INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Department of Quantitative Methods for Management and Economics/ Department of Information Science and Technology

Machine learning models to predict electricity consumption and the impacts of COVID-19 in Portugal

Ana Sofia Lobato Sucena

Master in Data Science

Supervisor:

PhD, Anabela Ribeiro Dias da Costa, Assistant Professor

ISCTE-IUL

Supervisor:

PhD, Diana Elisabeta Aldea Mendes, Associate Professor

ISCTE-IUL

BUSINESS SCHOOL

TECNOLOGIAS E ARQUITETURA

Department of Quantitative Methods for Management and Economics/ Department of Information Science and Technology

Machine learning models to predict electricity consumption and the impacts of COVID-19 in Portugal

Ana Sofia Lobato Sucena

Master in Data Science

Supervisor:

PhD, Anabela Ribeiro Dias da Costa, Assistant Professor

ISCTE-IUL

Supervisor:

PhD, Diana Elisabeta Aldea Mendes, Associate Professor

ISCTE-IUL

Dedication to my Family, Teachers, and Friends.

Acknowledgments

I want to express my sincere gratitude to my parents because, without their unconditional support and love throughout my life, I would not be who I am today or have accomplished what I have.

I'd also like to express my gratitude to my brother and friends, who have always been there for me no matter what.

The professors Anabela Costa and Diana Aldea Mendes, my thesis supervisors, deserve a special thank you for their patience, understanding, and positive attitude throughout the entire process.

Resumo

Esta tese analisa a forma como os modelos de "*machine learning*" e os dados provenientes de fontes de dados públicas podem ser utilizados para prever o consumo de eletricidade em Portugal. Boas previsões são cruciais para uma gestão eficiente do setor energético, nomeadamente devido ao aumento da procura global de energia. Portugal apresenta um ótimo caso para a previsão de consumo, uma vez que depende significativamente de importações de energia e sofre de pobreza energética.

O estudo utiliza uma metodologia baseada em dados para analisar doze anos de padrões de consumo energético e analisar a forma como a pandemia do COVID-19, os padrões climáticos e o PIB afetam o consumo de eletricidade. Foram estudados cinco modelos preditivos - SARIMA, SARIMAX, VAR, SVR e LSTM – e os seus indicadores de desempenho em dois períodos diferentes (um para os doze anos de analise, incluído durante o Covid-19, e outro apenas para dados antes do Covid-19). Assim, este estudo permite avaliar a prestação dos modelos de machine learning em periodos estáveis e não estáveis

O estudo reconhece as suas limitações, como a falta de dados na era pós-COVID, mas continua a fornecer informações úteis para o desenvolvimento e a gestão de políticas energéticas.

Palavras-chave: Consumo de eletricidade, Pandemia COVID-19, Modelos preditivos

Abstract

This thesis analyzes how data from public data sources and machine learning models can be used to forecast electricity consumption in Portugal. Accurate forecasts are crucial for efficient energy management, given the rising global demand for energy. Portugal presents a compelling case for consumption projections since it significantly relies on energy imports and suffers from poverty.

The study uses a data-driven methodology to analyze twelve years of consumption patterns and examine how the COVID-19 pandemic, weather patterns, and GDP affect electricity use. Five predictive models were studied - SARIMA, SARIMAX, VAR, SVR and LSTM - and their performance indicators in two different periods (one for the twelve years of analysis, including during Covid-19, and the other only for data before Covid-19). Thus, this study makes it possible to evaluate the performance of machine learning models in stable and non-stable periods.

The study acknowledges its limitations, such as the lack of data in the post-COVID era, while providing valuable insights for developing and managing energy policies.

Keywords: Electricity consumption, Machine learning, COVID-19 pandemic, Predictive models

Index

Tables Index

Figures Index

Equations Index

Abbreviations

ADF - Augmented Dickey-Fuller

AIDS - Acquired Immunodeficiency Syndrome

AIC - Akaike Information Criterion

ANN - Artificial Neural Network

ARMA - Autoregression Analysis and Moving Average

ACF - Autocorrelation Function

BIC - Bayesian Information Criterion

BNN - Bayesian Neural Network

CART - Classification and Regression Trees

CMCC - Fondazione Euro-Mediterraneo Sui Cambiamenti Climatici

DF - Dickey-Fuller

EU - European Union

FPE - Akaike's Final Prediction Error

GDP - Gross Domestic Product

GWH - Gigawatt Hours

GRNN - Generalized Regression

HIV - Human Immunodeficiency Virus

HQIC - Hannan–Quinn Information Criterion

IEA - International Energy Agency

KNN - K-Nearest Neighbor Regression

KPSS - Kwiatkowski–Phillips–Schmidt–Shin

KWh - Quilowatt Hour

LR - Linear Regression

LSTM - Long Short-Term Memory

MAE - Mean Absolute Error

MLE - Maximum Likelihood

MLP - Multi-Layer Perceptron

MSE - Mean Squared Error

NECP - National Energy and Climate Plans

NNLS - Non-Linear Least Squares

OECD - Organization for Economic Cooperation and Development

OLS - Gaussian Ordinary Least Squares

PACF - Partial Autocorrelation Function

PHEIC - Public Health Emergency of International Concern

PIB - Produto Interno Bruto

PP - Philips Perron

RBF - Radial Basis Function

REN - Redes Energéticas Nacionais

RMSE - Root Mean Square Error

SARIMA - Seasonal Autoregressive Integrated Moving Average

SARIMAX - Seasonal Auto-Regressive Integrated Moving Average with Exogenous Factors

SDGs - Sustainable Development Goals

SVM - Support Vector Machine

SVR - Support Vector Regression

US - United States

USD - US Dollar

VIF - Variance Inflation Factor

WHO - World Health Organization

XAI - Explainable Artificial Intelligence

CHAPTER 1

Introdution

With the rapid development of the economy and society, economic growth was accompanied by a large amount of energy consumption (Nwulu & Agboola, 2012). Worldwide energy consumption is rising fast because of increased human population, continuous pressures for better living standards, emphasis on large-scale industrialization in developing countries, and the need to sustain positive economic growth rates (OECD, 2011). Electricity has become one of the most important forms of energy to man, being one of the primary forms of energy that modern life is built upon. It affects a society's quality of living and efficiency (Kavaklioglu, 2011) and its versatility has led to an almost limitless set of applications, from transport, heating, lighting, communications, etc.

Most recent state-of-the-art shows that the world's energy consumption increased rapidly between 1995 and 2015, going from 8,588.9 to 13,147.3 million tons (Mtoe) (Ahmad et al., 2020; Dong et al., 2020). All reputable organizations anticipate a sharp increase in energy consumption shortly due to a rapid change in the demand for coal to power production and industry (Ahmad et al., 2020).

All recent European efforts have been to completely decarbonize the economy by the year 2050, a goal that depends on energy efficiency and electrification of the energy supply (Hank et al., 2020). Portugal is a country that is heavily dependent on energy imports (78%), with oil and petroleum products constituting the primary energy source (68%), mainly used in the transport sector, where electrification will be gradually increased until 2050 (IEA, 2022). Other decarbonized options, such as biofuels and synthetic (green) fuels, cannot be completely ruled out, and they should be considered for the decarbonization targets to be met in 2050.

One of COVID-19's notable effects is on the world's climate, which will partially improve due to the reduction in energy usage (Saadat et al., 2020). Critical fluctuations in energy and consumption requirements have been caused by the pandemic's altered productive proportion of Gross Domestic Product (GDP) output.

Given this, a sound forecasting technique is essential for accurate investment planning of energy production/ generation and distribution. Identifying adequate and essential information for a decent prediction is a common challenge in developing accurate forecasts. If the information level is insufficient, the forecasting will be poor; similarly, if the information is useless or redundant, modelling will be difficult or skewed (Kaytez, 2020).

The research questions guiding this thesis investigation are focused on comprehending and predicting Portugal's electricity consumption while taking the COVID-19 pandemic's effects into account through machine learning. These investigations include looking at past consumption patterns, analyzing the effects of pandemics, figuring out how exogenous factors affect consumption, and assessing the suitability of various machine learning models.

The thesis is divided into five chapters, each of which focuses on a different area of the research.

The introductory chapter provides an overview of the research problem and its importance.

The chosen methodology, a thorough literature review, and Portugal's distinctive energy landscape are all covered in the Literature Review chapter.

The data utilized in this thesis is discussed, analyzed, and, when necessary, transformed in the Data Analysis chapter. This is important in order to comprehend the data's patterns, trends, and correlations.

The models used (SARIMA, SARIMAX, VAR, SVR, and LSTM) are discussed, and the stationarity of the data is evaluated, the models are then compared in the Modeling chapter, allowing the best model to be chosen.

The thesis concludes with a summary of findings, and recommendations for future research in the area.

CHAPTER 2

Literature Review

2.1. Methodology

Over the past 15 years, the number of papers published on forecasting electric power has increased exponentially (Vivas et al., 2020). The studies are typically site-specific, and the outcomes heavily depend on the type of model used, the length of the prediction, and a variety of other attributes of the data and models. This significant constraint makes it challenging to generalize the findings (Enders et al., 2015).

In order to identify the main streams of relevant literature, a multi-step approach was followed to identify articles of scientific value. First, the search strategy was determined by conceptualizing the topic, "Prediction of electricity consumption using machine learning". SCOPUS was used for this analysis, limiting the search to only documents in the "final" publication stage and written in English. The exclusion criteria was used to select papers that related electricity consumption to economic growth and COVID-19. Title and abstract screening was done, and all the documents irrelevant to the theme were excluded. Additionally, the references of the analyzed articles were screened to identify further articles of relevance.

An overview of how the scope was ultimately focused in Figure 2.1:

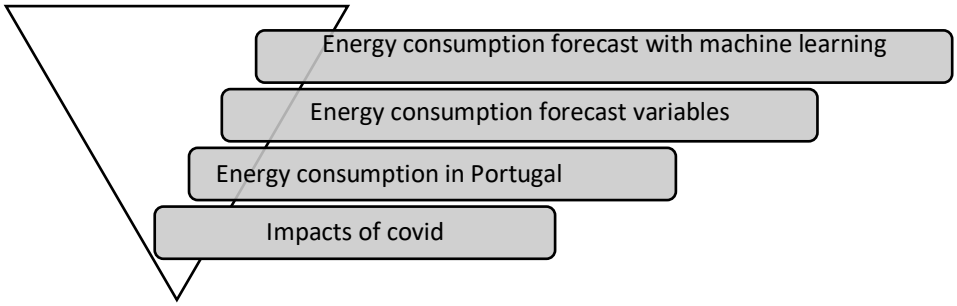

Figure 2. 1 - Research scope

2.2. Why energy forecast is important

The Sustainable Development Goals (SDGs) set by the United Nations were signed by nations from both developing and growing economies, officially known as the Agenda for Global Development of Sustainable Development Goals (SDGs) 2030 and Corporations (United Nations, 2015; Omer and Noguchi, 2020; Ahmad and Zhang, 2020). Energy has drawn much attention as a key pillar for ensuring

sustainable and inclusive economic development for every country (Pietrosemoli and Rodrguez-Monroy, 2019).

The amount of power used is a major concern for the electricity sector regarding planning and strategic considerations. Because in contrast to fuels, electricity is difficult and pricy to store (Vishnuraja, et al, 2020), because it is economically non-storable, and the stability of the power system needs a continuous balance between the production and consumption of electricity (Weron, 2014). Therefore, over the years, it has become vital for accurate and efficient mechanisms to model and predict energy consumption. (Nwulu & Agboola, 2012)

Energy forecasting and planning are also important in energy sector development and policy formulation (Ahmad & Chen, 2020). Forecasting electricity use can assist policymakers in making effective plans for economic growth. This is possible through energy conservation by avoiding excessive electricity consumption through enhanced operational strategy (Khan et al., 2020). Allows energy distribution companies to predict their customers' energy needs and provide accordingly. By, for example, giving them information about upcoming periods of high energy demand or authorizing them to tailor their tariffs to customer consumption, consumption prediction enables those companies to improve their processes. (Vishnuraja et al., 2020)

It is commonly recognized that countries' consumption levels are growing by the day, yet the resources required to produce energy are restricted in both quantity and range. Furthermore, the global distribution of power generation is unbalanced, with some places having more significant energy reserves in quantity and range than others. Such an unequal distribution of energy resources is true not only for reserves but also for consumption levels, and there is an imbalance in the use of energy resources around the world, not only between regions but also between countries (Pradeep, K., & Likhita, T., 2022). When an economy's energy resources run out, it has two options: accept "poor economic growth with current power resources" or strive to "enhance growth by using imports to meet the uncovered portion for the need of the power resources". Energy costs impact countries' production patterns, and budget balances, whether direct or indirect. As a result, ensuring the economy's performance and contributing to long-term growth requires sufficient purchases of power resources at a low cost (Pradeep, K., & Likhita, T., 2022).

Examining economic aspects of energy consumption, mainly income and price elasticities, is very important for policymakers. These factors can help policymakers understand two important concepts: how, historically, the relationship between electricity use and its economic drivers has changed and how projected changes in these drivers will affect future consumption. This information would be helpful to demand-side policymakers for managing energy consumption using tools such as energy prices, tax rates and tariffs. It could also be useful for supply-side policymakers in planning the level of energy supply required to meet future demand adequately (Atalla and Hunt, 2016; Hasanov et al., 2016a). The protection of the environment would also require benefits from this information. (Hasanov, 2021)

2.3. Relationship between energy consumption and economic growth

The empirical research on the connections between energy use and economic growth can be inconclusive and cannot be used to make universally applicable policy recommendations (Shahbaz et al., 2015).

However, as the driving force of modernization, energy has a very close and essential relationship with economic growth. Energy consumption promotes economic growth, making it possible to improve energy utilization efficiency (Alam & Murad, 2020). Meanwhile, energy exhaustion and the environmental problems caused by energy consumption also restrict economic growth.

So, looking at the relationship between energy consumption and economic growth is important. Despite the fact that the "GDP growth-energy relationship" study (Martins et al., 2022) showed a clear connection between GDP growth and energy consumption, it has been possible to observe a decoupling between economic growth and energy consumption over the past few decades while also observing an increase in the electrification of the economy. Electricity power supply, the basic industry in almost every country and region, can affect economic development. On the one hand, economic growth depends strongly on the electricity power supply; the power supply shortage would negatively affect economic growth. On the other hand, the fast growth of the economy would also stimulate the power consumption demand. The common situation is that the amount of electricity power consumption of a certain city or even a region always has a close relationship with the economic development of the corresponding area. Therefore, interpreting the characteristics of previous regional electricity power consumption and past economic data, and finding the interrelation between them, would be of great reference value in doing the short-term forecasting work together for both. The combined forecasting work will undoubtedly be more rational and accurate in giving future electricity power consumption tendencies, compared with the work merely based on electricity power consumption data. (Li et al., 2020).

When using multivariate time sequence cointegration theory and the established Vector autoregression (VAR/VEC) model to analyze the long-term equilibrium relationship among GDP, energy consumption, and its components, it was shown by Zhang et al., in 2017, that cointegration between GDP and energy consumption and between GDP and electricity consumption may exist. Namyan & Popp (2012) analyzed panel data of 93 countries from 1980-2006, and found that the increasing energy consumption would reduce GDP in the long term. Xiaoli & Haoran (2013) studied the non-linear relationship between energy consumption and economic growth based on the Panel Smooth Transition Regression (PSTR) model, which showed that energy consumption would be more sensitive to economic growth changes with the economy's further growth.

The dynamic relationship between energy consumption, electricity consumption, and economic growth shows that GDP maintains a positive response when facing the disturbance of energy consumption or electricity consumption, which shows that economic growth largely depends on energy consumption.

From a long-term perspective, a two-way causality exists between energy consumption and economic growth, which shows that high energy consumption is closely related to economic growth. Then, the regression equation shows that global economic growth is important for promoting energy consumption. The steady growth of the global economy will inevitably require more energy consumption.

