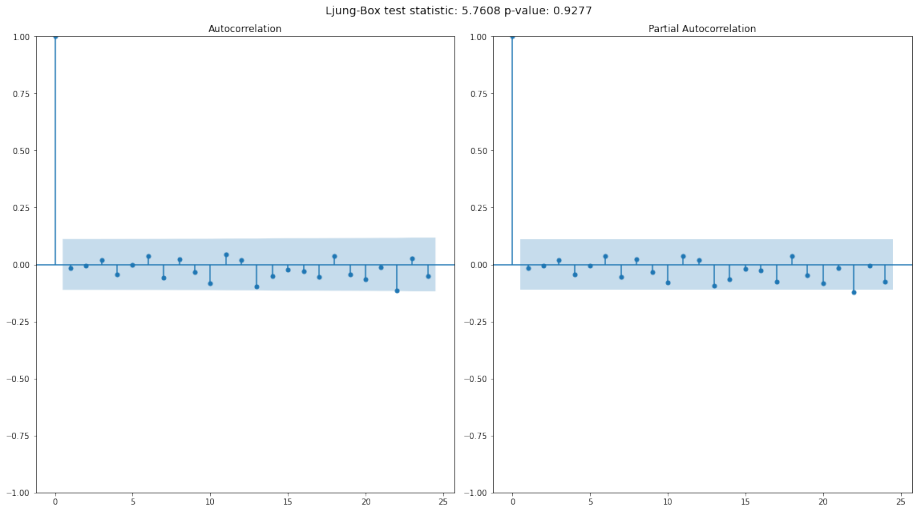
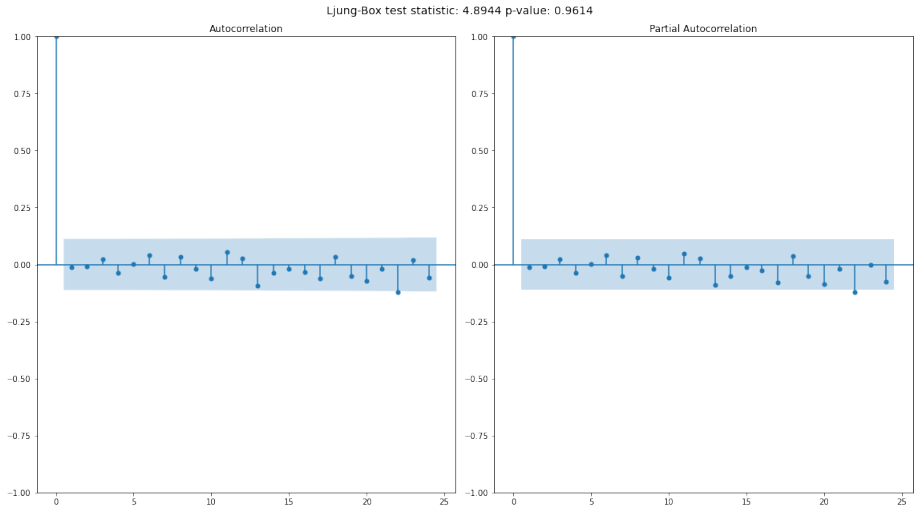


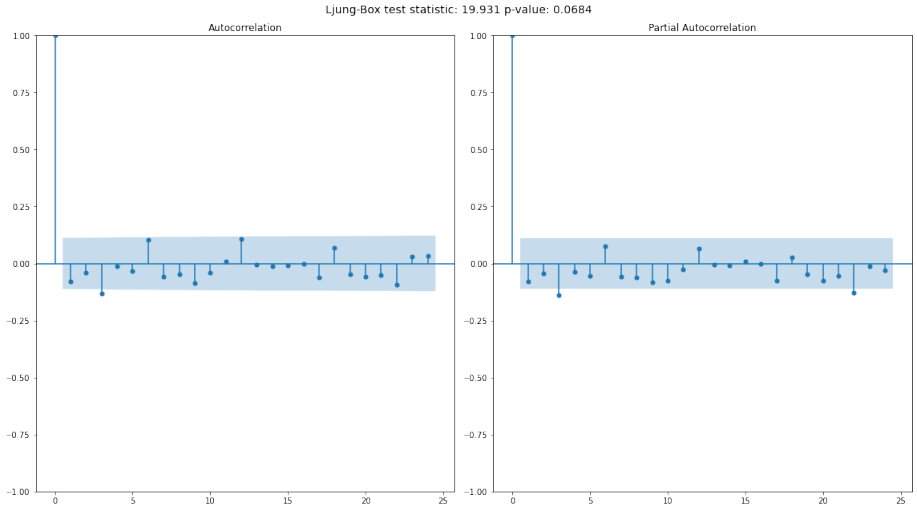
Fig. 11. Ljung-Box test for multivariate ARIMAX(2,1,5) for h=1



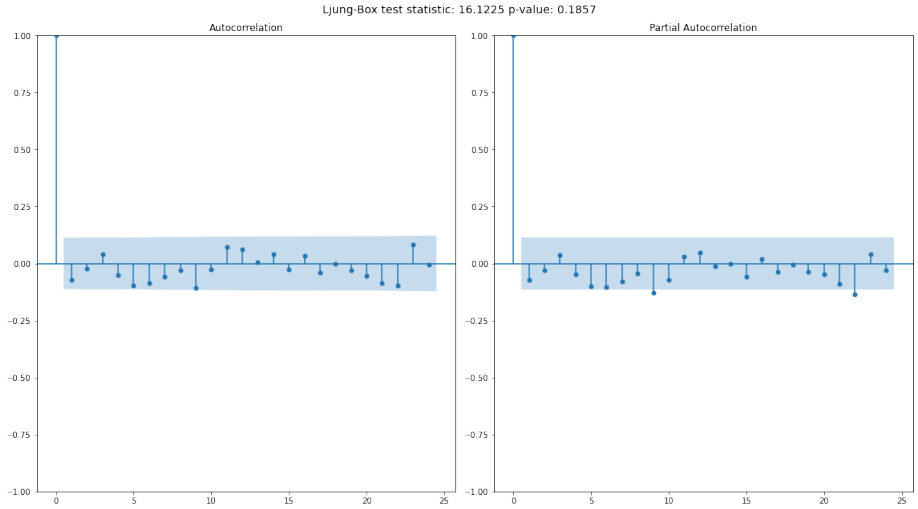
(a) Univariate SARIMA(2,1,2)x(1, 0, 1, 12) for h=3



(b) Univariate SARIMA(2,1,2)x(1, 0, 1, 12) for h=6



(c) Multivariate ARIMAX(2,1,5) for h=3



(d) Multivariate ARIMAX(4,1,5) for h=6

Fig. 12. Ljung-Box test results for ARIMA models both for univariate and multivariate cases for h=3 and h=6 forecasting horizons together with autocorrelation (ACF) and partial autocorrelation (PACF) function plots.

Finally, the data for model training were split into training and test datasets, where the split date was January 2023.

3.4. Univariate forecasts

Univariate model test results are presented in tables 9 and 10. In these models Lithuanian HICP values are forecasted, excluding any covariates. Here *h* refers to the number of forecast horizons (1, 3 or 6 months ahead). The evaluation metrics RMSE and SMAPE were calculated both for the first differences and differences inverted to the real HICP values. Looking at the results, it is clear that the linear models outperform the transformer models in the short-term forecasting horizons (h=1 and h=3). Naturally, both RMSE and SMAPE errors increase as the forecast horizon increases from 1 to 3 months. However, as the look-ahead period reaches 6 months, transformer forecasts have a lower error than linear ones.

Among linear models, ARDL has a lower forecast error than SARIMA (2,1,2) (1,0,1,12) in forecasting monthly price level changes in Lithuania. As the figure 13 highlights, the SARIMA model (a) overestimates the real HICP values in the three-month forecasts, although it reflects all jumps and plummets as in real values. Meanwhile, ARDL model (b) predicts values close to the real forecasts, but fails to address any HICP fluctuations.

