A Comparative Study for Energy Price Prediction Using Different Machine Learning Algorithms

by

Beksy Saji George

Submitted to the Department of Computing

in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

South East Technological University, Ireland

August 2023

Beksy Saji George

Author . . . .

Department of Computing

August 16, 2023

Certified by. . . . .

Lei Shi

Assistant Professor

Thesis Supervisor

Accepted by . . . .

Nigel Whyte

Chairman, Department Committee on Graduate Thesis

DECLARATION

- I declare that all of the content in this work, with the exception of those duly acknowledged, is entirely mine.

- I have cited the sources of all quotations, paraphrases, summaries of information, tables, diagrams, or other material; including software and other electronic media in which intellectual property rights may reside.

- I have provided a complete bibliography of all works and sources used in the preparation of this submission.

- I am aware that failing to adhere to the University's anti-plagiarism policies is a serious offense.

Student Name: Beksy Saji George

Student Number: C00290800

Signature: ______________________

Date: 16.08.2023

ACKNOWLEGEMENTS

I am grateful and humbled as I reflect on the outcome of my dissertation, which seemed to be a journey that was filled with blessings and support. Firstly, I express my greatest appreciation to the Heavenly father for his grace and guidance during this project.

I express my sincere gratitude to my supervisor, Dr. Lei Shi, whose consistent support, effective guidance, and extensive knowledge in the field have been of immense value to me.

I am thankful to my family's unwavering support, whose love and prayers have provided me with steady strength throughout my academic endeavour.

I would also like to express my sincere gratitude to all individuals who have made direct or indirect contributions to the successful completion of this dissertation. The valuable insights, feedback, and encouragement provided by everyone have significantly enhanced to achieve the objectives of this project.

CONTENTS

List of Abbreviations

API - Application Programming Interface

ARIMA- Autoregressive Integrated Moving Average Model

AR - Auto-Regressive Model

CSV – Comma Separated Values

EDA - Exploratory Data Analysis

ENTSOE - European Network of Transmission System Operators for Electricity

EPF - Energy Price Forecasting

FIT - Feed-in Tariff

GARCH - Generalized Auto-Regressive Conditional Heteroskedastic Model

IoT – Internet of Things

ISO - International Organization for Standardization

IQR - Inter Quartile Range

LASSO - Least Absolute Shrinkage and Selection Operator

LightGBM – Light Gradient Boosting Machine

LMP - Locational Marginal Prices

LSTM – Long Short-Term Memory

MAE – Mean Absolute Error

ML – Machine Learning

RES - Renewable Energy Sources

RMSE – Root Mean Squared Error

TSO - Transmission System Operator

XGBoost – Extreme Gradient Boosting

List of Figures

Abstract

A technique for estimating future electric or alternative energy source costs is called energy price prediction. The cost of power fluctuates substantially on daily basis on the international market, making both consumers and manufacturers unable to plan ahead. The importance of utilizing various models to predict future electricity prices has increased globally in recent years. We could deploy cutting-edge the use of machine learning systems that can precisely predict future electricity costs by using historic and current market data. By assessing the approaches and models used in machine learning algorithms, this work attempts to analyse the complicated difficulties related to the forecast of power costs. It will aim to analyse how each approach is used for identifying market trends, studying data sets, and creating prediction models. We may learn more regarding what the potential holds for the price of electricity and how we might use machine learning to create precise predictions by examining the many types of algorithms that use machine learning used to forecast energy prices.

1. Introduction

1.1 Background and Motivation

The environment around the world's electricity markets has seen significant changes over the past few decades, which are defined by market liberalization and the increasing importance of energy production from renewable sources. Energy price forecasting has been the topic of extensive research thanks to this liberalization trend and an increase in the accessibility and quality of data on the price of energy and related time series (Cochell *et al.* 2012). For market participants, governments, and researchers, understanding the mechanics of price generation and creating trustworthy forecasting models has grown crucial.