From the short-term perspective, if global energy consumption and economic growth deviate from the long-term equilibrium relationship, it will be adjusted at the speed of 83.49%. The adjusting speed is fast, which shows that the long-term equilibrium relation between energy consumption and economic growth is stable. (Wang et al., 2018)

2.4. Impacts of financial crises (covid-19)

The connection between energy and the economy in Europe is crucial, critical, and frequently complex. It isn't easy to properly incorporate connections between energy prices, growth, competitiveness, and employment, as well as relationships between the structure of the energy industry and energy pricing (World Economic Forum, 2012).

There are several ways that an infectious disease might impact the economy (i.e., direct and indirect). Estimates of the macroeconomic effects of the 2003 SARS outbreak revealed that despite having a relatively low number of deaths and cases, it did have a significant impact on economies by sharply reducing the consumption of various goods and services, increasing business operating costs, and reassessing the risks of a region that resulted in higher risk premiums (Lee and McKibbin, 2004). Siu and Wong (2004), Hai et al. (2004), Chou et al. (2004), and Bloom et al. (2005) calculated the potential economic impact of highly contagious diseases with no vaccine (such as pandemic influenza, HIV/AIDS, and SARS) (Aktar et al., 2021). It is abundantly clear that COVID-19 is a highly contagious and economically devastating pandemic.

The outbreak of the COVID-19 pandemic has led to a significant economic shock to the world that was never experienced for decades. The Organization for Economic Cooperation and Development (OECD) and the World Trade Organization have indicated that the COVID-19 pandemic is the largest warning for each nation globally since the 2008–2009 global financial crisis (Sruthi, 2021). Some experts even claim that since World War II, the world has not experienced such an unusual state of emergency (Chakraborty and Maity, 2020).

Furthermore, it created a lot of uncertainty for the global electrical market because several nations have imposed restrictions to stop the virus's spread. This has affected economies globally. Although countries have implemented different strategies to prevent the spread of the virus, lockdowns have become the inevitable end for all countries. Except for basic necessities, all economic sectors in Europe have experienced an abrupt closure due to an unusual fall in demand and a decline in the supply of goods. The energy sector is not an exception; historical consumption patterns have changed as people's and industries' daily routines have significantly changed; transportation is kept to a minimum; and curfews have increased the importance of cycling and walking (Hauser et al., 2021).

For European countries, Bahmanyar et al. (2020), determined that different lockdown measures and their population activities in countries significantly changed the electricity consumption profile.

Electricity demand has been significantly reduced worldwide due to lockdown measures being imposed and followed, leading to Global electricity demand decreased by 2.5%. Industries were shut down during this time, and individuals were urged to stay inside for their safety. As a result, domestic power consumption increased while industrial power usage declined. (Gulati et al., 2021)

Ruan et al. (2020), analyzed the impact of COVID-19 on the US electricity sector. In the entire US market, a decrease of 6.36–10.24% in April and 4.44–10.71% in May was observed in electricity consumption. It also concluded that a reduction in electricity consumption is associated with an increase in COVID-19 cases and the duration of stay-at-home.

Prol and Sungmin (2020), examined the effect of COVID-19 on electricity consumption in the EU countries and USA states. In this study, it was observed that electricity consumption decreased by 3–12% in 5 months.

Carvalho et al. (2020), examined the impact of the COVID-19 pandemic on Brazil's electricity consumption models and consumption between January 1 and May 27, 2020. Electricity consumption varied by region, and while a decrease of 7% was observed in the residential area, a decrease of 20% to 14% was observed in other regions.

Ghiani et al. (2020), reported that electricity consumption in Italy decreased by 37% compared to the previous year. Halbrügge et al. (2021), analyzed the effects of the COVID-19 outbreak on the electricity sector in Germany and Europe. It was observed that after COVID-19 lockdowns started, less electricity was consumed in Germany, France, Italy, Spain, and Sweden than in prior years.

Many of the improvements recognized for 2020 are feared to be temporary since they do not correspond to long-term, systemic changes in the economy, transportation, or energy systems. Emissions will increase after the lockdowns are lifted, closed factories reopen, and the world economy is back to normal. (McCloskey and Heymann, 2020).

Since the effects of COVID-19 are changing day by day, it is to be noted that the results are likely to change when the same analyses are performed for different periods (Gulati et al., 2021).

2.5. Portugal energy panorama

Portugal has overcome the protracted economic slump it endured after the 2008 global crisis. GDP increased to USD 340 billion in 2019 from USD 325 billion in 2008 before the financial crisis. As opposed to 7.6% in 2008, the unemployment rate decreased from a peak of 16.2% in 2013 to 6.5% in 2019. With total final consumption per GDP falling by 8% between 2014 and 2019, Portugal is demonstrating signs of decoupling economic development from energy demand, accelerating ongoing structural reforms away from energy-intensive businesses.

Additionally, Portugal has attained a high degree of electrification. In 2019, electricity met 25% of the world's ultimate energy demand, 56% of the demand for energy in buildings, and 25% of the demand for energy in industry. Portugal has also attained significant levels of renewable energy, which in 2019 accounted for 30.6% of the country's total final energy demand. Renewable energy sources accounted for 54% of electricity output, primarily due to hydropower and wind generation, while bioenergy is widely used in buildings and industry. However, recent years have seen only a modest increase in renewable energy.

The proportion of renewable energy in gross final energy demand increased by 3.8% between 2014 and 2019. To reach Portugal's 2030 goals, a sustained deployment of renewable energy is required in every sector. In 2019, imports of fossil fuels (43% oil, 24% natural gas, and 6% coal) comprised 76% of Portugal's primary energy supply. Coal, natural gas, and all types of oil are imported.

The Covid-19 epidemic had a significant impact on Portugal, causing its GDP to shrink by 8.4% in 2020—the most since 1936. Portugal and the European Union (EU) have made significant efforts to alleviate the pandemic's effects and promote the resumption of economic growth.

2.6. Portugal Energy Panorama Challenges

Portugal has issues with energy poverty because of its frequently inadequately insulated building stock and relatively high energy prices. The EU Energy Poverty Observatory stated in 2018 that 19.4% of Portugal's population reported being unable to maintain a sufficient level of heating in their homes (the EU average was 7.3%) and also noted difficulties with cooling.

The government has established social tariffs for electricity and natural gas that offer discounts on portions of the distribution tariffs to lower households' electricity and gas bills that meet certain socioeconomic criteria. The government places a high priority on energy affordability. In December 2020, 34 709 homes (2.4% of all households connected to the gas network) and 752 965 households (14% of all households) got the natural gas social tariff.

Portugal heavily relies on energy imports due to the strong demand for fossil fuels (mainly oil and natural gas) and the scarcity of native fossil resources. Portugal had one of the highest rates of energy import dependency among IEA members in 2019, at 74%. By increasing the percentage of renewable energy sources in its energy supply, particularly for electricity, Portugal has made headway in lowering its reliance on energy imports.

The National Energy and Climate Plans (NECP) and “Roteiro para a Neutralidade Carbónica 2050” (RNC2050) set aggressive goals for reducing energy import dependence to under 65% by 2030 and under 19% by 2050, respectively. Solid and long-lasting steps must be taken to reduce the demand for fossil fuels, particularly in the transportation sector, where oil accounted for 94% of energy demand in 2019, as well as in the industrial sector, where oil and natural gas combined for 51% of energy demand in 2019.

2.7. Energy Consumption Forecasting Techniques

The planning of energy forecasting models and their real-time applications started in 1960 (Nguyen, 2005). Scholars mainly analyzed the relationship between energy consumption and economic growth at the national level, and most of them used the classical econometric theory. Since the 1990s, the

investigation of the link between non-stationary time series has been supported theoretically by the joint adoption of cointegration theory and the error correction model. Since then, the technology in the energy forecasting research area has also developed rapidly with advances in forecasting theory and machine learning. (Ahmad & Chen, 2020)

The electrical sector is becoming decarbonized, decentralized, and digitalized, increasing the importance of artificial intelligence and data analytics. Factors such as the declining costs of information and communication technology and the advances in computing power lead to rising data availability and new opportunities for analysis. Additionally, the growing share of renewable energy sources in the electricity system and the rise in the number of active actors in the electricity system adds complexity and new needs for data analytics (Scheidt et al., 2020). With these opportunities, further studies have emerged in the energy sector, namely, electricity.

In the recent past, numerous prediction models have been applied. The goals of the model network mechanism, the available data, and the energy planning operation are typically considered when choosing a prediction model (Ahmad et al., 2020). But no one-size-fits-all approach exists. The most appropriate method always depends on the context. Scheidt et al. (2020), propose some guiding questions that researchers can ask themselves to find a suitable approach: What kind of data is available? Is time series data used? How volatile is it? How important is the interpretability of results? Can a hybrid method that combines several strategies help address the problem's various characteristics? What level of computing complexity is acceptable? These factors enable a suitable Data Science approach to be customized to each use case.

The range of applications for energy forecasting models is pretty broad. Benchmarking forecasting models are helpful for energy markets, electricity pricing scenarios, ancillary service markets, market regulation, and price balancing. (Klæboe, Fleten & Eriksrud, 2015, Ahmad & Chen, 2020).

Modern time series forecasting techniques primarily rely on historical future predictions. The uniqueness of the energy consumption indicators is the existence of multidirectional trends, seasonal and cyclical fluctuations, and structural breaks, which impose specific criteria for selecting appropriate methodologies and models. The complexity, interpretability, and forecasting precision of the systems for extrapolating past data to the future constantly increases (Kalimoldayev et al., 2020).

Traditionally, regression analysis has been the most popular modeling technique in predicting energy consumption (Tso et al., 2007). Technology advancements prompted the active development of machine learning forecasting methodologies, even though statistical techniques based on the Gaussian Ordinary Least Squares (OLS), non-linear least squares (NNLS), and maximum likelihood (MLE) estimation are still widely utilized (Kalimoldayev et al., 2020). As the advancement in artificial intelligence is getting bigger, these techniques can handle large amounts of data efficiently and effectively and address the diverse nature of electrical load (Gulati et al., 2021). Because they can extract features from the series and incorporate them into the models without specifying the parameters, as with typical statistical methods, a variety of artificial intelligence techniques that are now in use showed good experimental results (Kalimoldayev et al., 2020). These include Artificial Neural Networks (ANNs), Fuzzy Logic Systems, Multi-Layer Perceptron (MLP), Bayesian Neural Network (BNN), Generalized Regression Neural Networks (GRNN), K-Nearest Neighbor regression (KNN), Classification and Regression Trees (CART), Support Vector Machine (SVM) (Bishop, 2006).

These models and hybrid systems overcome the problems of irregularity and complexity in modern energy systems (Sehgal & Pandey, 2015). They are primarily used in real-time applications and achieve higher forecasting accuracy (Debnath & Mourshed, 2018; Ahmad & Chen, 2020). By combining various models, ensemble approaches allow us to improve forecasting accuracy. If used correctly, artificial neural networks can be a robust choice since they can extract and model unseen relationships and features (Ahmad et al., 2020).

For a better understanding of the most common machine learning models, it is given a brief definition of them. The Support Vector Machines (SVM) model forecasts (Sapankevych and Sankar, 2009) based on support vectors divide into two streams, one relying on classification and assignment of some current state to the cluster representative and the other relying on model fitting with support vectors regression. The Random Forests (Abuella and Chowdhury, 2017) randomly sample the training set to provide inputs to tree-like models and combine them to give the final result. Forecasting based on random forests exploits the average of the output of the individual tree models to provide the final result. Deep learning (Abdulaziz and Edwards, 2017) can use Long Short-Term Memory networks with more than three layers. These recurrent neural networks learn long-term dependencies between time steps of sequential data. Hence, the feature vector represents a generic "sequence" from which the system learns the relation with the target data. A linear ARMA model identifies and follows time-series trends. The input is the observed time series rather than sparse features (Luis, Esteves & Da Silva, 2020)

As an example of comparison between these models, Ahmad & Chen (2020), studied the daily consumption data of a dwelling composed of two people. These data made it possible to know that the Linear Regression and Support Vector Regression obtained 85.7% accuracy, partly due to the inclusion of the previous day in the training process, with Random Forest being the model with the worst result, 79.9%. However, just because RF and SVR fit the dataset's variables (day, weekday, week, presence, and so on) better than the other models in this comparison does not imply that they are superior. For example, the Kernel-based extreme machine learning models and wavelet transforms give high forecast accuracy and shorter computation time than conventional forecasting models (Zhang & Li, 2018).

Still, numerous studies report better model fitting but worse forecasting accuracy of these methods than statistical models. Supervised-based ML approaches are susceptible to output and input datasets, and their accuracy relies on time-series models. The non-stationary time series with higher deviation amounts may lead to lower forecasting performance (Ahmad & Chen, 2020). Regarding their improved interpretability and characterization of the uncertainty around the point forecasts, the researchers note the need for improvement and continued development of machine learning models (Kalimoldaye, et al., 2020)

Since electrical energy is one of the most needed energy types worldwide and it is a reflection of the development in an economy, the consumption amount of this type of energy increases every day since electricity can be produced from many energy sources, it does not cause any harm to the environment, its area of use is wide and changes alongside with population increases, and the technology advances (Cihan, 2022). So, one way to improve the performance of these models is by having relevant variables. Many researchers have addressed several features like temperature, rainfall, humidity, holidays, population, and other possible scenarios for more accurately predicting electricity consumption (Gulati et al., 2021).

CHAPTER 3

Data Analysis

The goal of this chapter is to understand and explore the data in order to have a better comprehension of the variables to include in the predictive models. Python programming language and Jupyter Notebook - a free computational environment - were used for visualization, data analysis, and modeling.

3.1 CRISP-DM Framework

For the identification of the problem, algorithm selection, and predictive model definition, it was used the CRISP-DM framework. This stands for Cross-Industry Standard Process for Data Mining and is an industry-independent process model that defines the steps and tasks one must do to develop a successful Data Mining project. It consists of six iterative phases, from business understanding to deployment. The general schema of the CRISP-DM framework is shown in Figure 3.1.

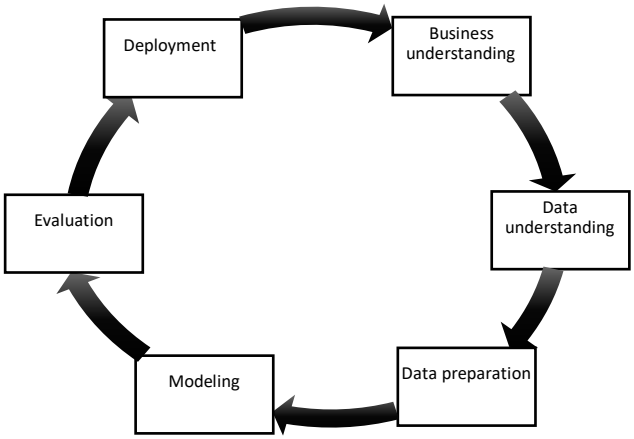

Figure 3. 1 - CRISP-DM Methodology

The first step is business understanding. This initial phase focuses on understanding the project objectives and requirements. The second phase is data understanding, which begins with initial data collecting and includes efforts to understand the data, find data quality issues, and get initial insights from the data. The following phase is data preparation, which consists of all operations related to constructing the final dataset (data that will be input into the modeling tool(s)) from the initial raw data.

Various techniques are chosen and employed throughout the modeling phase, and their parameters are adjusted to ideal values. At the project's evaluation stage, the models presented in the literature review will be evaluated for the various datasets, and a comparison will be made between them to understand which is the best model for predicting electricity consumption in Portugal. Before proceeding with the final deployment of the model, it is critical to thoroughly examine the model and review the procedures used to develop it to ensure that it meets the business objectives. One significant goal is identifying whether any critical business issues have been overlooked. Finally, in the Deployment stage, the solution is ready for deployment, with additional monitoring and maintenance. Completing and delivering the thesis will be the deployment phase in this case.

3.2 Data understanding and data preparation

This study aims to investigate the effects of various feature selection and engineering strategies on the efficacy of machine learning models in predicting electricity consumption in Portugal and the potential impact of the COVID-19 epidemic on Portugal's electricity consumption patterns. Exploratory data analysis was performed to validate the literature review, which claims that electricity consumption follows predictable patterns on an annual and monthly basis. So, in this section, these trends will be investigated.