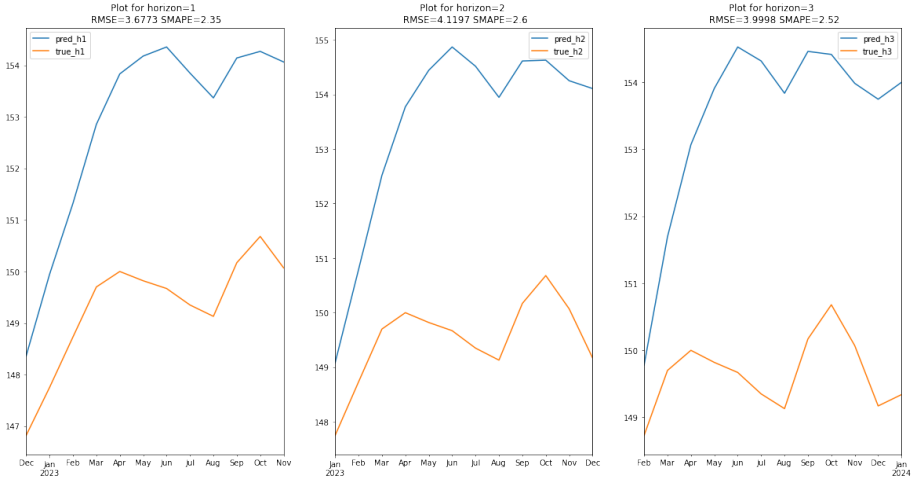
Among transformer models, Temporal Fusion Transformers (TFT) are more accurate than other types of transformer models in short-term (h=1 and h=3) forecasts. As the forecast horizon extends up to half a year, both the encoder and the TFT models share similar evaluation metric scores. Compared with linear models, it is interesting that transformer models tend to underperform real HICP values (see (c) and (d) in 13)

Table 9. Univariate linear model results.

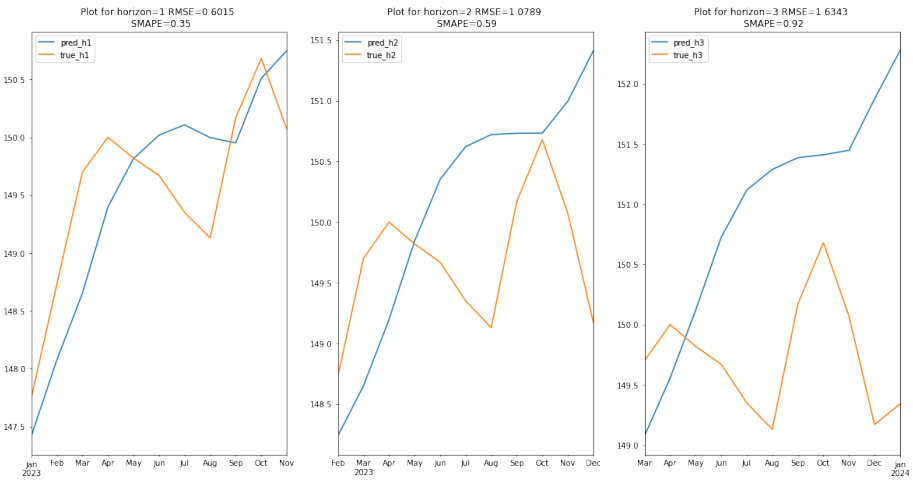
			SARIMA(2,1,2)x(1, 0, 1, 12)	ARDL
horizon=1	**RMSE**	**real (h=1)**	2.4196	**0.7746**
	SMAPE	**real (h=1)**	1.53	0.43
horizon=3	**RMSE**	**real (h=1)**	3.6773	**0.6014**
		real (h=2)	4.1196	**1.0788**
		real (h=3)	3.9997	**1.6342**
	SMAPE	**real (h=1)**	2.35	0.35
		real (h=2)	2.6	0.59
		real (h=3)	2.52	0.92
horizon=6	**RMSE**	**real (h=1)**	2.2880	**0.6558**
		real (h=2)	2.2719	**0.9295**
		real (h=3)	2.2932	**1.2026**
		real (h=4)	2.2204	**1.8128**
		real (h=5)	2.7231	**2.0829**
		real (h=6)	3.1221	**1.8975**
	SMAPE	**real (h=1)**	1.5	0.39
		real (h=2)	1.43	0.54
		real (h=3)	1.44	0.69
		real (h=4)	1.36	1.07
		real (h=5)	1.71	1.27
		real (h=6)	1.99	1.13

Table 10. Univariate transformer model results.

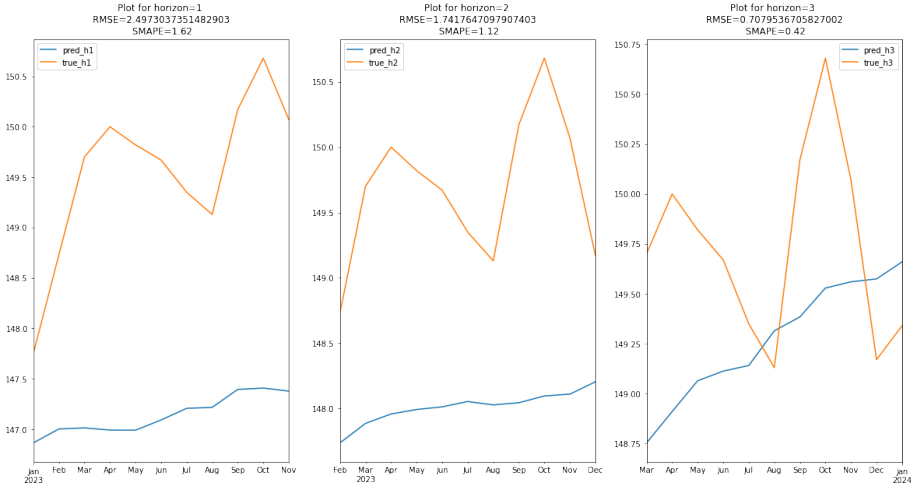
			Attention layer	Encoder layer	Transformer	TFT
horizon=1	**RMSE**	**real (h=1)**	1.7576	2.6771	2.6760	**1.4048**
	SMAPE	**real (h=1)**	1.09	1.74	1.74	0.83
horizon=3	**RMSE**	**real (h=1)**	2.4973	2.7456	2.7459	**2.3734**
		real (h=2)	1.7417	1.9168	1.9165	**1.4689**
		real (h=3)	0.7079	1.0297	1.0298	**0.7284**
	SMAPE	**real (h=1)**	1.62	1.78	1.78	1.55
		real (h=2)	1.12	1.24	1.24	0.89
		real (h=3)	0.42	0.62	0.62	0.43
horizon=6	**RMSE**	**real (h=1)**	2.3228	**2.044**	2.4781	1.8136
		real (h=2)	1.8678	**1.3911**	1.8109	1.1363
		real (h=3)	1.0809	**0.7305**	1.1125	0.6322
		real (h=4)	0.4455	**0.5524**	0.4555	0.8682
		real (h=5)	0.7109	**0.9163**	0.5872	1.132
		real (h=6)	0.6666	**0.8575**	0.5547	0.9705
	SMAPE	**real (h=1)**	1.5	1.32	1.61	1.16
		real (h=2)	1.22	0.89	1.18	0.66
		real (h=3)	0.67	0.43	0.68	0.37
		real (h=4)	0.23	0.29	0.25	0.47
		real (h=5)	0.4	0.5	0.32	0.65
		real (h=6)	0.38	0.47	0.33	0.49



(a) SARIMA(2,1,2)x(1, 0, 1, 12)



(b) ARDL



(c) Attention layer

(d) Temporal Fusion Transformer

Fig. 13. Univariate model forecasts for three months ahead after inverting first differences values to absolute HICP values.