After the fiscal year 2000, energy price prediction as a field of research grew in prominence, with a notable increase in the total number of published works. There are countless approaches to solving these forecasting problems thanks to the availability of openly available data on energy pricing as well as developments regarding technology and statistical methodology (Cochell *et al.*, 2012). While there has been a lot of research on short-term forecasting, particularly the ability to anticipate power costs up to one day in advance at an hourly resolution, there hasn't been much done to explore methods for over time price forecasting.

It should be emphasized that the majority of existing techniques for predicting electricity prices have concentrated on short-term forecasts, relying on fundamental models that reflect the system's behaviour and predicted pricing functions of market rivals. However, these models usually lack real time series data, which makes it difficult for them to create accurate hourly resolution estimates, which are essential for day-ahead markets (Wang *et al.*, 2017).

Longer-term power price forecasting is important because it can help market participants, policymakers, and researchers make better decisions. (Pallonetto, Jin, and Mangina,2012) Energy providers may enhance their generation and trading approaches with the help of accurate long-term forecasts, better aligning their operations with market dynamics. Consumers can benefit from better energy cost planning and management and governments can create enhanced energy regulations to promote sustainable and effective energy systems. Additionally, researchers can improve forecasting techniques and gain a better understanding of the shifting patterns of the electricity market.

The electricity industry has the issue of managing the risks and uncertainties associated with the market pricing in the context of the energy transition and the rising reliance on renewable energy sources. This difficulty is compounded by the fact that the electricity sector is increasingly dependent on renewable energy sources. (Gholipour Khajeh et al., 2017).This makes it important to think about how prices will change in the future when making investment decisions. Given that renewable energy assets require considerable up-front investments and have lengthy amortization periods, it is essential to accurately estimate future cash flows when assessing the economic feasibility of projects.

By focusing solely on the projections of future generations, Feed-in Tariff previously provided a straightforward technique for estimating future revenues. As a result, the procedure was quite simple to comprehend. (Gabrielli et al., 2022) However, the transition from FITs to market-based processes demands the inclusion of projected revenue prices in the future market. The need for electricity price forecasts (EPF) has increased as a result of this trend, particularly for funding renewable energy initiatives. It's critical to anticipate market changes in order to calculate the potential profit and risk of a renewable energy project.

1.2 Objective

The necessity to predict and schedule the costs related to energy usage gave rise to the idea of forecasting energy prices. As we all know, a number of variables, such as shifts in demand and supply, fuel costs, weather, political turmoil, and regulatory rules, can cause significant fluctuations in energy prices. However, these modifications have an impact on the cost, accessibility, and viability of renewable power sources in national and international energy markets. (Stefenon et al.,2017)

This study's aim is to offer useful information to stakeholders in the rapidly changing energy sector. Here, the suggested methodology has an opportunity to increase stakeholders' understanding of future electricity pricing, allowing them to improve their strategies and make better-informed choices while employing renewable energy.

1.3 Problem Statement

Because of different factors affecting power generation, such as renewable energy sources, deregulated markets, and shifting demand patterns, the energy market has progressively grown more complex. As a result, it has become extremely difficult to anticipate hourly electricity rates while keeping an eye on energy and climate factors. Even though numerous studies have been carried out using different machine learning algorithms, it is still unclear which model is the best for properly forecasting hourly electricity costs.

This project seeks to conduct a comparison analysis for predicting power costs utilizing several machine learning and deep learning algorithms, primarily LSTM (Long Short-Term Memory), XGBoost, and Light Gradient Boosting Machine (LightGBM) algorithms, in order to eliminate these inconsistencies.

The Research Question for this thesis is:-

What is the best machine learning model for predicting future hourly electricity costs based on energy and weather features?

With the help of performance metrics analysis, this study will attempt to determine how well each model performs. The study will next provide an exhaustive overview of the data set comprising the energy and meteorological features. The following parts will meticulously review the numerous research articles to identify the gaps in those studies. The outcomes and conclusions of this investigation will be described after properly handling the data.