The database is a compilation of online data from various sources, with some changes made to obtain better data analysis. The data, as well as the transformations, are as follows:

- Data on monthly electrical consumption - was retrieved from the REN - *Redes Energéticas Nacionais* - website. This company is responsible for ensuring mainland Portugal's uninterrupted electricity and natural gas supply.
    - Time period: from 2010 until 2022.
    - Transformations: Month_Year column to DateTime format and GWh transformed to KWh
- Data on annual electrical consumption by type - were retrieved from the PODATA (database on present-day Portugal with official and certified statistics on the country and Europe) website.
    - Time period: from 1994 until 2023
    - Transformations: Month_Year column to DateTime format, and the entries before 2010 were removed

- The weather data, created through collaborations between the International Energy Agency (IEA) and the Fondazione Euro-Mediterraneo Sui Cambiamenti Climatici (CMCC), found on the IEA website, showcases global data on weather-related variables.
    - Time period: monthly data from 2010 to 2022
    - Transformations: entries for countries besides Portugal were excluded. There was a wrong entry for 2013-02-01 in the Wind100int column, so the next value replaced it.
- The monthly GDP data was found on the Federal Reserve Economic Data (FRED) website.
    - Time period: monthly data from 2010 to 2022
    - Transformations: Replace column name “observation_date” with “Mês_ano” and change the type to Datetime format and column name “PRTLORSGPNOSTSAM” for “GDP”

The dataset variables and information about them are presented in Table 3.1.

Table 3. 1 - Data variables

Serie	Description	Measure Unit
KWh	Consumption refers to the net production of the central stations / per month.	KWh
Temperature	This parameter is the temperature of air at 2m above the surface of land, sea or inland waters	°C
Cloud	This parameter is the proportion of a grid box covered by a cloud. Total cloud cover is a single-level field calculated from the cloud occurring at different model levels through the atmosphere.	%
Precipitation	This parameter is the accumulated liquid and frozen water, comprising the rain and snow that falls to the earth's surface. It is the sum of largescale precipitation and convective precipitation.	mm/h
Wind100int	It is the horizontal airspeed, one hundred meters above the earth's surface.	m/s
Daylight	Minutes of sunlight.	%
Humidex	Humidity-corrected 2-metre temperature	ºC
GDP	Gross domestic product at market prices is the final result of the production activity of resident producer units.	Euro
Electricity consumption by type of consumption (kWh)	Hydroelectric, nuclear, and conventional thermal power stations produce energy from wave, tidal, wind, and solar photovoltaic sources.	kWh (kilowatt-hour)

The data collection consists of 151 instances and nine attributes after the transformations: time, KWh, Temperature, Wind100int, Daylight, Precipitation, Humidex and GDP. This data collection includes the period from 01/01/2010 to 08/01/2022.

3.3 Exploratory Analysis

On January 30, 2020, the World Health Organization (WHO) declared the Covid-19 outbreak a public health emergency of international concern (PHEIC). On March 11, 2020, the Portugal Prime Minister

decreed the closure of all public and private educational establishments from March 16 until at least April 9. The first lockdown in Portugal ended in May 2020 when the Portuguese government declared a state of emergency. The second lockdown lasted from January 15, 2021, to March 15, 2021, with slow openings and trade and service restrictions. On May 1, 2021, the Portuguese government lifted the State of Emergency and declared a Calamity throughout the country. The WHO ended its PHEIC declaration on May 5, 2023, but it is still referred to as a pandemic as of June 2, 2023. Considering this information, this study will assess the period of Covid-19 in Portugal between March 2020 and April 2021.

For this purpose, a new variable was created to analyze the impact of this pandemic in Portugal. It is a categorical variable with the value "1" when the date is between March 2020 and April 2021, representing the existence of Covid. If the variable takes the value "0", it describes the non-existence of Covid.

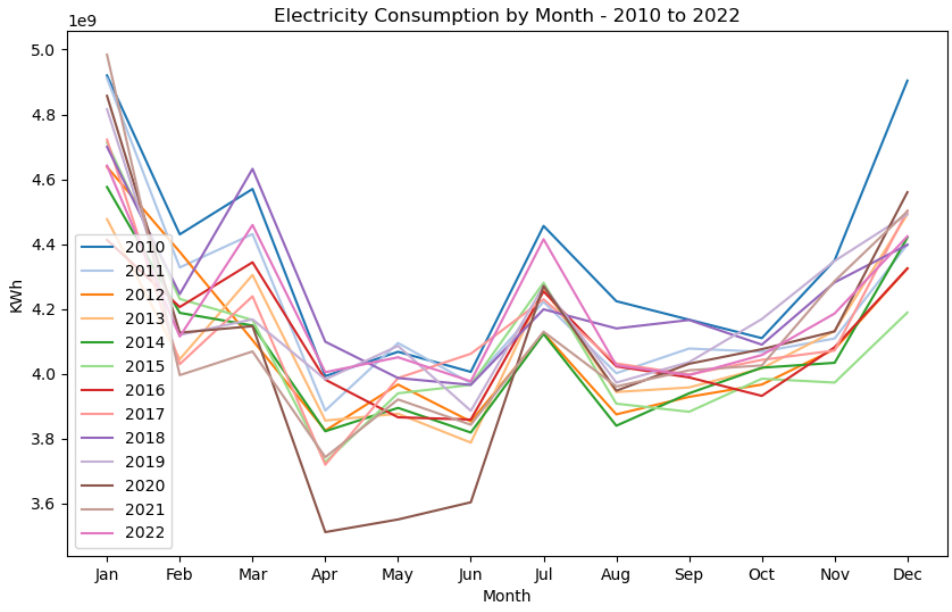

Figure 3. 2 - Electricity consumption in Portugal in the last 12 years, by year and month

Figure 3.2 shows a monthly overview of each year's electricity consumption in 2010-2022. Some conclusions about consumption patterns can be drawn.

- From November to March, consumption values are at their highest, with January having the highest consumption of all years.
- In 2010, consumption peaked in December, which did not happen in subsequent years.
- The months with the lowest electricity consumption are April through October.

- Over the past 12 years, consumption has consistently increased in July, followed by a return to lower levels in August.
- It is noticeable that in 2020, there was a significant decrease in March, and the lowest electricity consumption value occurred in April of that year. The values returned to normal in July of the same year.

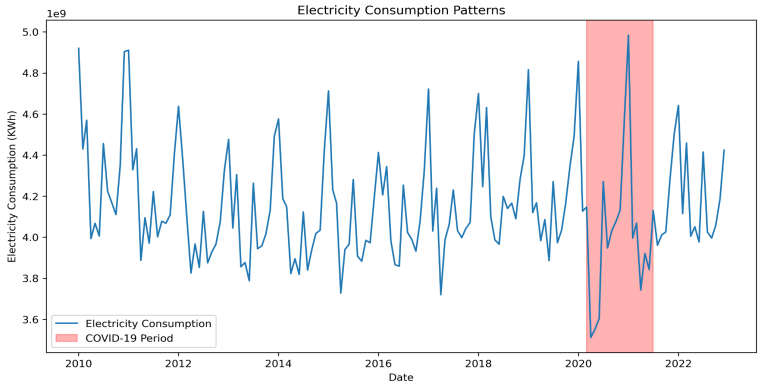

Figure 3. 3 - Monthly view of electricity consumption and the Covid-19 period

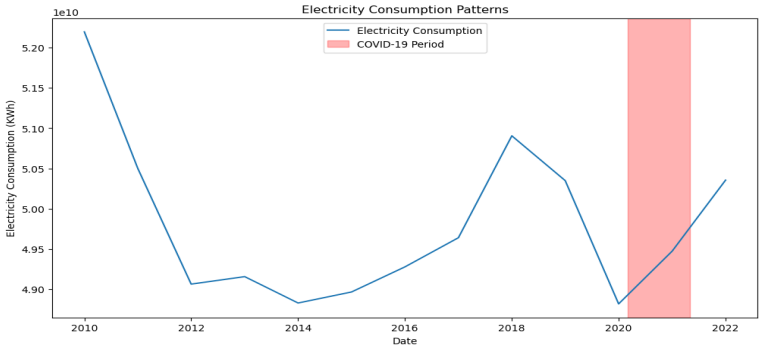

Figure 3. 4 - Annual view of electricity consumption and the Covid-19 period

Figure 3.3 illustrates a monthly view of electricity consumption in Portugal, from which it is possible to understand better the trend between hot and cold months and how much COVID-19 affected consumption at the start of 2020.

Figure 3.4 provides a broader perspective on annual consumption between 2010 and 2022, suggesting that while consumption values experienced a significant decline at the beginning of 2020, they quickly recovered in the following months, appearing in 2021 with a higher average annual value than in 2020.

The average electricity consumption before, during, and after COVID-19 was calculated to provide a better understanding of the impact of COVID-19 in this situation, and the results are shown in Table 3.2.

Table 3. 2 - monthly and yearly average electricity consumption

	Before COVID-10	During COVID-19	After COVID-19
Average monthly electricity consumption (KWh)	4 162 729 508.2	4 044 500 000.0	4 151 650 000.0
Average yearly electricity consumption (KWh)	49 789 454 545.5	49 471 000 000	50 354 000 000

It is possible to look into the annual electricity consumption by type with a yearly perspective of the average amount of electricity used. Figure 3.5 and Table 3.3 show that during COVID-19, government buildings' electricity use dropped significantly (1 776 750 787 KWh), and street light consumption dropped by 159 840 520 KWh. All other categories of energy usage increased.

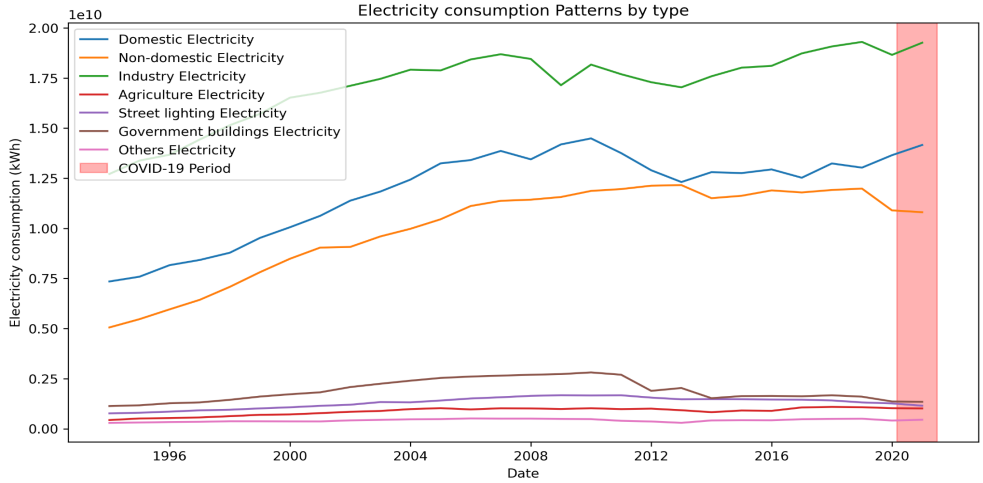

Figure 3. 5 - Annual electricity consumption by type

Table 3. 3 - Annual average electricity consumption by type

	Before COVID-10	During COVID-19	Difference	
Domestic Electricity	11 803 818 218	14 160 391 660	+	2 356 573 442
Non-domestic Electricity	9 986 767 297	10 805 171 602	+	818 404 305
Industry Electricity	17 077 219 095	19 262 138 645	+	2 184 919 550
Agriculture Electricity	867 845 754	1 015 287 423	+	147 441 669
Street lighting Electricity	1 310 989 040	1 151 148 520	-	159 840 520
Government buildings Electricity	1 922 971 561	146 220 774	-	1 776 750 787
Others Electricity	418 814 468	449 768 659	+	30 954 191

It is crucial to comprehend how weather variables behave and their patterns because, as was seen in the previous chapter, doing so can improve the performance of models.

Figures 3.6 and 3.7, which represent Daylight and Temperature, show somewhat similar trends, with low values in the winter and high values in the summer. There hasn't been a significant shift in this pattern since 2010. It is clear from the relationship between these factors and electricity usage that consumption is higher when temperature and solar radiation are at their lowest levels and *vice versa*.

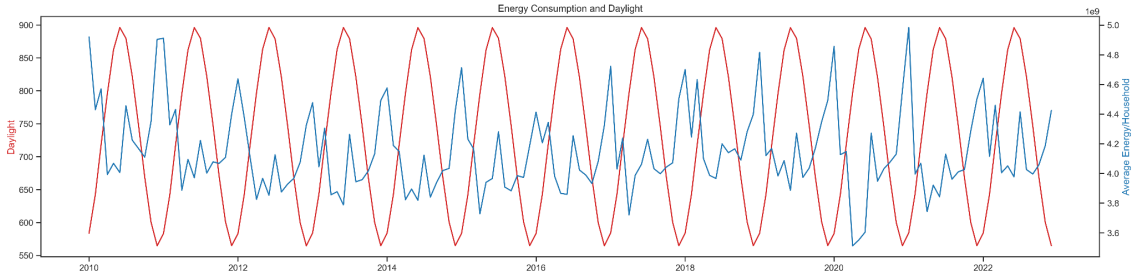

Figure 3. 6 - Monthly electricity consumption and Daylight

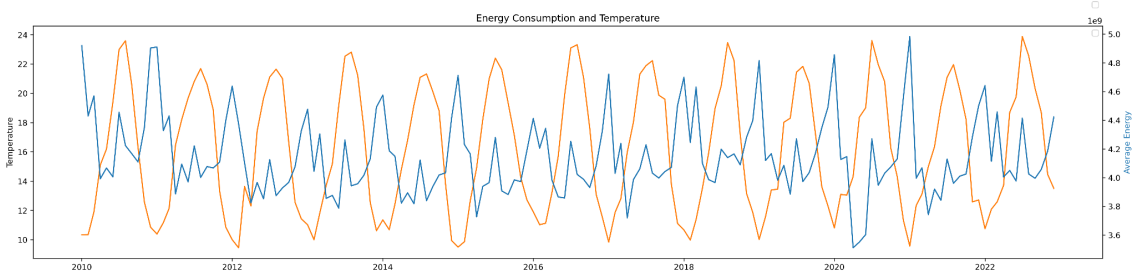

Figure 3. 7 - Monthly Electricity Consumption and Temperature

Figures 3.8, 3.9, 3.10 and 3.11 show the opposite behavior of the previous figures for Humidity, Precipitation, Clouds, and Wind, with the highest values occurring in the coldest and the lowest in the hottest months. Since 2010, the behavior of these variables has been relatively consistent. Consumption of electricity is positively correlated with the factors, as expected.