3.5. Multivariate forecasts

The results of multivariate model forecasts are highlighted in tables 11 and 12. Evaluation metrics RMSE and SMAPE were calculated both for first differences and differences inverted to real HICP values. Out of three linear models (VAR, ARIMAX and ARDL), VAR model visibly outcompetes the other two models in one month forecasts. However, when it comes to forecasting HICP values three and six months ahead, ARIMAX models are more accurate. Apparently, adding covariates to ARIMA model improved its forecasting capacity, compared to univariate forecasts.

Among transformer models, absolute HICP forecasts alone for one month ahead were most accurately predicted by encoder model (RMSE 1.08). Subsequently, as forecasts were made three months ahead, full transformer models, consisting of encoder and decoder layers, had the lowest RMSE scores, compared to alternative models. However, short-term forecasts fall short in accuracy compared to linear models. However, likewise to univariate forecasts, as the forecasting horizon covers six months, TFT models once again prove more accurate than any linear model. Consequently, the results suggest that transformer model lack accuracy in short-term (1-3 months) forecast in comparison with classical forecasting models, but make up in medium-term forecasts.

Three- and six-month multivariate forecasts are displayed in figure 14. What stands out in the graphs is that ARIMAX model addresses fluctuations in real HICP values, while transformer models (c) and (d) are less predictive of shocks in monthly HICP values.

The results univariate and multivariate forecasts suggest several points. Firstly, classical linear models, such as ARIMA, VAR and ARDL, are more accurate in short-term (up to 3 months) forecasts. However, as the forecasting horizon increases (in this case, up to 6 months), transformer models, in particular Temporal Fusion Transformer, are more precise than linear models. It means that transformers stand as a suitable alternative in medium- and long-term forecasts. Secondly, multivariate forecasts have better evaluation metrics than univariate ones. This finding could induce further research into the leading indicators of Lithuania's or other European countries' HICP. In our models, HICPs from other European countries together with two Lithuania's economic indicators (unemployment rate and Producer Price Index) and dummy variables were used in the research. As the literature review revealed, factors impacting consumer prices in Europe are still not sufficiently researched.

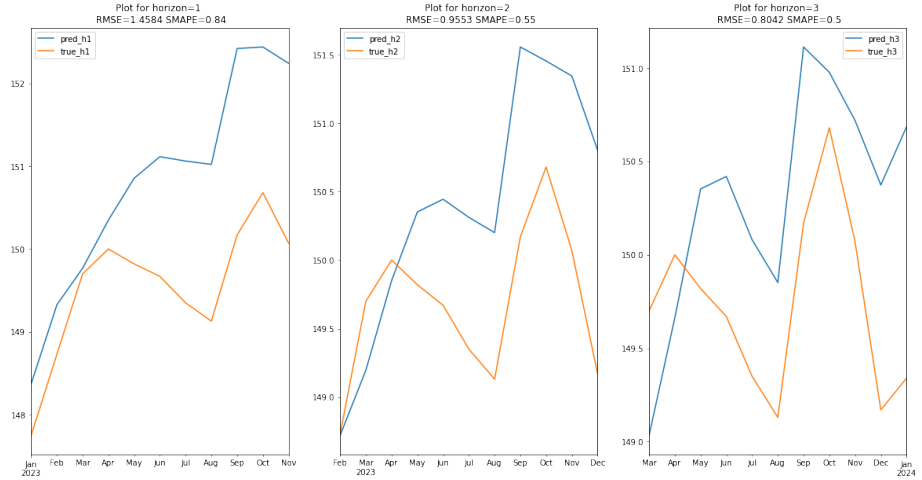
Thirdly, in the original article by Lim et al., where Temporal Fusion Transformers were introduced, the transformer achieved state-of-the-art results. The model defeated ARIMA models in electricity, occupancy rate in traffic, product sales and stock volatility multi-horizon forecast tasks. Nevertheless, it is important to note that the model in the original paper was trained on large datasets with hourly or daily samples ranging from 100 to 500 thousand observations. Meanwhile, the economic data is usually available on a monthly, if not quarterly or yearly, basis. Therefore, in domains, where data granularity is low, classical linear models would be still a more preferred alternative to neural network models, such as transformers. Fourthly, transformers were trained to make predictions on 3 month lags. The choice was made to create a level playing field with linear models which, according to their information criteria, required only 3 lags for forecasting. Given more lags, for instance, 6 or 12 months, transformers could perhaps provide even more accurate forecasts.

Table 11. Multivariate linear model results.

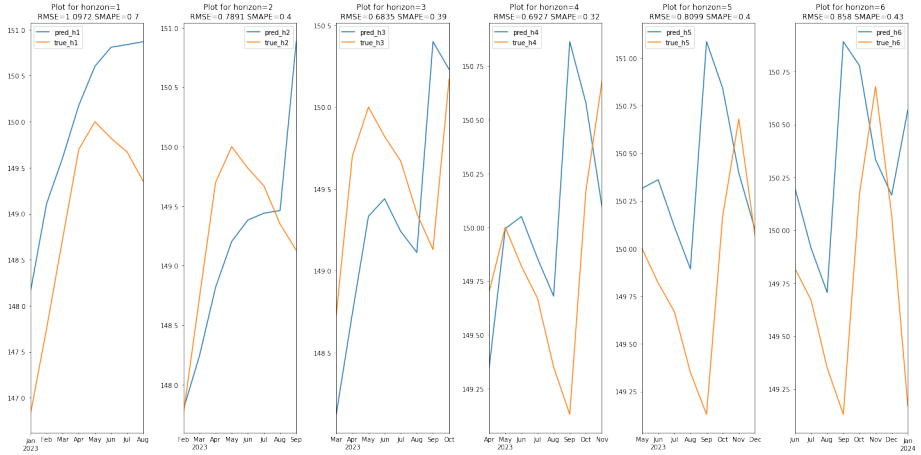
			VAR	ARIMAX(2,1,5)	ARDL
horizon=1	**RMSE**	**real (h=1)**	**0.6757**	1.7462	1.1794
	SMAPE	**real (h=1)**	0.39	1.0	0.7
horizon=3	**RMSE**	**real (h=1)**	0.5274	**1.4583**	1.1201
		real (h=2)	1.9229	**0.9553**	4.0218
		real (h=3)	3.4882	**0.8042**	2.5536
	SMAPE	**real (h=1)**	0.31	0.84	0.65
		real (h=2)	1.06	0.55	2.55
		real (h=3)	1.94	0.5	1.47
			VAR	**ARIMAX(4,1,5)**	**ARDL**
horizon=6	**RMSE**	**real (h=1)**	0.3853	**1.0971**	0.3343
		real (h=2)	0.7626	**0.7891**	0.8071
		real (h=3)	1.5568	**0.6834**	0.8129
		real (h=4)	2.5584	**0.6926**	1.1106
		real (h=5)	3.5479	**0.8098**	1.711
		real (h=6)	4.3036	**0.858**	2.0028
	SMAPE	**real (h=1)**	0.22	0.7	0.2
		real (h=2)	0.41	0.4	0.46
		real (h=3)	0.85	0.39	0.47
		real (h=4)	1.46	0.32	0.64
		real (h=5)	2.09	0.4	1.01
		real (h=6)	2.52	0.43	1.21

Table 12. Multivariate transformer model results.