2. Background and related work

This literature review explores the methodology, data sources, and models utilized in this subject in an effort to give a thorough evaluation of the most recent studies on energy price forecasting. This review examines the strengths, shortcomings, and other trends concerning energy pricing predictions and summarizes the findings of numerous studies that have been combined. People will be able to discover more about forecasting methods and how to use them in a dynamic energy market with the use of the data gleaned from this study.

2.1 Understanding Energy Price Forecasting (EPF)

In the subject of forecasting known as energy price forecasting (EPF), prices in the commercial power markets are projected over a variety of time frames (Jedrzejewski et al.,2022) Instead of the current flat rates, real-time power pricing schemes may be beneficial for both economies and the environment. They can specifically assist end users in saving money on power by enabling them to respond to shifting costs throughout the day. (Leon-Garcia and Mohsenian-Rad ,2010).As a result, the price of energy has a significant effect on a country's economic development. The accurate prediction of electricity prices is crucial because it enables both enterprises and consumers to effectively plan and control their spending.

Furthermore, forecasting electricity costs is now essential in the current world due to a number of factors, such as the expansion of green energy sources, the liberalization of power markets, and the requirement for efficient energy management. According to Sridharan and Li, trustworthy and accurate projections of power costs are crucial for assisting customers, regulators, and market participants in making decisions. Algorithms for machine learning are computer programs that use data to develop predictions and make educated decisions (Fałdziński *et al.* 2020). These algorithms make use of the available data to find patterns and then forecast what will happen in the future. Due to their use of data-driven insights, machine learning algorithms have the potential to estimate power costs more accurately and consistently than previous techniques. This literature study examines pertinent studies and research papers to help explain why it is crucial to forecast power costs in the contemporary period.

2.2 Electricity Market Liberalization and Price Fluctuations

Over the last few decades, there has been a noticeable trend toward deregulation in the energy industry in many nations around the world. Deregulation aims to make it easier to integrate market-based and competitive procedures into the formerly monopolistic energy sector. This change has effectively turned electricity from a basic requirement for human survival into a traded good, offering market actors a variety of opportunities and challenges.

Deregulation typically entails dividing the electricity industry into separate segments, including generating, transmitting, and distributing. (Hryshchuk and Lessmann, 2018). As a result, various market forces, the interaction of demand and supply, and other factors now play a role in determining energy pricing rather than just being controlled by regulated monopolies.

Additionally, all stakeholders now face higher degrees of uncertainty as a result of the liberalization of the power markets, particularly in regard to price volatility. (Kumaappan and Anbazhagan, 2013).Market-driven methods have taken the role of the conventional regulated pricing structure, which provided secure and predictable pricing, and allow for rates to change in response to shifting market conditions and external influences. For market participants, including electrical producers, customers, investors, and policymakers, the increased volatility poses a number of difficulties. To understand how deregulation influences energy price volatility and to develop effective forecasting approaches to lower the risks involved, researcher’s Li and Flynn have conducted extensive research. Numerous factors, including the following, might affect price volatility in unregulated markets:

2.2.1 Dynamic Supply and Demand:

Within deregulated markets, the relationship between the supply and demand for electricity is critical in determining price changes. Energy can be considerably impacted by a number of variables, such as seasonal variations, demand patterns, the capacity to produce electricity, the cost of fuel, and unplanned outages. As a result, understanding and developing models for these processes is crucial for accurate price prediction (Sridharan, Tuo and Li ,2022).

2.2.2 Market dynamics and competitions:

In the modern world, deregulation often results in competitive dynamics within the energy industry, affecting market structure and competitiveness. Through strategic offers, market influence, and a response to market signals, the actions of different market actors, such as generators, however, retailers, and merchants, have the power to affect pricing (Hryshchuk and Lessmann, 2018).In order to effectively identify pricing trends, it is essential to examine market structure and competition dynamics.