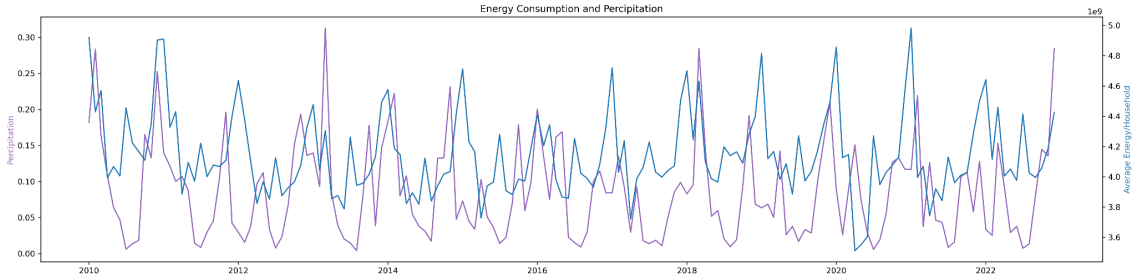

Figure 3. 8 - Monthly electricity consumption and Precipitation

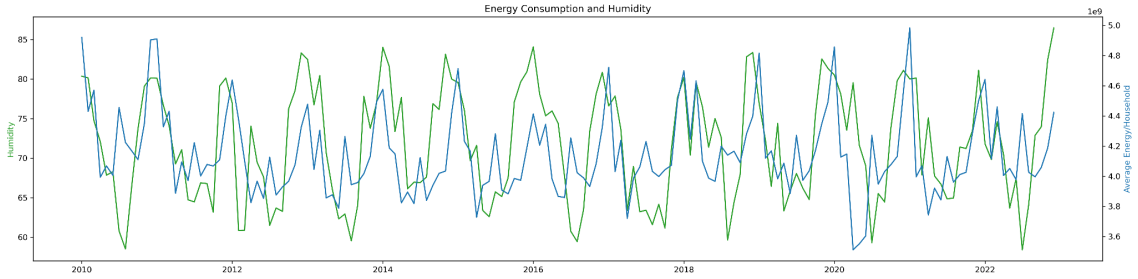

Figure 3. 9 - Monthly electricity consumption and Humidex

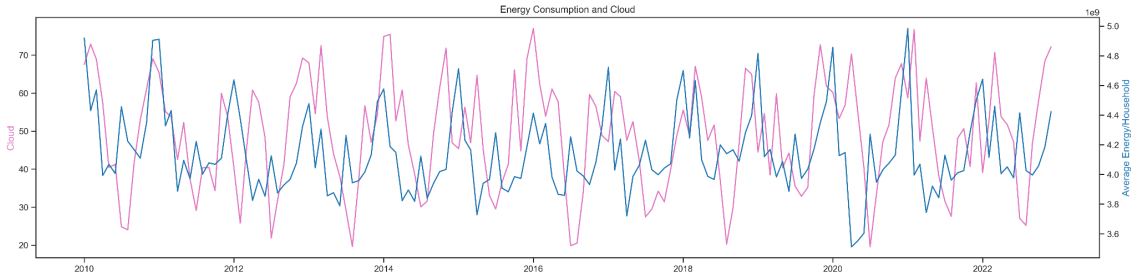

Figure 3. 10 - Monthly electricity consumption and Cloud

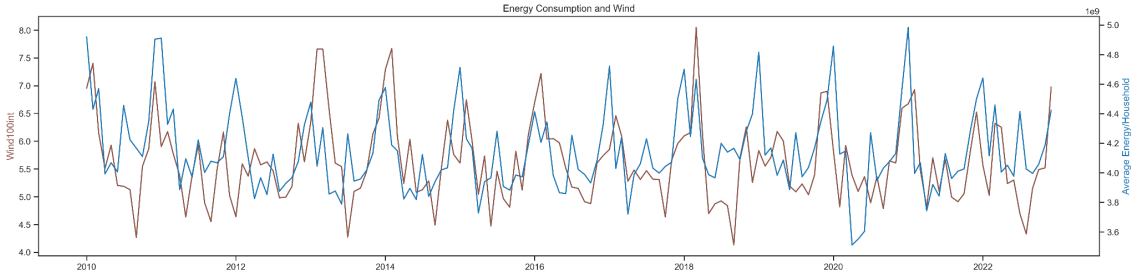

Figure 3. 11 - Monthly electricity consumption and Wind

The literature review said that energy consumption is closely related to economic growth since the steady growth of the global economy will require more energy consumption, which can be observed in Figure 3.12.

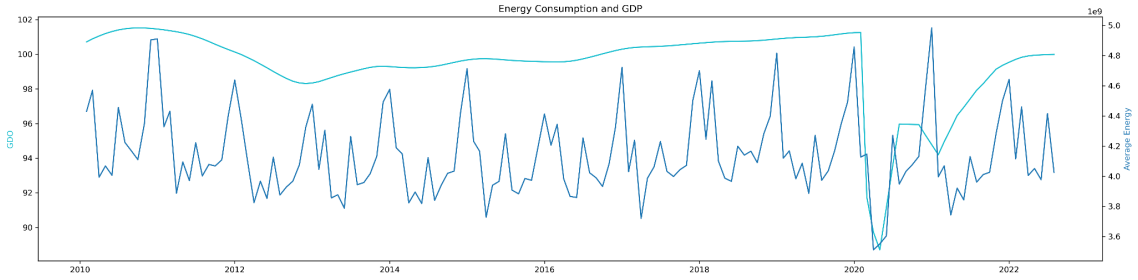

Figure 3. 12 - Monthly electricity consumption and GDP

Even though GDP fluctuates less than power consumption, it is still easy to see that starting in 2011, GDP decreased slightly, just like electricity consumption. Beginning in 2013, the GDP grew gradually along with the progressive increase in power usage. Both variables see a sharp decline in value in 2020, but by 2022, they have already returned to "normal" levels.

3.4 Feature engineering

As seen, energy consumption patterns experienced some disruptions due to the pandemic, and understanding these changes is crucial for accurate forecasting, so to analyze the impact of the COVID-19 pandemic on the performance of the models, the work was divided into two parts:

In the first part of the analysis (total period), it was considered the entire dataset, including data from before and after the COVID-19 pandemic. This provides an overview of how the variables and predictive models performed when trained on a dataset spanning the pre-pandemic and pandemic periods.

In the second part (pre-COVID period), the focus was exclusively on the dataset containing values from before the start of the COVID-19 pandemic. So, the analysis aims to evaluate the effectiveness of predictive models in capturing trends and prevalent seasonality patterns before the pandemic.

Robust predictive models are crucial for predicting electricity consumption in both stable (before COVID) and non-stable (period where COVID existed) contexts. This section describes feature engineering, an essential stage in developing predictive models that bridge the gap between raw data and model-driven insights. It includes creating, modifying, and choosing features that impact the performance of a predictive model.

The availability of date-time information allows for creating additional columns for data visualization and as features for predictive models. The new columns correspond to the year, month, day of the year, week of the year, and quarter, and the pair-plot (a grid of scatterplots that allows one to visualize the relationships between multiple pairs of variables) of KWh and the new date variables are shown in Figure 3.13.

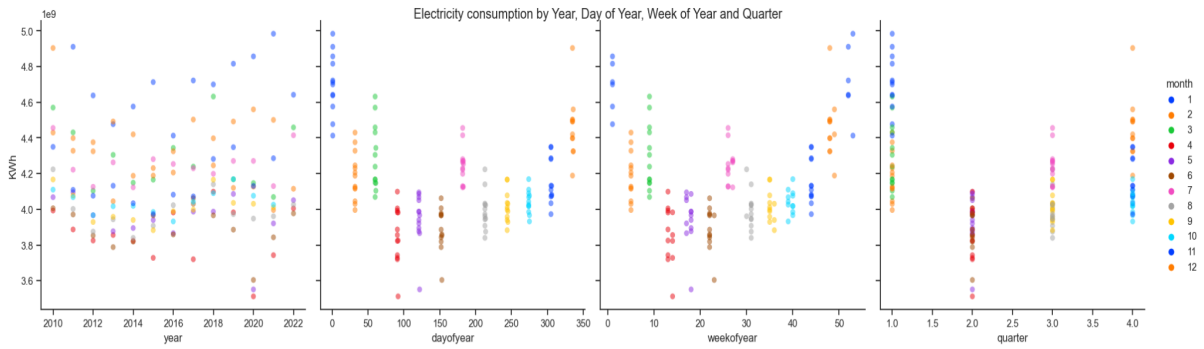

Figure 3. 13 - time series features based on time series index

This aggregation of data on energy consumption by time period and the correlation between variables is essential for identifying patterns and behaviors in energy consumption that are important for the modeling phase.

The correlation between all variables was studied to help understand how different factors affect the predictive results. Because it is critical to analyze the relationship between variables to create a robust predictive model, the method for identifying those relationships must also be applied correctly. The Pearson Correlation Coefficient was calculated for each period in order to analyze the linear relation between numerical time series.

Eta - correlation ratio was also used to examine the correlation between the categorical variables "covid", “year”, “month”, “day of the year”, “week of the year”, and “quarter” and the non-categorical variable "KWh" over the entire period. This statistical measure was chosen because, when one variable is categorical and the other is not, it allows quantifying the degree to which the dispersion in the variable of interest, in this case, "KWh," can be attributed to changes within various categories of the categorical variables. Eta can have values between 0 and 1, with values closer to 0 indicating that no single category significantly influences the variable "KWh" more than any other. On the other hand, values nearer to 1 imply that one or more categories are substantially different from the others and have a more significant impact on the variable "KWh". Table 3.4 shows the results of this test for the entire period.

Table 3. 4 - Correlation Matrix for categorical variables for the entire period

	KWh	**covid**	**quarter**	**month**	**year**	**dayofyear**	**weekofyear**
KWh	1.000000	0.962279	0.973001	0.974022	0.959979	0.965818	0.968596
covid	0.962279	0.962990	0.066003	0.107928	0.812861	0.412924	0.357672
quarter	0.973001	0.066003	1.000000	1.000000	0.084324	1.000000	1.000000
month	0.974022	0.107928	1.000000	1.000000	0.054366	1.000000	1.000000
year	0.959979	0.812861	0.084324	0.054366	1.000000	0.269800	0.272239
dayofyear	0.965818	0.412924	1.000000	1.000000	0.269800	1.000000	0.729416
weekofyear	0.968596	0.357672	1.000000	1.000000	0.272239	0.729416	1.000000

The correlation matrix for the variables, including energy consumption (KWh) and various temporal factors such as quarter, month, year, day of the year, and week of the year, reveals a strong positive correlations with quarter, month, year, day of the year, and week of the year, emphasizing the profound influence of time-related variables on energy consumption. Furthermore, the very high correlation between KWh and Covid suggests a significant impact of pandemic-related changes on energy demand.

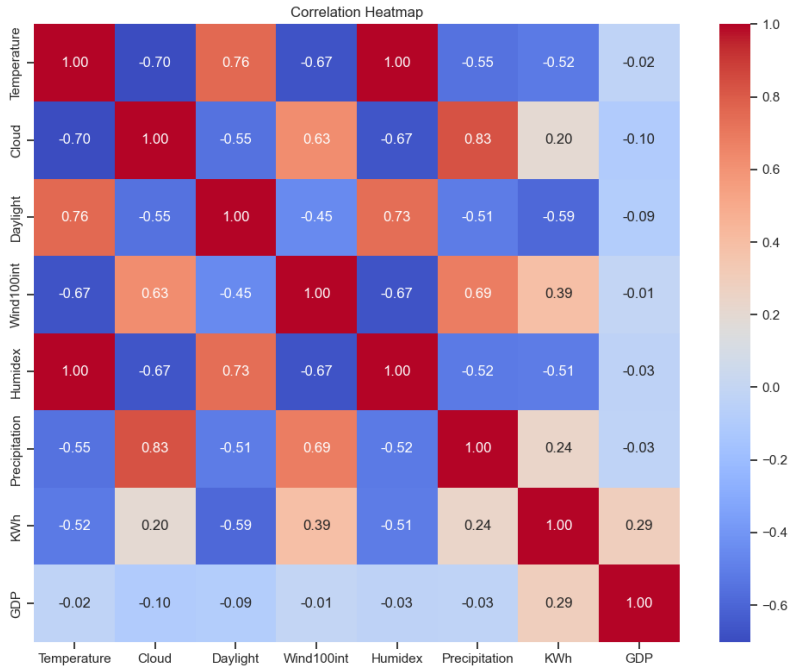

Figure 3. 14 - Variables correlation heatmap of the entire period

Throughout the entire data period, it was observed intriguing correlations among electricity consumption and various meteorological and economic variables (Figure 3.14). Firstly, temperature exhibits a negative correlation with energy consumption (KWh), implying that as temperatures rise, energy usage tends to decrease. This finding may be attributed to energy-efficient cooling systems in warmer weather or other unmeasured factors influencing the relationship. Conversely, wind intensity shows a positive correlation with energy consumption, indicating that windier days are associated with slightly higher energy usage, possibly due to increased heating or cooling demands. Additionally, GDP, a measure of economic activity, exhibits a weak positive correlation with energy consumption, suggesting that as the economy grows, energy consumption tends to increase, reflecting the energy demands of expanding industrial and commercial sectors. Daylight has a negative correlation with energy consumption, signifying that longer daylight hours are linked to reduced energy usage. This relationship aligns with the idea that ample natural daylight can reduce the need for artificial lighting and heating. Cloud cover and precipitation display weak positive correlations with energy consumption, suggesting that cloudier or rainy days might lead to slightly higher energy usage, possibly due to increased indoor activities. The humidex, a combined measure of temperature and humidity, also exhibits a moderate negative correlation with energy consumption, indicating that as humidex levels rise, energy consumption tends to decrease.

These findings collectively illustrate the intricate interplay of meteorological and economic factors on energy consumption, emphasizing the need for comprehensive modeling and analysis to fully understand these relationships and inform energy management strategies.

For the categorical variables before Covid, the correlation matrix is shown in table 3.5

Table 3. 5 - Correlation Matrix for categorical variables for the pré-covid period

	KWh	quarter	month	year	dayofyear	weekofyear
KWh	1.000000	0.971918	0.977008	0.965896	0.975900	0.962599
quarter	0.971918	1.000000	1.000000	0.130136	1.000000	1.000000
month	0.977008	1.000000	1.000000	0.093707	1.000000	1.000000
year	0.965896	0.130136	0.093707	1.000000	0.303815	0.309857
dayofyear	0.975900	1.000000	1.000000	0.303815	1.000000	0.742323
weekofyear	0.962599	1.000000	1.000000	0.309857	0.742323	1.000000

The correlation matrix among energy consumption (KWh) and time-related variables—quarter, month, year, day of the year, and week of the year—reveals remarkable associations. Energy consumption (KWh) displays strong positive correlations with all temporal factors, indicating pronounced seasonality and consistent patterns in energy usage across different time intervals. Quarter and month exhibit perfect positive correlations, implying mutual redundancy. At the same time, the year variable moderately correlates with the day and week of the year, illustrating their roles in marking the annual cycle. The robust relationships in this matrix underscore the significant influence of time-related factors on energy consumption trends, emphasizing the importance of seasonality in energy management and forecasting.

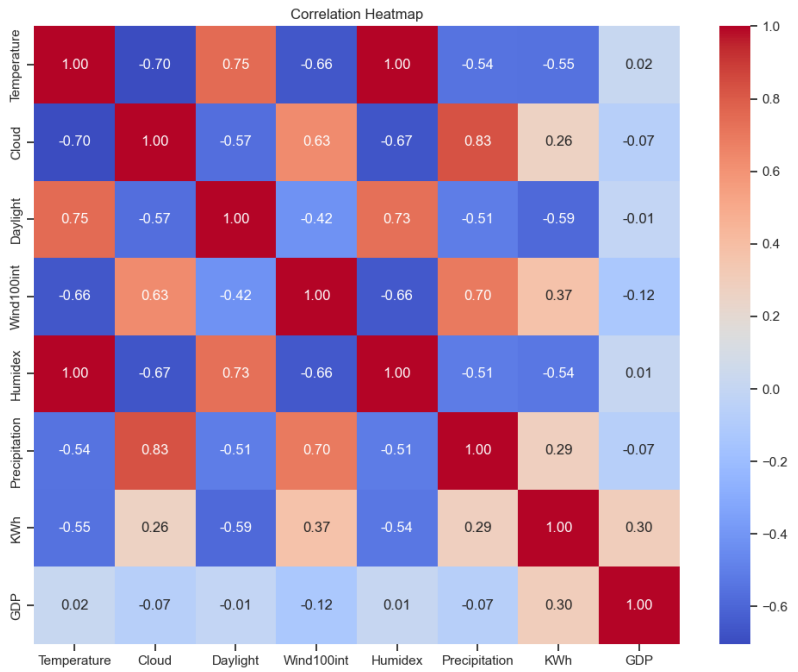

Figure 3. 15 - Variables correlation heatmap of the pre-covid period

Before the COVID-19 pandemic, a deeper analysis of the correlations matrix revealed interesting insights, as can be observed in Figure 3.15. Temperature displays a moderate negative correlation with energy consumption, suggesting lower energy use during warmer periods. Cloud cover has a weak positive correlation with energy consumption, while longer daylight hours negatively correlate with energy use. Wind intensity shows a moderate positive correlation, likely linked to increased heating or cooling needs during windy conditions. The humidex exhibits a moderate negative correlation, indicating reduced energy consumption during high humidity. Precipitation has a weak positive correlation with energy consumption, possibly due to increased indoor activities on rainy days. Lastly, the correlation between Gross Domestic Product (GDP) and energy consumption is weakly positive, reflecting the energy demands of economic growth. These insights underscore the intricate interplay of meteorological and economic factors influencing energy usage patterns.