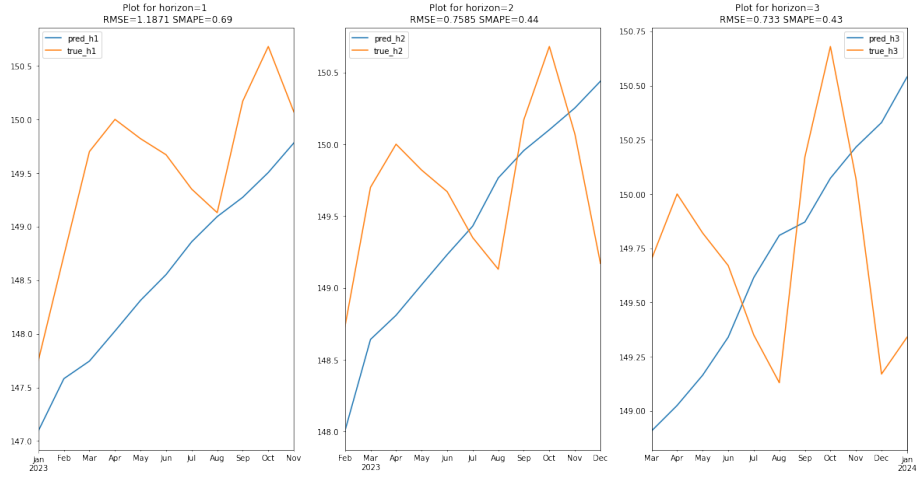
			Attention layer	Encoder layer	Transformer	TFT
horizon=1	**RMSE**	**real (h=1)**	1.4408	**1.0800**	1.2320	1.4757
	SMAPE	**real (h=1)**	0.86	0.61	0.7	0.9
horizon=3	**RMSE**	**real (h=1)**	1.5473	1.5052	**1.1870**	1.6475
		real (h=2)	1.0266	0.8653	**0.7584**	1.5654
		real (h=3)	1.0348	0.8971	**0.7329**	1.0183
	SMAPE	**real (h=1)**	0.95	0.94	0.69	1.03
		real (h=2)	0.61	0.5	0.44	0.97
		real (h=3)	0.53	0.48	0.43	0.52
horizon=6	**RMSE**	**real (h=1)**	1.5964	1.8553	1.9774	**0.802**
		real (h=2)	1.0441	1.1402	1.0228	**0.6455**
		real (h=3)	0.5106	0.5579	0.6136	**0.8574**
		real (h=4)	1.0072	0.7081	0.754	**0.9579**
		real (h=5)	1.3173	1.2829	1.4536	**0.5969**
		real (h=6)	1.03	1.05	0.9	**0.85**
	SMAPE	**real (h=1)**	1.03	1.05	0.9	0.85
		real (h=2)	0.99	1.18	1.26	0.44
		real (h=3)	0.6	0.7	0.58	0.38
		real (h=4)	0.3	0.33	0.34	0.47
		real (h=5)	0.57	0.38	0.42	0.5
		real (h=6)	0.77	0.76	0.86	0.31



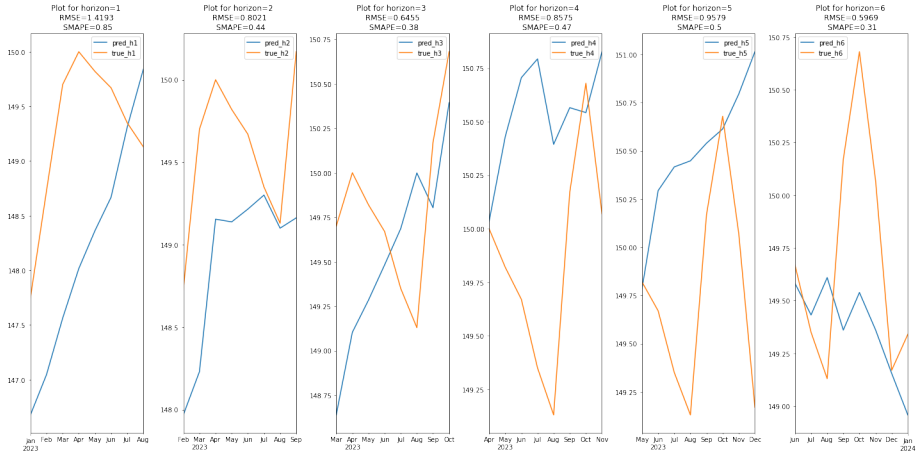
(a) ARIMAX(2,1,5) for h=3



(b) ARIMAX(4,1,5) for h=6



(c) Full transformer model (encoder + decoder layers) for h=3



(d) TFT forecasts for h=6

Fig. 14. Multivariate model forecasts.

3.6. Understanding Temporal Fusion Transformer Forecasts

On a separate note, this section will provide an example of transformer prediction interpretations. As [20] elaborates, LIME and SHAP tools, the latter tool having been previously used to select country features, are not suitable to explain time series forecast models. In the meantime, TFT models leverage their architecture in two ways. Firstly, the **multi-head attention** evaluates the **importance of past values**. Secondly, **Variable Selection Network** layer explains **feature importance**. Therefore, we will take advantage of TFT model architecture to explain its predictions for the test period: January, 2023, - January, 2024.

The figure 15 explains the last six-month ahead predictions for the test dataset. The attention score (gray line) peaks at the third lag. It suggests that the model considers the third lag at most in making the future predictions. Because Lithuania's HCIP values do not vary much across three lags, the model forecasts stable values for the next 6 months, whereas in reality they fluctuated.

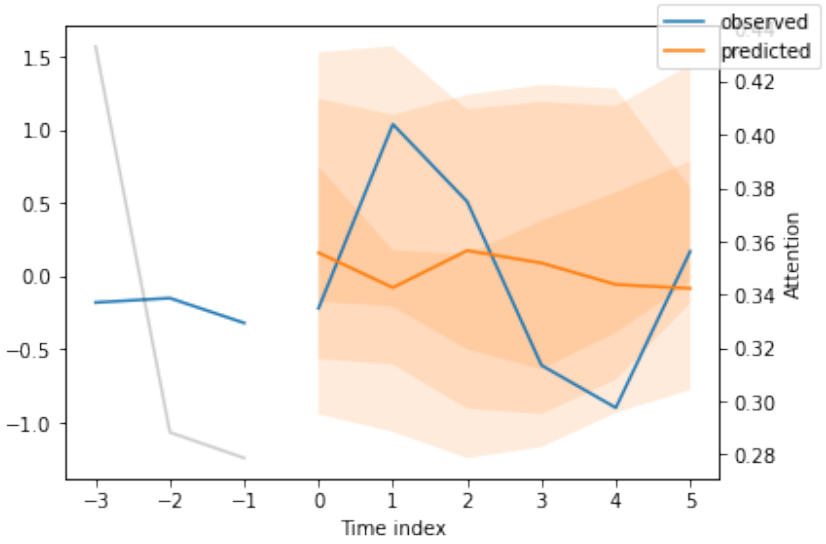


Fig. 15. TFT prediction for August, 2023 (time idx 0), - January, 2024 (time idx 5), period. Gray line represents attention scores to time lags.

Features importances for encoder and decoder layers for the test dataset are displayed in figure 16. It is apparent that the model places high importance on time related features, in this case, reported months. Similarly, the encoder layer pays more attention to Lithuania's harmonised consumer and producer price index values and their rolling averages to make future predictions. While for us this fact is not surprising (past values of the variably clearly impact its future values), it shows that the model managed itself to find an autoregressive relation between Lithuania's HCIP past and the target values. Furthermore, for decoder the most important variable is the three-month rolling average of Lithuania's unemployment rate. It again confirms the underlying idea behind Phillips curve that price levels (or their inflation rates) are related to the unemployment in general. On the other hand, the explanation of important features related to HCIPs in other European countries (Croatia for encoder or Czechia for decoder) require deeper economist insights.