2.2.3 Fuel Costs and Generation Mix:

Electricity prices in deregulated markets are significantly influenced by the cost of fuel, which includes a variety of energy sources such natural gas, coal, and green energy (Higgs et.al,2015).Fuel prices have had a significant economic impact on the amount of electricity consumed, which causes price volatility and affects the supply-demand balance generally. Therefore, a thorough understanding of the relationship between fuel and power price is necessary to enable accurate forecasting.

2.2.4 The environmental and Legal Factors:

Power prices can be significantly impacted by environmental policies and regulations such as carbon pricing, renewable energy goals, and emission reductions. To fully analyze the impact on price volatility, these factors must be included in forecasting models as they add additional inconsistencies and costs.

2.3 Efficient Market Operation

As demand and supply dynamics determine pricing in liberalized power markets, market efficiency is essential. Power price forecasting accuracy is crucial since it enables market participants to make wise decisions and enhance their operations.

The study by Weron (2014) demonstrates how crucial it is to forecast power prices in order to increase market liquidity. Market participants can prepare for price shifts and adjust their strategies when projections are accurate. They are given a better approach to manage risks as a result, and it is now simpler for them to choose when to execute deals. For instance, by altering their offers in accordance with their predictions of the pricing, generators can enhance their bidding methods. This enables them to maximize their profits while guaranteeing the most affordable energy supply. Price projections provide market participants with information on the levels and trends of energy prices in the future, enabling them to make efficient use of their resources (Li and Flynn ,2004). For instance, suppliers can determine the most cost-effective combination of sources of energy and contracts to satisfy anticipated demand. Traders can locate opportunities to profit by leveraging predicted price discrepancies between several locations or times.

The research by Gray et al. implies that by ensuring that everyone shares the same level of knowledge, precise power price predictions also contribute to the market's efficiency. The playing surface is levelled, and open competition is promoted when everybody has access to reliable pricing forecasts. This increases market efficiency by making it simpler to determine prices and allocate resources (Goretti, G. (2020). Knowing how much electricity will cost in the future also makes it easier to integrate renewable energy sources (RES) into the system.

The availability and cost of renewable energy become increasingly crucial to market players as the usage of renewable energy sources (RES) increases (Chan, Gray and van Campen 2008).Investors and consumers of RES can determine the viability of their projects and make wise investment decisions with the help of precise pricing estimates. Furthermore, the effectiveness of liberalized power markets depends in large part on accurate price forecasts. It enables market participants to maximize the efficiency of their operations, make informed decisions, and make effective use of their resources (Rodriguez and Anders 2004). The market gets more liquid, bidding strategies improve, and the market functions better when accurate price projections are used. These benefits both consumers as well as industry stakeholders by increasing the stability and efficiency of the energy markets as a whole.

2.4 Traditional Approaches in Price Forecasting Models

Traditional econometric models were widely used for electricity price prediction before the development of ML approaches. Energy markets, according to Pandey and Upadhyay (2016), vary throughout parts of the world, making it difficult to develop an ideal price forecasting method that can be used everywhere. Instead, the particular market type affects the method choice. Moreover, many businesses rely on traditional approaches to direct their decision-making processes when it comes to electricity price forecasting. In an effort to forecast future power prices, it is stated that a variety of analytical techniques, including seasonal evaluation, autoregressive approaches, and generalized additive models, are used (Lehna, Scheller and Herwartz 2022). This research discovered that in order to identify repeating seasonal trends and predict future price fluctuations, the process of seasonal analysis entails looking at historical trends of electricity prices.

Forecasts for future prices are created from previous price data via autoregressive models (Pallonetto, Jin and Mangina 2022). Therefore, the use of additive general models serves the objective of identifying correlations between price changes and outside factors, such as weather patterns, changes in demand, and overall economic conditions, among others. These models have a lot to offer as tools for predicting power costs, but their ability to give businesses a competitive edge in the electrical market is frequently constrained by their poor accuracy and dearth of all-encompassing information (Goretti, G. (2020).