After examining correlations, it is crucial to conduct VIF (Variance Inflation Factor) tests for multicollinearity. When two or more independent variables in a regression analysis have a high degree of correlation, multicollinearity occurs, resulting in unstable and unreliable regression coefficients. Consequently, it becomes challenging to accurately determine the specific impact of each independent variable on the dependent variable.

The Variance Inflation Factor (VIF) tests were carried out to check the multicollinearity among particular variables in each studied time period. The results can be found in the appendix (Table 1). Following the thorough analysis that included correlation and VIF evaluations, the following independent variables were chosen as the final set for each period:

- Entire period: KWh, Wind100int, Precipitation, GDP, Covid, Month, Cloud and Daylight;
- Before covid period: KWh, Wind100int, GDP, Month, Cloud and Daylighg

CHAPTER 4

Modeling

Time series forecasting methods mainly focus on past patterns to predict the future. The presence of multidirectional trends, seasonal and cyclical changes, and structural gaps in energy consumption indicators requires selecting appropriate models (Kalimoldayev et al., 2020). This chapter builds on the previous one using predictive models for the analyzed data. Each predictive model is briefly described, and the chapter concludes by comparing and recommending the best model for implementing accurate energy consumption forecasts using public data for the Portugal case.

Based on the literature review (Chapter 2), it was considered the set of the most successful models. The three categories below can be used to group the five selected models:

- Classical methods: SARIMA, SARIMAX, and VAR
- Machine learning method: Support Vector Regression (SVR)
- Deep learning algorithms: Long Short-Term Memory (LSTM) neural network

For each of the periods explained in the previous chapter, the following was done:

- Evaluate the stationarity of each variable in the dataset and transform the data accordingly.
- Train and test each of the five selected predictive models on the respective datasets.
- Evaluate and compare the model's performance using appropriate metrics such as Mean Absolute Percentage Error (MAPE), Coefficient of Determination (R^2), and Root Mean Square Error (RMSE).

The predictive models applied to the dataset were developed with the open-source Python programming language (version 3.9.13) with the usage of the following libraries: Pandas; Numpy; Matplotlib; Seaborn; Statsmodels; Keras; Tensorflow, and Scikit-Learn.

The data was sequentially split into training and test sets for each period. Table 4.1 outlines how the data was divided.

Table 4. 1 - Train and test split for the 2 periods

	Entire period	**Pre-covid**
Train	127	105
Test	24	16
Nr of variables	8	6

The key evaluation metric used to determine the effectiveness of the different predictive models was the Mean Absolute Percentage Error (MAPE). MAPE measures the average percentage difference between predicted and actual values. Lower MAPE values indicate better performance, providing a straightforward assessment of predictive accuracy. The formula for MAPE is as follows:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Actual_t - Predicted_t|}{|Actual_t|}$$

(1)

Where:

- *n* is the total number of data points.
- $Actual_t$ represents the actual value at time *t*
- $Predicted_t$ represents the predicted value at time *t*

For the model comparison, in addition to MAPE it was used the Coefficient of Determination (R^2), and the Root Mean Square Error (RMSE).

The "R-squared" value, also referred to as R^2, measures the goodness-of-fit of regression models. It ranges from 0 to 1, with higher values indicating better fit, and it denotes the percentage of variance in the dependent variable that can be accounted for by independent variables. The formula for R^2 is as follows:

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(Actual_t - Predicted_t)^2}{\sum_{t=1}^{n}(Actual_t - \underline{Actual})^2}$$

(2)

Where:

- *n* is the total number of data points.
- $Actual_t$ represents the actual value at time *t*
- $Predicted_t$ represents the predicted value at time *t*
- $\underline{Actual}$ is the mean of actual values.

The model's accuracy in predicting absolute values can be determined by the RMSE, which computes the average magnitude of errors between predicted and actual values. Higher predictive accuracy is reflected by lower RMSE values. The formula for RMSE is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Actual_t - Predicted_t)^2}$$

(3)

Where:

- n is the total number of data points.
- $Actual_t$ represents the actual value at time t
- $Predicted_t$ represents the predicted value at time t

4.1 Stationarity unit root and stationarity tests

An iterative modeling methodology includes looking at trend and seasonality patterns, identifying and estimating model parameters using the maximum likelihood or non-linear least squares approach, and testing the model's adequacy and forecast accuracy (Hyndman and Athanasopoulos, 2018).

A time series is stationary if it has constant mean, variance, and covariance over time, i.e., the mean of a stationary time series does not change, causing cycles to eventually "die out", and if the series fluctuates around a constant long-term mean, it will have a time-independent, finite variance.

A series can be non-stationary for various reasons, including a linear or non-linear trend, seasonality, and a non-constant variance. However, there are ways to make them stationary; for example, if the series has a linear trend, the first or second difference operator (non-seasonal difference) is applied.

The data need to be stationary for applying the VAR model, so the data was examined in this direction. The stationarity concept, primarily concerned with the statistical characteristics of numerical data over time, does not apply to Month and Covid in the same way because they are qualitative variables. Two key features of the time series data can be seen through the exploratory analysis and reviewed literature. Therefore, the GDP trend and the seasonality of Daylight variables were handled before analyzing the remaining time series. This was done for both considered periods. Daylight variables (in both datasets) displayed distinct seasonal patterns that showed consistent fluctuations. The seasonality was eliminated to enable a focus on underlying patterns and connections in the data. Also, a clear linear trend was visible for the GDP time series, which was removed accordingly.

For time series stationarity, there are some tests (called unit roots tests) - such as Dickey-Fuller (DF) - which only contains a lag, Augmented Dickey-Fuller (ADF) - is an improvement of DF since it allows the addition of more lags to guarantee the independence of the variable -, and Philips Perron (PP) - allows to apply in series where the variance is not constant.

The unit root tests ADF and PP were used and present the following test hypotheses:

▪ H0: Existence of unit root (non-stationary time series)

▪ H1: No unit root (stationary time series).

To decide whether to reject or not the null hypothesis, the p-value indicator is used, which will be compared with the significance levels - α (1%, 5% and 10%). A significance level of 5% was considered for all tests. If the p-value is less than α, H0 is rejected, and the time series is stationary. If the p-value exceeds α, H0 is not rejected, and we conclude that the series has a unit root(is non-stationary).

If the ADF and PP tests do not agree, i.e., one rejects the null, and the other does not reject H0, the KPSS test is used, which is a stationarity test, where the hypotheses are:

- H0: Stationary time series
- H1: Non-stationary time series.

Since the mean and the variance of some of the series are not constant, the ADF, the PP, and the KPSS tests were performed for both datasets. The month and COVID variables were not tested because they are qualitative exogenous variables by default. Therefore, they are used in the models as they are. The stationarity results of the remaining variables can be observed in the following tables (Table 4.2).

Table 4. 2 - Unit root tests (ADF and PP) and stationarity test (KPSS) of the entire period

	ADF	PP	Conclusion
KWh	0.0479	0.000	stationary
Wind100int	0.0	0.000	stationary
Precipitation	0.0	0.000	stationary
GDP_detrended	0.0	0.006	stationary
Daylight_deseasonal	0.0	0.029	stationary
Cloud	0.0429	0.000	stationary

For the dataset of the entire period, after removing the trend for GDP and removing the seasonality for Daylight, all variables were stationary.

Now, for the dataset containing values before COVID, the results of the tests ADF, PP, and KPSS can be observed in Table 4.3.

Table 4. 3 - Unit root tests (ADF and PP) and stationarity test (KPSS) of the pre-covid period

	ADF	PP	KPSS	Conclusion
KWh	0.0463	0.000	-	stationary
cloud	0.0	0.000	-	stationary
Daylight_deseasonal	0.0	0.072	0.0228	stationary
Wind100int	0.0	0.000	-	stationary
GDP_detrended	0.0114	0.001	-	stationary

For the dataset of the pre-COVID data, the tests concluded that all series were stationary; however, the KPSS test was performed on the variable Daylight_deseasonal since the other unit root tests do not agree.

4.2 SARIMA and SARIMAX

Autoregressive Moving Average (ARMA) or Autoregressive Integrated Moving Average (ARIMA) models, which employ the Box-Jenkins approach, are among the most frequently applied traditional "ad-hoc" time series techniques (Box et al., 2016). The ARIMA model has parameters p (the autoregressive order or lag of the model), d (the integration or differencing order), and q (the moving average order) that forecast the future values of a time series based on a linear combination of its previous values and disturbances. The model is explained in Equation (4).

$$Y_t = \mu + \sum_{j=1}^{p} \phi_j \left(Y_{t-j-\mu}\right) + \sum_{j=1}^{p} \psi_j \varepsilon_{t-j} + \varepsilon_t \qquad (4)$$

Where:

- Y_t represents the value of the time series at time t
- μ is the constant term or mean of the time series
- $\sum_{j=1}^{p} \phi_j \left(Y_{t-j-\mu}\right)$ is the autoregressive (AR) component of the ARIMA model. The AR component involves lagged values of the time series (Y) weighted by coefficients ϕ_j.
- p represents the order of the autoregressive component, which indicates how many lagged values are included in the model.
- $\sum_{j=1}^{p} \psi_j \varepsilon_{t-j}$ is the moving average (MA) component of the ARIMA model. The MA component involves lagged values of the error term ε (the difference between the observed value and the predicted value) weighted by coefficients ψ_j.
- ε_t represents the error term at time t

To handle the trend in the data, the integrated term "i" is used, turning the non-stationary data into stationary and naming the model ARIMA.

It is clear from the previous sections that seasonal patterns can be found in the data. The ARIMA model is extended to a Seasonal ARIMA (SARIMA) model to account for seasonality, denoted as (p, d, q)(P, D, Q)s, where *s* is the seasonal duration, which is commonly set to 12 for monthly series. SARIMA (p, d, q)(P, D, Q)s capture both non-seasonal and seasonal components of time series data, where p, d, q, P, D, and Q are non-negative integers that reflect the polynomial order of the model's autoregressive (AR), integrated (I), and moving average (MA) components (Vagropoulos et al., 2016). SARIMA models outperformed typical ARIMA models regarding forecast accuracy (Verdejo et al., 2017).

SARIMAX models, an extension of SARIMA, go a step further by including exogenous variables in the model, joining the historical energy consumption data. The additional information has the potential to enhance forecasting performance significantly.

For both periods, it was analyzed the Autocorrelation function (ACF) and Partial Autocorrelation Function (PACF) to visually determine the values for the Autoregression component (AR) and the Moving average component (MA). The ACF can be described as the linear relationship between two data points in time as a function of their time difference; the PACF of a specific lag is the partial correlation between the time series and itself at that lag when all the in-between information is considered constant (Su, 2018). The results can be seen in the appendix (in Figure 1 and 2)

The model selection process for both periods also involved utilizing automated methods for parameter tuning, resulting in the identification of the best SARIMA model for each case. To assess the model's predictive performance, the best model was applied to the test dataset, and the forecasts were generated with the "dynamic=False" setting, allowing each prediction to be based on the entire historical data up to that point.

Table 4.4 provides the best-tested models for the two periods, the hyperparameters, the exogenous variables, and the Mean Absolute Percentage Error.

Table 4. 4 - Parameters of SARIMA and SARIMAX models

Model	Variables	MAPE
	Entire period	
SARIMA(1,0,0)(0,1,1,12)	-	3.995
SARIMAX(1,0,1)(1,0,1,12)	'Daylight' , 'Wind100int', 'Precipitation', 'GDP', 'covid'	2.254
SARIMAX(1,1,1)(0,1,1,12)	GDP	2.256
SARIMAX(0,1,0)(1,0,1,12)	**GDP, 'Wind100int'**	**2.139**
	Pre-covid period	
SARIMA(1,0,1)(1,0,0,12)	-	3.596
SARIMAX(0,0,1)(1,0,1,12)	'Daylight' , 'Wind100int', 'GDP', 'month', 'Cloud'	2.778
SARIMAX(1,0,0)(1,0,1,12)	GDP, Cloud	2.617
SARIMAX(1,0,0)(1,0,1,12)	**GDP, Wind100int**	**2.409**

Among the various predictive models evaluated for forecasting Portugal's electricity consumption in the non-stable market, the SARIMA(1,0,0)(0,1,1,12) model, which does not include any additional predictor variables, exhibited a Mean Absolute Percentage Error (MAPE) of 3.995%. While this model provides a baseline for forecasting, it is clear that incorporating additional variables can significantly improve predictive accuracy. In this period analysis, the SARIMAX(1,0,1)(1,0,1,12) model, which includes the predictor variables 'Daylight,' 'Wind100int,' 'Precipitation,' 'GDP,' and 'covid,' achieved a MAPE of 2.254%. This result highlights the importance of considering multiple factors, including weather-related variables, economic indicators (GDP), and external factors like the COVID-19 pandemic, to enhance the accuracy of electricity consumption forecasts. Additionally, the SARIMAX(1,1,1)(0,1,1,12) model, which solely relies on GDP as the predictor variable, performed similarly well with a MAPE of 2.256%, underlining the significant role of economic conditions in shaping energy demand during periods of instability. Moreover, the SARIMAX(0,1,0)(1,0,1,12) model, incorporating GDP and 'Wind100int' as predictor variables, achieved the lowest MAPE of 2.139%. This indicates that a combination of economic indicators and wind conditions can also effectively predict electricity consumption patterns during non-stable market conditions.

During the pre-COVID era, a relatively stable market, the SARIMA(1,0,1)(1,0,0,12) model exhibited a MAPE of 3.596% without any additional predictor variables. However, it's worth noting that this model's performance is lower than those in the non-stable market analysis, emphasizing the different dynamics at play during stable and unstable market conditions. Among the pre-COVID models that included additional predictor variables, the SARIMAX(0,0,1)(1,0,1,12) model, featuring 'Daylight,' 'Wind100int,' 'GDP,' 'month,' and 'Cloud' as predictors, achieved a MAPE of 2.778%. This demonstrates that a combination of weather-related variables, economic indicators, and temporal factors can be valuable for forecasting electricity consumption during stable periods. Additionally, the SARIMAX(1,0,0)(1,0,1,12) model, considering GDP and 'Cloud' as predictor variables, achieved a MAPE of 2.617%, while the SARIMAX(1,0,0)(1,0,1,12) model with GDP and 'Wind100int' as predictors achieved the lowest MAPE of 2.409%. Again, these results highlight the importance of considering economic indicators and specific weather-related variables when forecasting electricity consumption during stable market conditions.

These findings emphasize the need for a multidimensional approach when forecasting electricity consumption, as the relative importance of different factors may vary depending on the market's stability. Incorporating a combination of economic indicators, weather-related variables, and external factors can lead to more accurate predictions in both stable and non-stable market conditions.

4.3 VAR

Christopher Sims pioneered the multilinear Vector Autoregressive (VAR(p)) model in 1980, which is a generalization of the ARMA models introduced by Box and Jenkins in 1978 (Fotsing et al., 2014). The introduction of VAR models constituted a significant advance in investigating dynamic interactions within connected time series data. These models are made up of equations for each variable that explain its evolution based on its lag values as well as the lag values of other variables.

VAR models are distinguished from standard dynamic simultaneous equation models by their symmetrical treatment of all variables as endogenous. In contrast to these models, VAR models do not impose "incredible identification restrictions" and provide a theory-free technique for finding economic linkages.