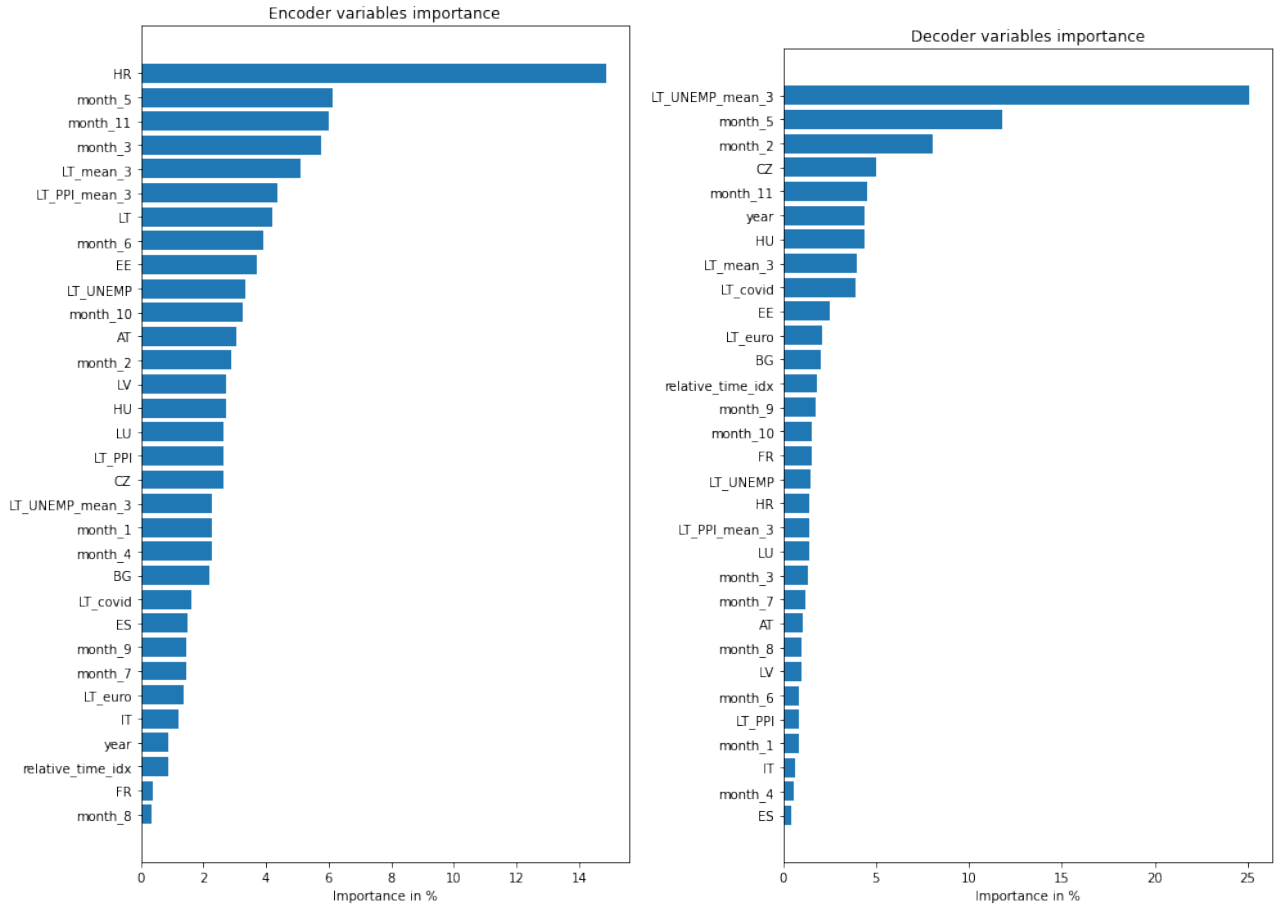


Fig. 16. Feature importances for Temporal Fusion Transformer Encoder and Decoder

3.7. Discussion and recommendations

The analysis of feature importance shows that Temporal Fusion Transformers could be leveraged to understand complex relations among multivariate features used in model training, their impact, and how previous time lags impact future forecasts. Therefore, TFT models could address the issues raised by [25], when time series machine learning model forecasts lack interpretations in their decisions.

In the big data era, consumer price indices are increasingly likely to be calculated even on a higher frequency. [9] recalls the Dutch experience, where in the 1990s the authorities began discussions to calculate CPIs using transaction data from supermarkets. If the country relied on 29 price collectors in 2000 to collect information about product and service prices in shopping outlets, 20 years later there were no more shop visits, and data for CPI primarily have been weekly collected from the databases of data providers, i.e. supermarket chains. With the advent of Covid-19 pandemic, another data collection method is being implemented: Web scraping the major retailer websites. Scraping allows daily data collection and reporting Consumer Price Indices potentially on a daily basis. As we already know, Temporal Fusion transformers benefit from training on abundant data. It explains why transformer forecasts in short-term were less accurate than ARIMA, VAR, and ARDL models. Lithuania has been cooperating with major retailers on scanner data since 2020 [28]. If consumer price indices were published each day with the help of innovative data collection methods, transformer models could produce more accurate forecasts weeks, if not months ahead, and defeat traditional

linear models also in short-term forecasts, as was the case in [26] findings. More real-time data and precise forecasts would allow governments to monitor inflation trends in the economy and take swift preventive actions, for instance, by adjusting interest rates (according to [2], at the moment the European Central Bank decides them for the euro zone every six weeks) or setting price caps in critical scenarios at shorter notice.

Conclusion

1. The literature review revealed that Consumer Prices Indices are predominantly forecasted by researchers using univariate models, because of two reasons. Firstly, in essence, covariate variables cannot always explain the variation in the rate of change of price levels. Secondly, multivariate models failed to produce as accurate forecasts as univariate ones. Among the reviewed authors, the ARIMA models outperformed other linear or neural network models.

2. The aim of this experiment was to compare the forecast accuracy between linear and transformer models in both univariate and multivariate settings. Transformers use a self-attention mechanism, according to which data points in time series could receive different importance in forecasting the future values. The forecasting capabilities of transformers were tested using different alternatives of the models: including only an attention layer; including encoder layer; including full transformer layer, consisting of encoder and decoder; Temporal Fusion Transformer model, which addresses weaknesses of transformer models in time series forecasting tasks and suits multi-horizon forecasts.

3. Harmonised Consumer Price Indices from 29 European countries for the period 1998-2024 together with Unemployment and Price Producer Index in Lithuania and two dummy variables data were used to forecast Lithuanian Harmonised Consumer Price Index. In order to reduce the number of variables, four elaborated feature selection techniques were used by training several ensemble and linear models: feature importance, Recursive Feature Elimination, Permutation Feature Importance, and Shapley additive explanation. As a result, 11 countries were selected for multivariate models: Latvia, Estonia, Spain, France, Hungary, Luxembourg, Austria, Croatia, Italy, Bulgaria, and Czechia. In addition, a few more variables were created following feature engineering: year, month, three month rolling means of Harmonised Consumer Price Index, unemployment rate as well as Price Producer Index.

4. The forecasting capabilities of different transformer alternatives were compared with three linear models: ARIMA, VAR, as well as ARDL. It was found that classical models are more precise in short-term (1-3 months) forecasts, whereas transformers have higher accuracy once the forecasting horizon expands to 6 months. Therefore, it is deduced that transformers could be taken advantage in medium- and long-term forecasts. Moreover, multivariate models in all forecasting horizons proved to defeat univariate models, thus inducing further research on factors impacting Harmonised Consumer Price Indices. Due to multi-head attention mechanism applied in the Temporal Fusion Transformer architecture, it is possible to obtain information about feature importance in the forecasting models. In Lithuania's case, the analysis of the most important features highlighted that time related features, past Lithuania's HICP values and unemployment rates were important for encoder and decoder in future forecasts.

References

[1] Marcos Álvarez-Díaz and Rangan Gupta. "Forecasting US consumer price index: does nonlinearity matter?" In: *Applied Economics* 48.46 (Mar. 2016), pp. 4462–4475. DOI: 10.1080/00036846.2016.1158922. URL: https://doi.org/10.1080/00036846.2016.1158922.

[2] European Central Bank. *3. Central bank interest rates*. URL: https://www.euro-area-statistics.org/digital-publication/statistics-insights-money-credit-and-central-bank-interest-rates/bloc-3a.html?lang=en.