Few academics feel that using the information produced from standard models will help energy companies better understand future price variations. However, this methodology has several limitations (Zhang et al., 2022).Conventional models lack accuracy and a comprehensive understanding, which makes it difficult for businesses to gain a competitive advantage in the field of renewable energy. Corporations need to incorporate more advanced approaches into their forecasting models in order to maximize the effectiveness of their energy pricing plans.

2.4.1 Autoregressive Integrated Moving Average (ARIMA) model

The forecasting of energy costs, with an emphasis on the near future, has been the subject of a significant amount of research over time. Numerous approaches have been put forth, including non-stationary time series models like the generalized auto-regressive conditional heteroskedastic (GARCH) model as well as static time series approaches such as dynamic regression, transfer function, auto-regressive integrated moving average (ARIMA), and auto-regressive (AR) models (Pandey and Upadhyay, 2016). Most of these predictive models, however, used linear predictors, which were unable to fully account for the non-linear aspects of power price dynamics.

The ARIMA model is frequently used as a method for predicting electricity prices. Due to its ability to accurately capture time-varying dependencies and changes in energy price data, ARIMA models have grown in favour. They are widely used in the energy industry because they offer a clear and simple methodology (Contreras et al. 2002).However, the ARIMA model's linear structure places limitations on its capacity to accurately reflect the complex dynamics and non-linear relationships contained in the data on energy prices.

According to additional research, Autoregressive Integrated Moving Average (ARIMA) models have been widely used in conventional energy price prediction because of their simplicity, interpretability, and ability to capture temporal interdependencies. According to Stefenon et al. (2023), ARIMA models are particularly effective in identifying trends and seasonality in data on energy prices. However, it is crucial to recognize that these models have inherent flaws that may affect their accuracy and applicability in completely capturing the complex dynamics of the energy markets.

Researchers have looked into alternative approaches, such as neural networks and fuzzy neural networks, in order to get around the limitations of the ARIMA model in the area of forecasting electricity prices (Gholipour Khajeh et al., 2017).These models are particularly suited for the purpose of modelling the non-linear behaviour displayed by electricity prices because they have the ability to successfully capture non-linear input/output mapping functions (Cerjan et al., 2013). The effectiveness of these systems depends on the painstaking adjustment of programmable parameters, including the number of hidden nodes in neural networks, and the careful selection of appropriate input features.

Additionally, due to their significant effectiveness, neural networks and efficient metaheuristic algorithms have been used to a variety of modelling and optimization challenges (Sajjad et al., 2020).Predicting local marginal prices (LMPs) is an example of using a hybrid approach that combines a system for fuzzy inference and a least-square estimation (Ziela and Steinert 2018). Prices for power series have been broken down into components with lower volatility using wavelet transforms. Following that, the ARIMA approach is used to forecast these components. Additionally, hybrid models that include identical days or neural network methods have been proposed for LMP prediction (Esteves et al., 2015). Additionally, researchers have looked into the use of hybrid algorithms that combine different forecasting tools and weighted closest neighbour techniques.

Energy price prediction research has made strides, but there is still a critical need for the development of more accurate and durable approaches (Sajjad et al., 2020). Despite being widely used, the ARIMA model has inherent limitations when it comes to capturing complex dynamics and non-linear interactions. To increase the precision and dependability of electricity price estimates, academic researchers like Tschora et al. diligently investigate novel methodology and approaches. Through the incorporation of cutting-edge machine learning algorithms, the inclusion of pertinent extra information, and consideration of the non-linear nature of energy pricing data, further advances in the accuracy of energy forecasting models can be made.