VAR models have been used successfully to investigate the complex relationships between energy use and economic growth (Alsaedi & Tularam, 2019). In its basic form, a VAR consists of a set of K endogenous variables $yt = (y1t, \dots, ykt, \dots, yKt)$ for $k = 1, \dots, K$. After including ρ lags of the endogenous variables, the VAR(ρ) model may be defined as:

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + D_t + u_t$$

(7)

Where:

- $A_{1, \dots, } A_p$ are coefficients associated with the lagged values of the dependent variable y.
- y_t is the current value of the dependent variable at time t.
- y_{t-p} are lagged values of the dependent variable y at earlier time points.
- D_t represents any deterministic components or factors that affect the dependent variable at time t but are not related to the past values of y. These components can include things like trends, seasonality, and exogenous variables.
- u_t is the error term or residual at time t.

Another benefit of multivariate data is the possibility of analyzing Granger causality, which aids in determining the direction of causality among variables. Granger causality tests, if a variable, may be used to predict another variable. This is significant because it allows the variables to be eliminated and individually modeled if there is no relationship between them. On the other hand, if a causal relationship exists, the variables must be considered at the modeling stage. In doing so, the regression equation's zero-coefficient of past values null hypothesis is tested (variable X is not Granger causing variable Y, if all X coefficients in Y equation are null). The null hypothesis is rejected if the p-value from the test is less than the significance level of 0.05.

Granger causality tests were employed to investigate causal relationships between time series. The VAR model was applied to the data, enabling an in-depth exploration of the temporal interactions between variables. Various lag orders were considered, and model selection criteria, including AIC, BIC, FPE, and HQIC, were harnessed to pinpoint the most suitable model. Table 4.5 shows the order of the tested VAR models, the set of variables, and the MAPE metrics.

Table 4. 5 – Variables and parameters of VAR models

Model	Variables	MAPE
	Entire period	
Var(1)	KWh, covid, month	6.740
VAR(1)	KWh, month, GDP_detrended	5.346
VAR(1)	**KWh, GDP_detrended**	**5.250**
	Pre-covid period	
Var(1)	**KWh, month**	**4.763**
VAR(1)	KWh, Wind100int, month	4.747
VAR(1)	KWh, wind100int, Month, Daylight_deseasonal	4.784

When evaluating the performance of the VAR model, the choice of variables had noticeable effects on predictive accuracy.

Among the models considered for the entire period, three VAR(1) models stand out. The first, incorporating electricity consumption (KWh), the COVID-19 variable, and the month, resulted in a MAPE of 6.740%. While it provides reasonable predictions, there is room for improvement. The second model, which included electricity consumption, month, and GDP_detrended, achieved a lower MAPE of 5.346%. The inclusion of GDP_detrended as an economic indicator appears to enhance predictive accuracy. The third model, considering only electricity consumption and GDP_detrended, resulted in a MAPE of 5.250%. This underscores the importance of economic factors in electricity consumption prediction during the entire period.

In the pre-COVID era, the three VAR(1) models evaluated that performed the best were the model featuring electricity consumption and month achieved a MAPE of 4.763%, indicating that temporal factors play a significant role in electricity consumption forecasting during stable periods. The second model, including electricity consumption, Wind100int (wind conditions), and month, achieved a slightly lower MAPE of 4.747%. Wind conditions appear to have a modest impact on improving forecasting accuracy during stable periods. The third model, which included electricity consumption, Wind100int, month, and Daylight, resulted in a MAPE of 4.784%, suggesting that incorporating Daylight as an additional variable did not substantially improve forecasting accuracy during stable periods.

Again, the analysis demonstrates that the choice of variables and the market conditions significantly impact the performance of VAR models in predicting electricity consumption in Portugal. Economic indicators, such as GDP_detrended, play a crucial role during the entire period, while temporal factors and specific weather-related variables, like Wind100int, have more influence during stable market conditions.

4.4 SVR

Vapnik's (1995) Support Vector Machine (SVM) algorithm is based on statistical learning theory and structural risk minimization. Support Vector Machines (SVM) is a widely used technique that has shown effectiveness in various areas, including pattern recognition, classification, and non-linear regression tasks like electricity price predictions (Weron, 2014). SVM's central principle is to convert the original variables into a high-dimensional feature space and incorporate a penalty component in the error function to control complexity. The fundamental goal is to minimize forecast error on a training dataset while maintaining as flat a functional shape as possible. SVM leverages Kernel functions to address non-linear relationships in data (Deb et al., 2017). Kernel functions are critical in SVM because they allow the model to capture complicated patterns and non-linearities in the data. Linear, Polynomial, Gaussian Radial Basis, and Multilayer Perceptron are some of the most frequent Kernel functions used in SVM (Chou & Tran, 2018).

The Support Vector for Regression (SVR) is a variant of SVM specifically tailored for addressing non-linear regression problems. In cases of non-linear regression, the data suffer a transformation via a nonlinear kernel function, which maps the inputs to a higher-dimensional feature space. The overall performance of an SVM regression model hinges on the appropriate selection of kernel parameters (Shi et al., 2018). The equation for the SVR model is shown in equation (8).

$$f(x) = \sum_{i=1}^{N} (\beta_1^* - \beta_i) K(\chi_i - \chi) + b \qquad (8)$$

Where:

- $f(x)$ represents the predicted output or response variable for the input feature vector x.
- N is the number of support vectors in the SVR model.
- β_1^* is a coefficient associated with the support vector that has the most influence on the model's predictions. It's the weight given to the first support vector.
- β_i represents the coefficients associated with the other support vectors (i ≠ 1).
- $K(\chi_i - \chi)$ is the kernel function. The kernel function computes the similarity or distance between the input feature vector x and the support vectors χ_i.

- b is the bias term or intercept. It's an offset added to the sum of the weighted kernel function evaluations to make the prediction.

As before, a collection of variables was explored and tested in the context of energy consumption. The SVR model was trained with different hyperparameters for each case, since a grid search was applied to find the best ones. Table 4.6 shows the various models tested.

Table 4. 6 - Parameters of SVR models

Model	Variables	MAPE
	Entire period	
C:0.1, epsilon:0.01, gamma:0.01 , kernel: poly	'Daylight' , 'Wind100int', 'Precipitation', 'GDP', 'covid', 'month'	5.399
C:0.1, epsilon:0.1, gamma: 0.01, kernel: poly	'Daylight' , 'Wind100int', 'Precipitation',	4.205
C:1, epsilon:0.01, gamma: 1, kernel: poly	**'GDP', 'covid', 'month'**	**3.998**
	Pre-covid period	
kernel': 'poly', 'gamma': 0.01, 'epsilon': 0.01, 'C':0.1	'Daylight' , 'Wind100int', 'GDP', 'Cloud', 'month'	4.111
'kernel': 'poly', 'gamma': 0.01, 'epsilon': 0.1, 'C': 0.1	**'Daylight' , 'Wind100int', 'Cloud',**	**3.530**
'kernel': 'poly', 'gamma': 1, 'epsilon': 0.01, 'C': 0.1	'Wind100int', 'GDP', 'month'	3.609

Three different configurations of support vector regression (SVR) models were scrutinized in the entire period analysis. The first model included a comprehensive set of predictor variables, yielding a MAPE of 5.399%. With a simplified set of predictor variables, the second SVR model achieved a lower MAPE of 4.205%. The third SVR model, focusing solely on economic and temporal factors, performed even better with a MAPE of 3.998%.

In the pre-COVID period analysis, three SVR models were also considered, each with a polynomial kernel and different hyperparameters. The first model featured a moderate predictive accuracy with a MAPE of 4.111%. The second model, adopting a simplified approach, demonstrated improved forecasting accuracy with a MAPE of 3.530%. The third model underscored the significance of wind conditions, economic indicators, and temporal factors, achieving a MAPE of 3.609%.

4.4 LSTM

The Long Short-Term Memory (LSTM) Neural Network is one of the most advanced models used for time series forecasting, as explained by Korstanje in 2021.

The LSTM model provides an evolutionary breakthrough in neural network models, outperforming standard RNN constraints. While RNNs are frequently employed for time series prediction, they often struggle with long-term data-dependent problems (Hochreiter & Schmidhuber, 1997). Hochreiter and Schmidhuber introduced the LSTM model in 1997, fitted with memory units designed to address the inherent difficulty of recording long-term data dependencies. The self-linking memory unit that keeps the network's temporal state is the key feature of LSTM. Three gates control this memory unit: the input gate, the output gate, and the forget gate (Yu et al., 2019).

The input and output gates control the data flow into and out of the memory unit, while the forget gate allows for the selective retention of important information, allowing high-weighted output information to be sent from one neuron to the next. Neurons house memories within their pipelines in the LSTM architecture, allowing them to keep and manipulate earlier knowledge, update it, and effortlessly transfer it to succeeding layers or cells, all without losing crucial data (Feng et al., 2022) .

Two LSTM models were used and tested for time series forecasting: energy consumption depending only on its history and energy consumption depending on several variables.

In the first case, the energy consumption time series was preprocessed, including scaling using MinMaxScaler to bring it within the (0,1) interval. An LSTM-based deep learning model was then constructed using TensorFlow and Keras. The model architecture included an LSTM layer with 100 units and a rectified linear unit (ReLU) activation function, followed by a dense layer with one output unit. The model was compiled using the mean squared error (MSE) as the loss function and trained for 20 epochs.

In the second approach, we proceed similarly: the input was defined as a dataframe (variables and history of variables), and MinMaxScaler was applied to the variables. The LSTM model was trained for this multivariate approach, and its performance was assessed using similar evaluation metrics. It is important to note that these metrics, were calculated based on the normalized series.

Table 4.7 shows the characteristics of the employed models for both considered datasets.

Table 4. 7 - Parameters of LSTM models

Model	Variables	R^2
Entire period		
LSTM(100, activation = relu, input_shape = (12,1)= Sequential, Dense(1), Optimizer=adam, Loss=mse	KWh	0.534
Past_history=127, Step=1 Future_target=12, Lstm(16, activation=relu, Dense=12	**'Daylight' , 'Wind100int', 'Precipitation', 'GDP', 'covid', 'month'**	**0.883**
Pre-covid period		
LSTM(100, activation='relu', input_shape=(12, 1), Dense(1) optimizer='adam', loss='mse'	KWh	0.258
Past_history=80, Step=1 Future_target=6, Lstm(16), activation=relu, Dense=6	**'Daylight' , 'Wind100int', 'GDP', 'Cloud', 'month'**	**0.978**

For the entire period, the LSTM model, focusing solely on electricity consumption (KWh), yielded an R^2 value of 0.534 during the entire period. While this model displayed predictive capabilities, its accuracy left room for improvement. In contrast, the second LSTM model, with a more complex architecture, including parameters like past_history=127, step=1, and future_target=12, demonstrated a substantially higher R^2 value of 0.883. This model also incorporated a broader set of predictor variables, encompassing 'Daylight,' 'Wind100int,' 'Precipitation,' 'GDP,' 'covid,' and 'month.' The notable improvement in accuracy underscores the importance of including diverse variables in electricity consumption forecasting during the entire period.

During the pre-COVID era, it was observed different trends in model performance. The first LSTM model, which primarily focused on electricity consumption (KWh) while maintaining a simpler architecture, achieved a relatively low R^2 value of 0.258. This suggests that electricity consumption might be less predictable during this period using this model configuration. Conversely, the second LSTM model, with adjusted parameters (past_history=80, step=1, and future_target=6) and the inclusion of specific predictor variables such as 'Daylight,' 'Wind100int,' 'GDP,' 'Cloud,' and 'month,' delivered an outstanding R^2 value of 0.978. This ~~exceptionally~~ high accuracy highlights the effectiveness of this model configuration and the significance of carefully selected predictor variables in accurately forecasting electricity consumption, particularly in stable market conditions.

4.4 Comparison of models and model selection

To thoroughly compare the predictive model efficacy, two tables show the top-performing models for each period dataset along with the predictor variables and the corresponding MAPE, R^2 and RMSE values. In order to establish a benchmark for model evaluation and to help compare relative model performance, the mean of the test set values for each period was also computed.

Because the metrics for the LSTM model were calculated based on normalized values, a ratio between the RMSE and the test set mean was done to compare this model's predictive ability to the others.

Table 4. 8 - Results of the entire period dataset models

Model	Variables	R^2	RMSE	MAPE	Test mean
SARIMAX(1,1,1)(0,1,1,12)	GDP	0.746	142421992.5	2.139	4164958333.3
VAR(1)	KWh, GDP_detrended	0.133	280850282.1917	5.346	
SVR, C:1, epsilon:0.01, gamma: 1, kernel: poly	'GDP', 'covid', 'month'	0.150	260698279.6995	3.998	
LSTM, Past_history=127, Step=1 Future_target=12, Lstm(16, activation=relu, Dense=12	'Daylight' , 'Wind100int', 'Precipitation', 'GDP', 'covid', 'month'	0.883	14.079	7.605	

In examining Portugal's electricity consumption patterns over the entire period, it becomes evident that the predictive models exhibited varying degrees of success. When focusing on the critical metrics of RMSE and MAPE, the SARIMAX model with GDP as a predictor stands out as a reliable choice for accurate electricity consumption forecasting. Its low MAPE, around 2%, indicates it can provide forecasts with minor percentage errors compared to the test set mean value. In contrast, the VAR model, which includes KWh and GDP_detrended variables, shows less ability in prediction, as indicated by its high MAPE and RMSE, but is still well-performing a short-term forecast. The SVR model, with 'GDP,' 'covid,' and 'month' as predictors, presents a promising alternative, offering more reliable forecasts than the VAR model, although a more comprehensive assessment is required. The LSTM model, while displaying a solid fit to the data, exhibits the highest RMSE and MAPE values (around 7%), suggesting that further fine-tuning may be necessary to enhance its predictive accuracy and reduce forecasting errors.

Table 4. 9 - Results of the pre-covid period dataset models

Model	Variables	R^2	RMSE	MAPE	Test mean
SARIMAX(1,0,0)(1,0,1,12)	GDP, Wind100int	0.770	130164812.8	2.409	4250750000.0
VAR(1)	KWh, month	0.014	248666907.32	4.763	
SVR, 'kernel': 'poly', 'gamma': 0.01, 'epsilon': 0.1, 'C': 0.1	'Daylight' , 'Wind100int', 'Cloud',	0.275	231078789	3.530	
LSTM, Past_history=80, Step=1 Future_target=6, Lstm(16), activation=relu, Dense=6	'Daylight' , 'Wind100int', 'GDP', 'Cloud', 'month'	0.978	7.853	5.012	

During the pre-COVID era, the modeling landscape exhibited noteworthy differences in predictive performance compared to the entire period. When focusing on RMSE and MAPE as critical evaluation metrics, the SARIMAX model with GDP and Wind100int as predictors emerges as a reliable choice for forecasting electricity consumption. It demonstrates competitive performance with a relatively low MAPE, indicating the ability to provide accurate forecasts with small percentage errors. The VAR model, incorporating KWh and 'month' as predictors, performs a bit worse than the SARIMAX model, but the predictive accuracy is still competitive. The SVR model, with 'Daylight,' 'Wind100int,' and 'Cloud' as predictors, delivers accurate forecasts when considering RMSE and MAPE, making it a good alternative for forecasting electricity consumption. The LSTM model, despite its strong R^2, indicates potential issues with accuracy when assessed based on RMSE and MAPE, necessitating further refinement to enhance its predictive capabilities.

When comparing the two distinct periods, the pre-COVID era exhibited more favorable predictive performance in most models, signifying potential shifts in electricity consumption habits or external factors impacting consumption patterns. Comparing the entire and pre-COVID periods, the SARIMAX model emerges as a consistently reliable choice for electricity consumption forecasting, demonstrating strong accuracy in both contexts. Its robustness in capturing market fluctuations and stability in stable conditions make it a versatile option. Conversely, the VAR model exhibits limitations in both periods, suggesting its inadequacy for precise forecasting. The SVR model maintains moderate accuracy across different timeframes, while the LSTM model requires refinement to improve its forecasting capabilities in both stable and fluctuating market conditions.

The choice of model and predictor variables should be tailored to the specific period under consideration, as different time frames may require distinct modeling approaches. Therefore, the selection of the best model for each period is based on the models' respective strengths and the specific dynamics of each period.