[3] Oren Barkan et al. "Forecasting CPI inflation components with Hierarchical Recurrent Neural Networks". In: *International Journal of Forecasting* 39.3 (July 2023), pp. 1145–1162. DOI: 10.1016/j.ijforecast.2022.04.009. URL: https://doi.org/10.1016/j.ijforecast.2022.04.009.

[4] Christoph Bergmeir and José Manuel Benítez. "On the use of cross-validation for time series predictor evaluation". In: *Information sciences* 191 (May 2012), pp. 192–213. DOI: 10.1016/j.ins.2011.12.028. URL: https://doi.org/10.1016/j.ins.2011.12.028.

[5] Mike Bernico. *Deep Learning quick reference*. Packt Publishing, Mar. 2018.

[6] Anastasia Borovykh, Sander Bohte, and Cornelis W. Oosterlee. "Conditional Time Series Forecasting with Convolutional Neural Networks". In: - (Sept. 2018). URL: https://arxiv.org/abs/1703.04691.

[7] George E. P. Box et al. *Time Series analysis*. John Wiley Sons, June 2015.

[8] N. Carnot, V. Koen, and B. Tissot. *Economic forecasting*. Springer, Aug. 2005.

[9] Antonio Chessa. "Theory and Practice of HICP: Index calculation with scanner data". In: Mar. 2022.

[10] Sibanjan Das and Umit Mert Cakmak. *Hands-On Automated Machine Learning: A beginner's guide to building automated machine learning systems using AutoML and Python*. Packt Publishing, 2018. ISBN: 1788629892.

[11] Eralda Gjika Dhamo, Llukan Puka, and Oriana Zaçaj. "FORECASTING CONSUMER PRICE INDEX (CPI) USING TIME SERIES MODELS AND MULTI REGRESSION MODELS (ALBANIA CASE STUDY)". In: *10th International Scientific Conference "Business and Management 2018"* (2018). URL: https://api.semanticscholar.org/CorpusID:158594578.

[12] Weng Dong-Dong. "The Consumer Price Index Forecast Based on ARIMA Model". In: Aug. 2010. DOI: 10.1109/icie.2010.79. URL: https://doi.org/10.1109/icie.2010.79.

[13] Eurostat. "Glossary:Consumer price index (CPI)". In: Aug. 2018. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Consumer_price_index_(CPI).

[14] Eurostat. *Harmonised Index of Consumer Prices (HICP) METHODOLOGICAL MANUAL*. Tech. rep. 978-92-68-12009-5. Jan. 2024. DOI: 10.2785/055028.

[15] David Foster. *Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play*. O'Reilly Media, Incorporated, Mar. 2023. ISBN: 9781098134181.

[16] Paul Goodwin. *Profit from Your Forecasting Software: A Best Practice Guide for Sales Forecasters*. Mar. 2018. URL: https://www.amazon.com/Profit-Your-Forecasting-Software-Forecasters-ebook/dp/B07BNR8KBY.

[17] Sanjeev Gupta and Sachin Kashyap. "Forecasting inflation in G-7 countries: an application of artificial neural network". In: *Foresight* 17.1 (Mar. 2015), pp. 63–73. DOI: 10.1108/fs-09-2013-0045. URL: https://doi.org/10.1108/fs-09-2013-0045.

[18] IBM. *What is a Transformer Model?* URL: https://www.ibm.com/topics/transformer-model.

[19] Emmanuel Jurczenko. *Machine learning for asset management*. June 2020. DOI: 10.1002/9781119751182. URL: https://doi.org/10.1002/9781119751182.

[20] Nikos Kafritsas. "Temporal Fusion Transformer: Time Series Forecasting with Deep Learning — Complete Tutorial". In: (Feb. 2024). URL: https://towardsdatascience.com/temporal-fusion-transformer-time-series-forecasting-with-deep-learning-complete-tutorial-d32c1e51cd91.

[21] Mindaugas Kieža. *Pagrindines vartotojų kainas sąlygojančių veiksnių nustatymas Lietuvoje.* 2011.

[22] Evžen Kočenda and Alexandr Černý. *Elements of Time Series econometrics: an Applied Approach*. Karolinum Press, Jan. 2015.

[23] Anders Kock and Timo Teräsvirta. "Forecasting the Finnish Consumer Price Inflation using Artificial Neural Network Models and THree Automated Model Selection Techniques". In: *Finnish Economic Papers* 26.1 (Jan. 2013), pp. 13–24. URL: http://www.taloustieteellinenyhd
fi/wp-content/uploads/2014/09/fep12013_kock_and_terasvirta.pdf.

[24] Francesca Lazzeri. *Machine Learning for Time Series Forecasting with Python*. John Wiley Sons, Dec. 2020.

[25] Bryan Lim and Stefan Zohren. "Time-series forecasting with deep learning: a survey". In: *The Royal Society Publishing* 379.2194 (Apr. 2020), p. 20200209. DOI: 10.1098/rsta.2020.0209. URL: https://doi.org/10.1098/rsta.2020.0209.

[26] Bryan Lim et al. "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting". In: *International journal of forecasting* 37.4 (Oct. 2021), pp. 1748–1764. DOI: 10.1016/j.ijforecast.2021.03.012. URL: https://doi.org/10.1016/j.ijforecast.2021.03.012.

[27] State Data Agency of Lithuania. *Suderinto vartotojų kainų indekso sudarymo metodika*. Tech. rep. DĮ-300. Dec. 2023. URL: https://osp.stat.gov.lt/documents/10180/11840852/SVKI_metodika_2023-12.pdf/401d57a4-c5a3-4780-bc36-dc0f320d2081.

[28] State Data Agency of Lithuania. *Vartotojų Kainų Indekso Sudarymo Metodika*. Tech. rep. DĮ-372. Dec. 2021. URL: https://osp.stat.gov.lt/documents/10180/5118910/Vartotoj%C5%B3+kain%C5%B3+indeksai+%28VKI%29%2C+kain%C5%B3+poky%C4%8Diai%2C+svoriai%2C+vidutin%C4%97s+kainos+%5BLT%5D+605.html.

[29] Monika Marazaitė. *The impact of the Russian invasion of Ukraine on Lithuania's Harmonised Index of Consumer Prices*. 2023. URL: https://talpykla.elaba.lt/elaba-fedora/objects/elaba:192842955/datastreams/MAIN/content.

[30] Paul D. McNelis and Peter McAdam. "Forecasting Inflation with Thick Models and Neural Networks". In: *Social Science Research Network* (Jan. 2004). DOI: 10.2139/ssrn.533014. URL: https://doi.org/10.2139/ssrn.533014.

[31] Ana Naveckienė and Žilvinas Kalinauskas. "Application of vector autoregression model for Lithuanian inflation / Ana Čuvak, Žilvinas Kalinauskas." eng. In: *Ekonomika ir vadyba-2009 = Economics and management-2009 : tarptautinės mokslinės konferencijos programa ir santraukų rinkinys* (2009), p. 48.

[32] Nhu-Ty Nguyen and Thanh-Tuyen Tran. "Mathematical development and evaluation of forecasting models for accuracy of inflation in developing countries: A case of Vietnam". In: *Discrete dynamics in nature and society* 2015 (Jan. 2015), pp. 1–14. DOI: 10.1155/2015/858157. URL: https://doi.org/10.1155/2015/858157.

[33] Tien-Thinh Nguyen et al. "The consumer price index prediction using machine learning approaches: Evidence from the United States". In: *Heliyon* 9.10 (Oct. 2023), e20730. DOI: 10.1016/j.heliyon.2023.e20730. URL: https://doi.org/10.1016/j.heliyon.2023.e20730.

[34] Ebrahim Pichka. "What are Query, Key, and Value in the Transformer Architecture and Why Are They Used?" In: (Oct. 2023). URL: https://towardsdatascience.com/what-are-query-key-and-value-in-the-transformer-architecture-and-why-are-they-used-acbe73f731f2.

[35] Jeff Ralph, Rob O'Neill, and Joe Winton. *A practical introduction to index numbers*. John Wiley Sons, June 2015.