For the entire period, the SARIMAX model with GDP as a predictor emerges as the most suitable choice. This recommendation is based on the model's consistent accuracy and robustness in capturing electricity consumption patterns across fluctuating market conditions. With a low MAPE of 2.409%, the SARIMAX model demonstrates its reliability by providing forecasts with small percentage errors relative to the high test mean. Furthermore, the SARIMAX model's interpretability is an asset. GDP as the sole predictor variable aligns with economic principles, reinforcing the model's trustworthiness in explaining electricity consumption trends.

In contrast, for the pre-COVID period characterized by relative stability, the SVR model with 'Daylight,' 'Wind100int,' and 'Cloud' predictors is recommended. This choice is based on the model's accuracy, adaptability, and suitability for stable market conditions. The SVR model delivers forecasts with a MAPE of 3.530%, indicating its capability to provide predictions with smaller percentage errors than the test set mean.

CHAPTER 5

Conclusions and Future Work

Around the world, the energy sector faces several challenges, including rising consumption and efficiency, shifting supply and demand trends, and a lack of analytics required for efficient management. Understanding the future level of energy consumption is critical for both short-term and long-term planning. Many users, from government agencies to local development authorities to financial and trading institutions, are interested in a realistic forecast of future consumption portfolios.

Portugal, in particular, has experienced energy poverty, characterized by poorly insulated buildings and very high energy expenditures, affecting a significant percentage of its population. Besides that, Portugal has a high reliance on energy import, and when such an energy dependence rate is high, accurate energy consumption prediction (directly related to the energy efficiency indicator) is essential. This is because, through this, energy-related problems can be effectively addressed along with planning the stable growth of the economy. In response to these concerns, this study aimed to forecast Portugal's electricity consumption using publicly accessible data sources and machine learning models through stable (pre-covid) and non-stable (post-covid) periods. Its insights may be extrapolated to provide valuable guidance for developing energy policies and improving energy efficiency, with broad ramifications for energy conservation.

The methodology chosen for this study was the CRISP-DM, which has six action stages (business understanding, data understanding, data preparation, modeling, evaluation, and deployment). The study included a thorough investigation of Portugal's twelve-year electricity consumption trends, revealing underlying patterns like seasonality, trends, and notable anomalies.

The objective of this analysis was to comprehend and predict Portugal's electricity consumption, with a particular focus on the COVID-19 pandemic's impact, utilizing machine learning techniques. The investigation included a look at historical consumption patterns, a look at pandemic effects, an assessment of the impact of exogenous factors (such as weather factors - temperature, humidity, and precipitation - economic factors, and external - COVID - factors) on consumption, and a look at various machine learning models.

This was possible by using data compiled from various online sources and then adjusted for better data analysis. It includes COVID-19 data, weather data, GDP data, monthly electrical consumption data, and annual electrical consumption data by type. This data helped to understand and gather insights into the problem and ultimately helped with the predictive model selection. Moreover, when projecting total energy consumption, the period under consideration was divided into - the relatively stable market era before the COVID-19 epidemic and the unstable market with data before and after the epidemic.

When analyzing the data quality, the Pearson Correlation Coefficient and Eta-correlation ratio were used to identify the correlations between variables. Higher temperatures are linked to reduced energy consumption, potentially due to the efficiency of cooling systems, while windier days correspond to slightly increased energy usage, likely driven by heating or cooling demands. As reflected in GDP, economic growth is associated with higher energy consumption due to expanding industrial and commercial activities. Longer daylight hours reduce energy usage, as natural light reduces the need for artificial lighting and heating. Cloud cover and precipitation have weak positive correlations with energy consumption, possibly indicating increased indoor activities during cloudy or rainy days. Humid conditions are linked to energy savings, and these patterns persist even before the COVID-19 pandemic, emphasizing the intricate interplay of meteorological and economic variables in shaping energy consumption trends and providing valuable insights for energy management and forecasting. These thorough evaluations influenced each period's final set of independent variables, providing a solid foundation for the predictive models and subsequent analyses.

The structure for subsequent modeling was established during the data analysis phase, which revealed vital insights into variable relationships and energy consumption patterns. Stationarity, seasonality, parameter identification, and model validation were the main focuses of the iterative modeling process.

Five different predictive models—SARIMA, SARIMAX, VAR, SVR, and LSTM—were evaluated to determine how well they predicted electricity consumption. These conclusions highlight the significance of tailoring predictive models to the specific characteristics of the dataset and market conditions. The SARIMAX model with GDP as a predictor consistently delivers accurate forecasts, making it an excellent choice for capturing electricity consumption variations over time. In contrast, the VAR model struggles to maintain accuracy in both periods, indicating its less ability for this particular task. The SVR model's moderate accuracy and adaptability in the pre-COVID period emphasize the importance of considering different modeling approaches for stable and unstable market conditions. However, its performance should be evaluated cautiously, given the SARIMAX model's consistent superiority. The LSTM model's potential, indicated by a strong R^2, suggests that it may benefit from further fine-tuning and variable selection to enhance its predictive capabilities and mitigate the high RMSE and MAPE observed in both periods.

It is evident that no single model can adequately address the diverse dynamics of energy consumption. The appropriate model selection must align with the dataset's specific characteristics and temporal considerations.

As Portugal grapples with energy-related challenges, the findings of this study can serve as a valuable resource for policymakers, energy providers, and researchers seeking to enhance energy efficiency, address energy poverty, and plan for a sustainable energy future.

The thesis study advances energy forecasting by using publicly available data to estimate Portugal's electricity consumption.

First, it recognizes the crucial need to account for external shocks, as exemplified by the inclusion of the COVID-19 pandemic's impact on energy consumption. This consideration reflects the growing importance of adapting forecasting models to address unforeseen disruptions in energy markets.

Secondly, the study promotes a holistic modeling approach by employing various predictive models, including SARIMA, SARIMAX, VAR, SVR, and LSTM. This multifaceted strategy underscores the intricate nature of electricity consumption dynamics and highlights the value of incorporating diverse factors such as economic indicators, meteorological data, and exogenous variables.

Moreover, the meticulous variable selection and analysis process underscores the significance of understanding the nuanced relationships between different factors and electricity consumption. This insight can guide future research in more precise feature engineering and variable selection, thereby improving the accuracy of energy forecasting models.

Finally, the temporal aspect of the study, differentiating between stable and non-stable market periods, underscores the necessity of temporal considerations in energy forecasting. The findings emphasize the importance of adaptive modeling techniques that can adapt to changing consumption patterns over time, thus enhancing predictive accuracy.

However, it is essential to recognize some restrictions and set a course for upcoming research projects. First and foremost, it's crucial to understand that selecting datasets and variable configurations can significantly affect model analysis and predictive accuracy, especially when dealing with periods of data scarcity. The availability of enhanced and real-time data, including high-resolution weather information and more comprehensive economic indicators, can significantly contribute to more accurate predictions. Improved data quality and accessibility will be crucial for advancing forecasting capabilities.

Secondly, exploring advanced machine learning techniques like deep learning models, ensemble methods, and reinforcement learning could yield even more precise forecasts, especially in complex and dynamic energy markets.

Thirdly, researchers can focus on identifying and integrating additional exogenous variables that influence electricity consumption, such as policy changes, technological innovations, and social trends. This expanded scope of variables can further refine forecasting models.

In the future, when there is more data regarding the period after the COVID-19 pandemic, an identical study can be performed to understand the changes in the performance of these machine learning models to a dataset containing only pos-covid values.

Additionally, spatial analysis could be employed to predict electricity consumption at regional or local levels, acknowledging the variations in consumption patterns across different areas within Portugal. Furthermore, developing models with enhanced interpretability can aid stakeholders in gaining a deeper understanding of the drivers behind electricity consumption, facilitating better-informed decision-making. This study can also be expanded by incorporating more sophisticated data treatment methods for daily and monthly data. Investigating alternative models for particular data aggregations, like daily data, may also be a helpful strategy for enhancing predictive accuracy.

These research directions aim to address the issues brought about by external factors like the Covid-19 pandemic while improving the robustness and applicability of energy consumption forecasting models in a constantly changing energy environment. Finally, future research should consider sustainability aspects, including the transition to renewable energy sources, carbon emissions, and environmental impact, to align energy forecasting with evolving ecological and policy considerations.

In conclusion, this study represents a significant step forward in energy forecasting, addressing the challenges posed by external shocks and advocating for comprehensive modeling approaches. Future research should build upon these foundations by leveraging advanced techniques, incorporating additional variables, and embracing emerging technologies to enhance the precision and applicability of energy forecasting models in a rapidly evolving energy landscape.

References

Abuella, M., & Chowdhury, B. (2017, April). Random forest ensemble of support vector regression models for solar power forecasting. In 2017 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT) (pp. 1-5). IEEE. https://doi.org/10.1109/ISGT.2017.8086027

Ahmed, M. Z., Ahmed, O., Aibao, Z., Hanbin, S., Siyu, L., & Ahmad, A. (2020). Epidemic of COVID-19 in China and associated Psychological Problems. Asian Journal of Psychiatry, 51, 102092. https://doi.org/10.1016/j.ajp.2020.102092

Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An Empirical Comparison of Machine Learning Models for Time Series Forecasting. Econometric Reviews, 29(5-6), 594–621. https://doi.org/10.1080/07474938.2010.481556

Ahmad, T., & Chen, H. (2020). A review on machine learning forecasting growth trends and their real-time applications in different energy systems. Sustainable Cities and Society, 54. https://doi.org/10.1016/j.scs.2019.102010

Ahmad, T., & Zhang, D. (2020). A critical review of comparative global historical energy consumption and future demand: The story told so far. Energy Reports, 6, 1973-1991.

Aktar, M. A., Alam, M. M., & Al-Amin, A. Q. (2021). Global economic crisis, energy use, CO2 emissions, and policy roadmap amid COVID-19. Sustainable Production and Consumption, 26, 770-781.

Almalaq, A., & Edwards, G. (2017, December). A review of deep learning methods applied on load forecasting. In 2017 16th IEEE international conference on machine learning and applications (ICMLA) (pp. 511-516). IEEE. https://doi.org/10.1109/ICMLA.2017.0-110

Alam, M. M., & Murad, M. W. (2020). The impacts of economic growth, trade openness and technological progress on renewable energy use in organization for economic co-operation and development countries. Renewable Energy, 145, 382-390.

Alsaedi, Y. H., & Tularam, G. A. (2019). The relationship between electricity consumption, peak load and GDP in Saudi Arabia: A VAR analysis. Mathematics and Computers in Simulation. doi: 10.1016/j.matcom.2019.06.012.

Atalla, T. N., & Hunt, L. C. (2016). Modelling residential electricity demand in the GCC countries. Energy Economics, 59, 149-158. https://doi.org/10.1016/j.eneco.2016.07.02

Banos, R., Manzano-Agugliaro, F., Montoya, F. G., Gil, C., Alcayde, A., & Gómez, J. (2011). Optimization methods applied to renewable and sustainable energy: A review. Renewable and Sustainable Energy Reviews, 15(4), 1753-1766. https://doi.org/10.1016/j.rser.2010.12.008

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion. doi: 10.1016/j.inffus.2019.12.012.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Box G., Jenkins G., Reinsel G., Ljung G. (2016). Time Series Analysis: Forecasting and Control. John Wiley & Sons, Fifth edition, New Jersey.

Campbell, Alrick (2018). Price and income elasticities of electricity demand: Evidence from Jamaica. Energy Econ. 69, 19–32. https://doi.org/10.1016/j.eneco.2017.10.040

Carvalho, M., de Mello, B., Delgado, D., de Lima, K. M., de Camargo, C. M., dos Siqueira, C. A., ... et al. (2020). Effects of the COVID-19 pandemic on the Brazilian electricity consumption patterns. Int J Energy Res 2020. https://doi.org/10.1002/er.5877

Chakraborty, I., & Maity, P. (2020). COVID-19 outbreak: Migration, effects on society, global environment and prevention. Science of the Total Environment, 728, 138882.

Chatziantoniou, I., Gabauer, D., & de Gracia, F. P. (2022). Tail risk connectedness in the refined petroleum market: A first look at the impact of the COVID-19 pandemic. Energy Economics, 111, 106051.

Cialani, C., & Mortazavi, R. (2018). Household and industrial electricity demand in Europe. Energy policy, 122, 592-600. https://doi.org/10.1016/j.enpol.2018.07.06

Chen, B. J., & Chang, M. W. (2004). Load forecasting using support vector machines: A study on EUNITE competition 2001. IEEE transactions on power systems, 19(4), 1821-1830. https://doi.org/10.1109/tpwrs.2004.835679

Chou, J.-S., & Tran, D.-S. (2018). Forecasting Energy Consumption Time Series using Machine Learning Techniques based on Usage Patterns of Residential Householders. Energy. doi: 10.1016/j.energy.2018.09.144.

Cihan, P. (2022). Impact of the COVID-19 lockdowns on electricity and natural gas consumption in the different industrial zones and forecasting consumption amounts: Turkey case study. International Journal of Electrical Power and Energy Systems, 134. https://doi.org/10.1016/j.ijepes.2021.107369

Creutzig, F. (2022). Fuel crisis: Slash demand in three sectors to protect economies and climate. Nature, 606, 460–462.

CNA. (2021). Europe's Recent Electricity Price Surge Inevitable If It Wants Green Energy. Available online: https://www.channelnewsasia.com/commentary/europe-eu-electricity-price-green-energy-transition (accessed on December 19, 2021).

Deb C, Zhang F, Yang J, Lee SE, Shah KW. (2017). A review on time series forecasting techniques for building energy consumption. Renewable and Sustainable Energy Reviews, 74, 902-924.

Debnath, K. B., & Mourshed, M. (2018). Forecasting methods in energy planning models. Renewable and Sustainable Energy Reviews, 88, 297–325. https://doi.org/10.1016/j.rser.2018.02.002

Enders, W., & Jones, P. (2015). Grain prices, oil prices, and multiple smooth breaks in a VAR. Studies in Nonlinear Dynamics & Econometrics. https://doi.org/10.1515/snde-2014-0101

Feng, T., Zhou, Z., Yu, J., Liu, M., Li, M., Jia, H., & Yu, X. (2022). The comparative analysis of SARIMA, Facebook Prophet, and LSTM for road traffic injury prediction in Northeast China. Frontiers in Public Health, 10. doi: 10.3389/fpubh.2022.946563.

Fotsing, F. I. M., Donatien, N., Réné, T. (2014). Influence of socio-economic indicators on electricity consumption of low voltage customers in Cameroon. International Journal of Energy Power Eng., 34, 186–203.

Fulzele, R., Fulzele, V., & Dharwal, M. (2021). Mapping the Impact of COVID-19 Crisis on the Progress of Sustainable Development Goals (SDGs)—A focus on Global Environment and Energy Efficiencies. Materials Today Proceedings, 60, 873–879.

Ghiani E, Galici M, Mureddu M, Pilo F. Impact on electricity consumption and market pricing of energy and ancillary services during pandemic of COVID-19 in Italy. Energies 2020;13:3357. https://doi.org/10.3390/en13133357

Ghoddusi, H., Creamer, G. G., & Rafizadeh, N. (2019). Machine learning in energy economics and finance: A review. Energy Economics, 81, 709–727. https://doi.org/10.1016/j.eneco.2019.05.006

Gulati, P., Kumar, A., & Bhardwaj, R. (2021). Impact of Covid19 on electricity load in haryana (india). International Journal of Energy Research, 45(2), 3397-3409. https://doi.org/10.1002/er.6008

Halbrügge S, Schott P, Weibelzahl M, Buhl HU, Fridgen G, Schöpf M. How did the German and other European electricity systems react to the COVID-19 pandemic? Appl Energy 2021;285:116370. https://doi.org/10.1016/j.apenergy.2020.116370

Hasanov, F. J. (2021). Theoretical framework for industrial energy consumption revisited: The role of demographics. Energy Reports, 7, 2178-2200. https://doi.org/10.1016/j.egyr.2021.04.002

Hasanov, F. J., Bulut, C., & Suleymanov, E. (2016). Do population age groups matter in the energy use of the oil-exporting countries?. Economic Modelling, 54, 82-99. https://doi.org/10.1016/j.econmod.2015.12.018

Hasanov, F. J., Hunt, L. C., & Mikayilov, C. I. (2016). Modeling and forecasting electricity demand in Azerbaijan using cointegration techniques. Energies, 9(12), 1045.

Hasanov, F. J. (2021). Theoretical framework for industrial energy consumption revisited: The role of demographics. Energy Reports, 7, 2178-2200.

Hank, C., Sternberg, A., Köppel, N., Holst, M., Smolinka, T., Schaadt, A., ... Henning, H.-M. (2020). Energy efficiency and economic assessment for imported energy carriers based on renewable electricity. Sustainable Energy & Fuels. https://doi.org/10.1039/D0SE00067A

Hauser, P., Anke, C. P., López, J. G., Möst, D., Scharf, H., Schönheit, D., & Schreiber, S. (2020). The impact of the COVID-19 crisis on energy prices in comparison to the 2008 financial crisis. In IAEE energy forum/Covid-19 issue. Cleveland: International Association for Energy Economics.

Hyndman R., Fan S. (2015). Monash Electricity Forecasting Model. Monash University. https://robjhyndman.com/papers/MEFMR1.pdf

Hyndman R., Athanasopoulos G. (2018). Forecasting: Principles and Practice. OTexts, 2nd edition. https://www.otexts.org/fpp

Jaisumroum, N., & Teeravaraprug, J. (2017). The consideration of population and economic growth in Thailand's electricity consumption. Paper presented at the 2017 4th International Conference on Industrial Engineering and Applications, ICIEA 2017, 331-335. https://doi.org/10.1109/IEA.2017.7939232

Javid, Muhammad (2015). Carbon emissions and oil consumption in Saudi Arabia. Renewable and Sustainable Energy Reviews, 48, 105-111. https://doi.org/10.1016/j.rser.2015.03.072

Jianfeng Liu, Xiaowei Tang (2012). Discussion of energy consumption and economic growth: based on maximum entropy. Statistics and Information Forum, vol.144, no. 9, pp. 45-51.

Kalimoldayev, M., Drozdenko, A., Koplyk, I., Marinich, T., Abdildayeva, A., & Zhukabayeva, T. (2020). Analysis of modern approaches for the prediction of electric energy consumption. Open Engineering, 10(1), 350-361. https://doi.org/10.1515/eng-2020-0028

Kavaklioglu, K. (2011). Modeling and prediction of Turkey's electricity consumption using support vector regression. Applied Energy, 88(1), 368-375. https://doi.org/10.1016/j.apenergy.2010.07.021

Kaytez, F. (2020). A hybrid approach based on autoregressive integrated moving average and least-square support vector machine for long-term forecasting of net electricity consumption. Energy. https://doi.org/10.1016/j.energy.2020.117200

Khan, A., Chiroma, H., Imran, M., Khan, A., Bangash, J. I., Asim, M., ... Aljuaid, H. (2020). Forecasting electricity consumption based on machine learning to improve performance: A case study for the Organization of Petroleum Exporting Countries (OPEC). Computers and Electrical Engineering, 86. https://doi.org/10.1016/j.compeleceng.2020.106737

Khondaker, A. N., Rahman, S. M., Malik, K., Hossain, N., Abdur Razzak, S., & Khan, R. A. (2015). Dynamics of energy sector and GHG emissions in Saudi Arabia. Climate Policy, 15(4), 517-541. https://doi.org/10.1080/14693062.2014.937387

Klæboe, G., Eriksrud, A. L., & Fleten, S. E. (2015). Benchmarking time series based forecasting models for electricity balancing market prices. Energy Systems, 6(1), 43-61. https://doi.org/10.1007/s12667-013-0103-3

Lee, J. W., & McKibbin, W. J. (2004). Globalization and disease: The case of SARS. Asian Economic Papers, 3(1), 113-131.

Li, K., Yang, Z., Li, D., Xing, Y. Y., & Nai, W. (2020, June). A Short-Term Forecasting Approach for Regional Electricity Power Consumption by Considering Its Co-movement with Economic Indices. In 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC) (pp. 551-555). IEEE.

Luis, G., Esteves, J., & Da Silva, N. P. (2020). Energy forecasting using an ensemble of machine learning methods trained only with electricity data. Paper presented at the IEEE PES Innovative Smart Grid Technologies Conference Europe, 2020-October 449-453. https://doi.org/10.1109/ISGT-Europe47291.2020.9248865

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. PloS one, 13(3).

Martins, F., Moura, P., & de Almeida, A. T. (2022). The Role of Electrification in the Decarbonization of the Energy Sector in Portugal. Energies, 15(5), 1759.

Mundaca, G. (2017). How much can CO2 emissions be reduced if fossil fuel subsidies are removed?. Energy Economics, 64, 91-104. https://doi.org/10.1016/j.eneco.2017.03.01

Narayan, P. K., & Popp, S. (2012). The energy consumption-real GDP nexus revisited: Empirical evidence from 93 countries. Economic Modelling, 29(2), 303-308. https://doi.org/10.1016/j.econmod.2011.10.016

Nguyen-Truong, H. T., & Le, H. M. (2015). An implementation of the Levenberg-Marquardt algorithm for simultaneous-energy-gradient fitting using two-layer feed-forward neural networks. Chemical Physics Letters, 629, 40–45. https://doi.org/10.1016/j.cplett.2015.04.019

Nwulu, N. I., & Agboola, O. P. (2012). Modelling and predicting electricity consumption using artificial neural networks. Paper presented at the 2012 11th International Conference on Environment and Electrical Engineering, EEEIC 2012 - Conference Proceedings, 1059-1063. https://doi.org/10.1109/EEEIC.2012.6221536

OECD. (2011). Towards green growth. Retrieved from https://www.oecd.org/greengrowth/greening-energy/49157219.pdf

Omer, M. A., & Noguchi, T. (2020). A conceptual framework for understanding the contribution of building materials in the achievement of Sustainable Development Goals (SDGs). Sustainable Cities and Society, 52, 101869.

Pietrosemoli, L., & Rodríguez-Monroy, C. (2019). The Venezuelan energy crisis: Renewable energies in the transition towards sustainability. Renewable and Sustainable Energy Reviews, 105, 415–426. https://doi.org/10.1016/j.rser.2019.02.014

Pradeep, K., & Likhita, T. (2022, March). Machine-Learning Based Approach to Predict Energy Consumption of India States. In 2022 6th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1153-1156). IEEE.

Prol JL, Sungmin O. Impact of COVID-19 measures on short-term electricity consumption in the most affected EU countries and USA states. iScience 2020;23: 101639. https://doi.org/10.1016/j.isci.2020.101639

Rahman, M. N., Esmailpour, A., & Zhao, J. (2016). Machine learning with big data an efficient electricity generation forecasting system. Big Data Research, 5, 9-15. https://doi.org/10.1016/j.bdr.2016.02.002

Ruan G, Wu D, Zheng X, Zhong H, Kang C, Dahleh MA, ... et al. (2020). A cross-domain approach to analyzing the short-run impact of COVID-19 on the US electricity sector. Joule 2020;4:2322–37. https://doi.org/10.1016/j.joule.2020.08.017

Sapankevych, N. I., & Sankar, R. (2009). Time series prediction using support vector machines: a survey. IEEE Computational Intelligence Magazine, 4(2), 24-38. https://doi.org/10.1109/MCI.2009.932254

Samseth, E., Stockhausen, F., Veillard, X., Weiss, A. (2022). Five Trends Reshaping European Power Markets. Available online: https://www.mckinsey.com/industries/electric-power-and-natural-gas/our-insights/five-trends-reshaping-europeanpower-markets

Scheidt, F. V., Medinová, H., Ludwig, N., Richter, B., Staudt, P., & Weinhardt, C. (2020). Data analytics in the electricity sector – A quantitative and qualitative literature review. Energy and AI, 1. https://doi.org/10.1016/j.egyai.2020.100009

Sharma, A., & Kakkar, A. (2018). Forecasting daily global solar irradiance generation using machine learning. Renewable and Sustainable Energy Reviews, 82, 2254-2269. https://doi.org/10.1016/j.rser.2017.08.066

Shahbaz, M., Dube, S., Ozturk, I., & Jalil, A. (2015). Testing the environmental Kuznets curve hypothesis in Portugal. International Journal of Energy Economics and Policy, 5(2), 475-481.

Shin S-Y, Woo H-G (2022). Energy Consumption Forecasting in Korea Using Machine Learning Algorithms. Energies, 15(13), 4880. https://doi.org/10.3390/en15134880

Štreimikienė, D. (2012). World economic forum 2012. Intellectual Economics, 6(1), 806-810.

Sruthi, P. L., & Raju, K. B. (2021, November). Prediction of the COVID-19 pandemic with Machine Learning Models. In 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 474-481). IEEE.

Su Y.–W. (2018). Electricity demand in industrial and service sectors in Taiwan. Energy Efficiency, 11(6), 1541–1557. https://doi.org/10.1007/s12053-018-9615-y

The World Bank. (2022). Commodity Markets Outlook: The Impact of the War in Ukraine on Commodity Markets. Washington, DC, USA: The World Bank.

The World Bank. Global Economic Prospects. Available online: https://www.worldbank.org/en/publication/global-economicprospects (accessed on June 10, 2022)

Tso, G. K. F., & Yau, K. K. W. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy, 32(9), 1761-1768. https://doi.org/10.1016/j.energy.2006.11.010

Vagropoulos, S. I., Chouliaras, G. I., Kardakos, E. G., Simoglou, C. K., & Bakirtzis, A. G. (2016). Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting. In 2016 IEEE International Energy Conference (ENERGYCON) (pp. 1-6). doi: 10.1109/ENERGYCON.2016.7514029.

Verdejo H., Awerkin A., Becker C., Olguin G. (2017). Statistic linear parametric techniques for residential electric energy demand forecasting. A review and an implementation to Chile. Renewable and Sustainable Energy Reviews, 74, 512–521. https://doi.org/10.1016/j.rser.2017.01.110

Vishnuraja, P., Sangeetha, K., Abinaya, C., Deepika, R., & Dinesh, S. (2020). Predicting energy consumption using machine learning methods. International Journal of Advanced Science and Technology, 29(3 Special Issue), 1004-1010.

Vivas, E., Allende-Cid, H., & Salas, R. (2020). A systematic review of statistical and machine learning methods for electrical power forecasting with reported MAPE score. Entropy, 22(12), 1-24. https://doi.org/10.3390/e22121412

Walther, J., & Weigold, M. (2021). A systematic review on predicting and forecasting the electrical energy consumption in the manufacturing industry. Energies, 14(4). https://doi.org/10.3390/en14040968

Wang, R., Zhang, H., Shi, F., Zhang, Y., & Zhang, L. (2018). Empirical study of the relationship between global energy consumption and economic growth. Paper presented at the CIEEC 2017 - Proceedings of 2017 China International Electrical and Energy Conference, 394-399. https://doi.org/10.1109/CIEEC.2017.8388480

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. International Journal of Forecasting, 30(4), 1030-1081. https://doi.org/10.1016/j.ijforecast.2014.08.008

Yanwu Chen, Chengye Wu (2003). Analysis of long-term equilibrium relationship between GDP and energy consumption in Taiwan. Journal of Huaqiao University, 3, 26-31.

Yu R, Gao J, Yu M, Lu W, Xu T, Zhao M, et al. (2019). LSTM-EFG for wind power forecasting based on sequential correlation features. Future Generation Computer Systems, 93, 33–42. doi: 10.1016/j.future.2018.09.054.

Zakeri, B., Paulavets, K., Barreto-Gomez, L., Echeverri, L. G., Pachauri, S., Boza-Kiss, B., ... Pouya, S. (2022). Pandemic, War, and Global Energy Transitions. Energies, 15(17), 6114.

Zhang, C., Zhou, K., Yang, S., & Shao, Z. (2017). Exploring the transformation and upgrading of China's economy using electricity consumption data: A VAR–VEC based model. Physica A: Statistical Mechanics and Its Applications, 473, 144–155. doi:10.1016/j.physa.2017.01.004

Appendix

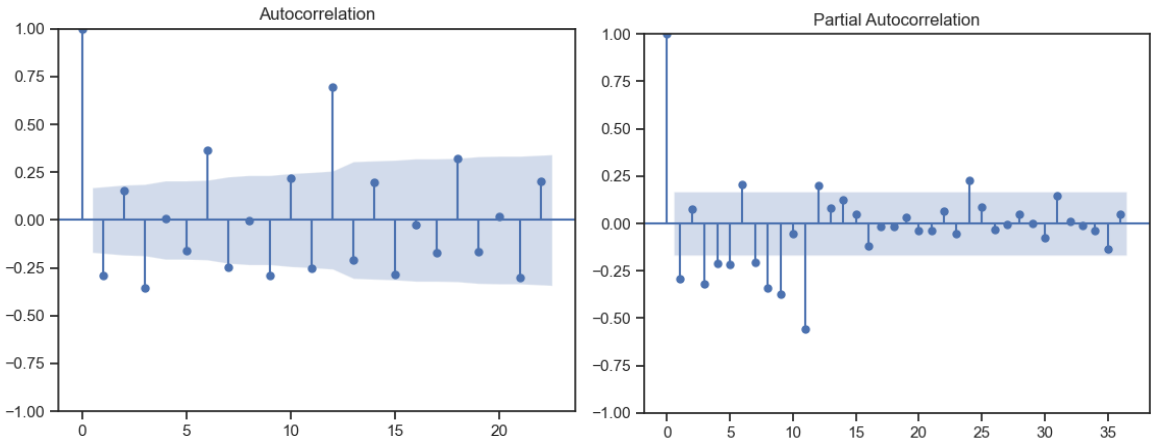

Figure 1 - ACF and PACF of the entire period dataset

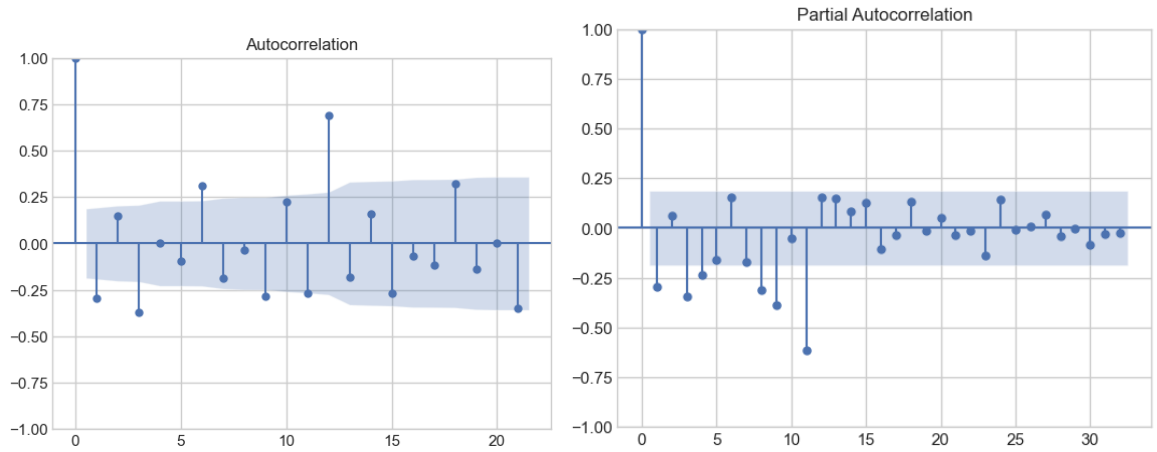

Figure 2 - ACF and PACF of the pre-covid period dataset

Table 1 - VIF tests for the entire and pre-covid periods

Entire period	Features	VIF
0	const	8122.246783
1	Temperature	254.103606
2	Cloud	5.469940
3	Daylight	4.378297
4	Wind100int	2.798599
5	Humidex	224.959510
6	Precipitation	4.644341
7	GDP	3.472925
8	covid	3.432013
9	month	1.917578
Pre-covid	Features	VIF
0	const	14525.692150
1	Temperature	245.494390

2	Cloud	5.470317