[36] Robert J. Rossana. *Macroeconomics*. Feb. 2011. DOI: 10.4324/9780203829271. URL: https://doi.org/10.4324/9780203829271.

[37] Ajibola Salami. *Cross Correlation with Two Time Series in Python*. Feb. 2022. URL: https://www.datainsightonline.com/post/cross-correlation-with-two-time-series-in-python.

[38] Emmanuel Sirimal Silva, Hossein Hassani, and Jesús Otero. "Forecasting Inflation Under Varying Frequencies". In: *Electronic Journal of Applied Statistical Analysis* 11.1 (Apr. 2018), pp. 307–339. DOI: 10.1285/i20705948v11n1p307. URL: http://siba-ese.unile.it/index.php/ejasa/article/view/16339.

[39] Geoffrey Smith. “Bank of England8217;s forecasting needs urgent revamp, says ex-US Fed chief Bernanke”. In: (Apr. 2024). URL: https://www.politico.eu/article/bernanke-boe-needs-to-open-its-mind-and-its-wallet-to-get-forecasting-right/.

[40] A. So et al. *The Data Science Workshop: A New, Interactive Approach to Learning Data Science*. Packt Publishing, 2020. ISBN: 9781838981266.

[41] Julius Stakėnas. *Forecasting Lithuanian inflation*. 2015. URL: https://www.lb.lt/uploads/publications/docs/wp_17.pdf.

[42] James H. Stock and Mark W. Watson. *Modeling inflation after the crisis*. Tech. rep. Oct. 2010. DOI: 10.3386/w16488. URL: https://doi.org/10.3386/w16488.

[43] Prabhanjan Narayanachar Tattar. *Hands-On Ensemble Learning with R: A beginner’s guide to combining the power of machine learning algorithms using ensemble techniques*. Packt Publishing, 2018. ISBN: 1788624149.

[44] Ashish Vaswani et al. “Attention is all you need”. In: *arXiv (Cornell University)* (Jan. 2017). DOI: 10.48550/arxiv.1706.03762. URL: https://arxiv.org/abs/1706.03762.

[45] Qingsong Wen et al. *Transformers in Time Series: A Survey*. 2023. arXiv: 2202.07125 [cs.LG].

[46] Dongdong Weng. “The Consumer Price Index Forecast Based on ARIMA Model”. In: *2010 WASE International Conference on Information Engineering* (Aug. 2010). DOI: 10.1109/icie.2010.79. URL: https://doi.org/10.1109/icie.2010.79.

[47] Mark Wynne and Fiona Sigalla. “The consumer price index”. In: *Federal Reserve Bank of Dallas Economic Review* 2 (1994), pp. 1–22.

[48] Richard Yamarone. *The Trader’s guide to key economic indicators*. John Wiley Sons, July 2012.

[49] Soffa Zahara, Sugianto Sugianto, and Muhammad Bahril Ilmiddaviq. “Consumer price index prediction using Long Short Term Memory (LSTM) based cloud computing”. In: *Journal of physics* 1456.1 (Jan. 2020), p. 012022. DOI: 10.1088/1742-6596/1456/1/012022. URL: https://doi.org/10.1088/1742-6596/1456/1/012022.

[50] Ailing Zeng et al. “Are Transformers Effective for Time Series Forecasting?” In: *arXiv* (Aug. 2022). URL: https://arxiv.org/abs/2205.13504.

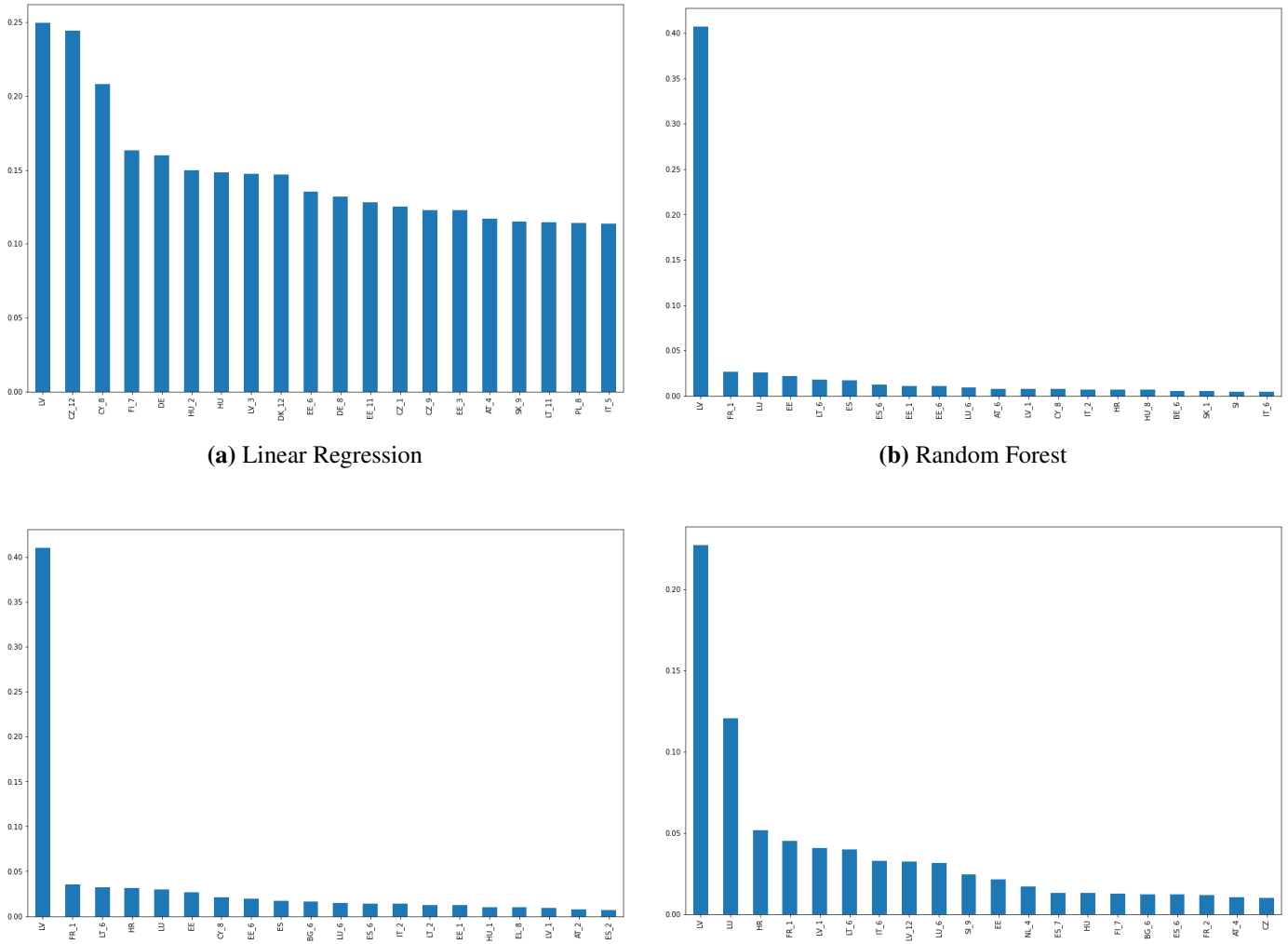
Appendices

Appendix 1. Top 20 country-lag features selected by Recursive Feature Elimination method

These 20 country-lag features were selected as the most important by Recursive Feature Elimination method:

- **Linear Regression**: LV, FI, FI_1, IT_2, HU_2, LV_3, LT_3, IT_4, AT_4, IE_6, LU_6, BE_8, DE_8, HU_8, SI_8, BE_9, CZ_12, FR_12, LV_12, PL_12
- **Random Forest**: EE, ES, HR, LV, LU, EE_1, FR_1, LV_1, IT_2, EE_6, ES_6, LT_6, LU_6, AT_6, LU_7, CY_8, HU_8, SI_8, FI_10, LT_12
- **Gradient Boosting**: EE, ES, HR, LV, LU, EE_1, FR_1, LV_1, HU_1, LT_2, LT_3, BG_6, EE_6, ES_6, IT_6, LT_6, LU_6, CY_8, LT_11, LV_12
- **XGBoost**: CZ, EE, HR, LV, LU, FR_1, LV_1, AT_1, FR_2, AT_2, HR_4, NL_4, BG_6, IT_6, LT_6, LU_6, FI_7, IT_10, IE_12, LV_12

Appendix 2. Top 20 country-lag features selected by feature importance method



(a) Linear Regression

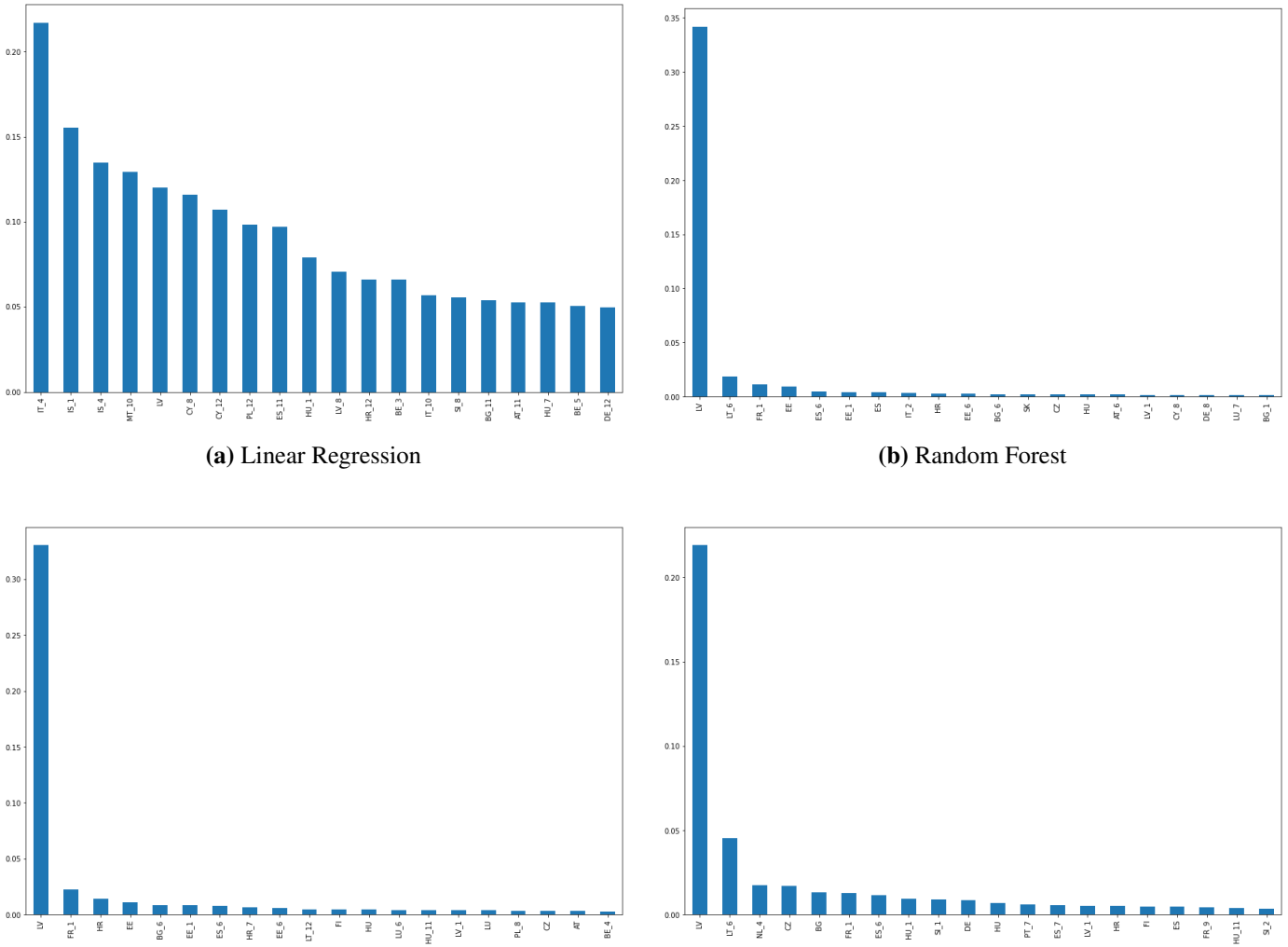
(b) Random Forest

(c) Gradient Boosting

(d) XGBoost

Fig. 17. Country-Lag Feature Selection by Feature Importance

Appendix 3. Top 20 country-lag features selected by Permutation Feature Importance method



(a) Linear Regression

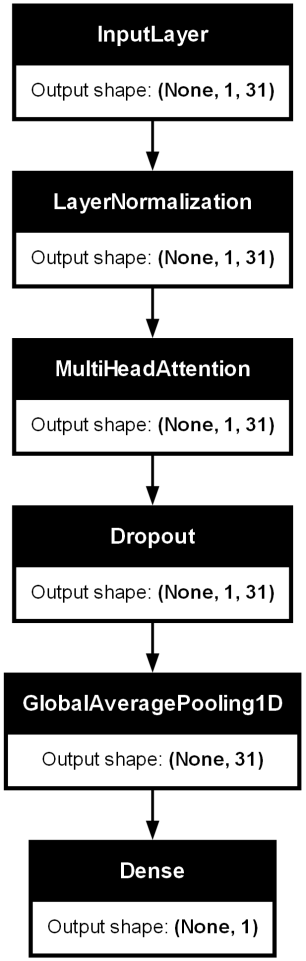
(b) Random Forest

(c) Gradient Boosting

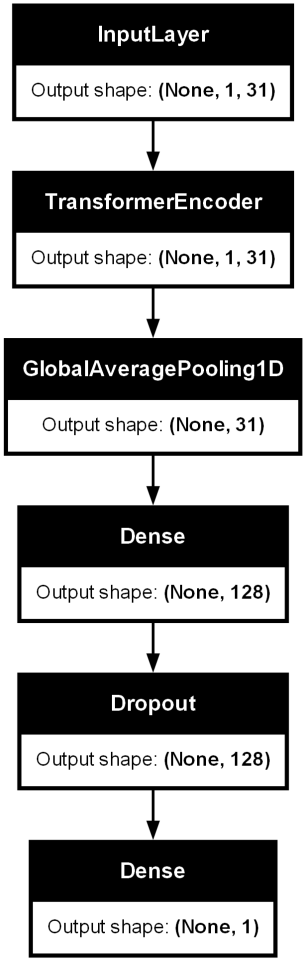
(d) XGBoost

Fig. 18. Country-Lag Feature Selection by Permutation Feature Importance

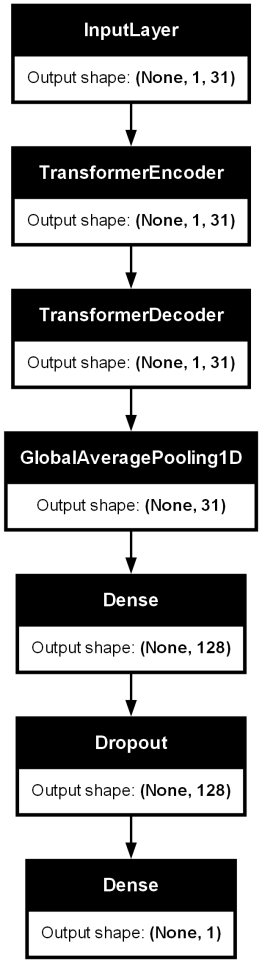
Appendix 4. Transformer model architectures



(a) Transformer model architecture including only Attention Layer.



(b) Transformer model architecture including Encoder Layer.



(c) Transformer model architecture for full transformer model with encoder and decoder Layers.

Fig. 19. Transformer model architectures