2.4.2 Autoregressive Models (AR)

In the field of predicting energy prices, autoregressive (AR) models are widely used as traditional econometric models (Pandey, Nivedita, and K. Upadhyay, 2016). These models presuppose that a parameter's current value and its historical values have a linear relationship. Because they are simple to grasp and can demonstrate how data shifts over time, they have been employed in numerous sectors, including economics, finance, and the energy sector (Yamashita et al., 2008). These models have also been widely employed in the field of power cost prediction to foresee both short-term and long-term changes in price. They have been effective at gathering seasonality and price trends that are seen in the energy markets. Additionally, due to their clarity and interpretability, AR models are appropriate for study and can serve as a baseline for forecasting systems that are more complicated.

The limitations of autoregressive models, however, affect their accuracy and usefulness for predicting energy prices. This approach's premise of linearity is one of its limitations because it might not apply to the energy markets observed in real-life situations (Esteves et al., 2015).As a result, non-linear correlations and intricate interdependencies between variables like demand and supply, weather, and shifts in governmental policy frequently have an impact on pricing (Stefenon et al., 2023).In addition, it might have trouble detecting non-linear patterns, leading to less-than-ideal predictions in energy markets which are known for their strong volatility and dynamism.

Researchers have looked into various stages of autoregressive (AR) models that use supplemental methods or adjustments in an effort to get beyond these restrictions (A. Al Metrik and A. Musleh 2022). The autoregressive integrated moving average (ARIMA) model is one illustration of a statistical modelling strategy that addresses non-stationarity in data (Sajjad et al., 2020). Both an autoregressive element and a distinction step are included in this model. By using distinguishing techniques that stabilize the time series' statistical characteristics, ARIMA models have the capacity to accurately capture both seasonal and trending components.

Additionally, to account for the influence of extraneous factors on energy costs, researchers have added exogenous variables to autoregressive (AR) models. These variables include a wide range of things, like changes in regulatory frameworks, weather patterns, economic indicators, and the generation of renewable energy. (Contreras et al., 2002). Autoregressive (AR) models can better reflect the intricate connections between energy prices and external influences that impact them through incorporating these supplementary variables (Moradzadeh *et al.* 2020). To go around these limitations, Lehna et al. (2022) have developed alternate autoregressive (AR) model iterations such the autoregressive integrated moving average (ARIMA). While autoregressive models have several drawbacks, they are nonetheless useful for predicting energy prices and provide the foundation for future developments in the field.

2.4.3 Regression Models

In order to evaluate the relationship between both independent and dependent variables, namely the historical electricity price, regression models are frequently used in the field of electricity price prediction. These models seek to accurately reflect the underlying trends and key variables that influence price volatility in the energy market (Ferreira et al., 2019). In fact, linear regression, which investigates the linear relationship between two variables by fitting a line to the data points, is the most prevalent regression model. You can also use other regression models, such polynomial regression and logarithmic regression.

Recent advancements in the field of EPF include the use of regularization approaches together with linear regression models that contain a large number of input features (Lago et al., 2021). Regularization is used to deal with instances where the number of regressors exceeds a certain threshold. The number of regressors can be effectively limited using techniques like the elastic net and the Least Absolute Shrinkage and Selection Operator (LASSO) (Pinho et al., 2022). Compared to traditional methods, regularized regression models have shown to perform better. In order to determine the linear relationship between two variables, including the price of electricity and the related demand for electricity, linear regression is the statistical method that is used (Maciejowska et al., 2016).This methodology provides useful information about how exogenous factors, such as changes in regulatory structures and variations in demand, affect power pricing.

Additionally, researchers have looked into the use of various calibration windows inside a single regression model and then combined their findings. According to Ferreira et al. (2019), this methodology makes it possible to combine data from several temporal intervals, potentially including different market patterns or seasonal patterns. By merging the outputs of different models, the researchers hope to improve forecasting precision and reinforce the trustworthiness of forecasts. To precisely fit a polynomial function to a particular dataset, for instance, the method of polynomial regression is used. This approach shows promise for predicting electricity costs in complex environments where linear linkages can fall short. In contrast, logarithmic regression seeks to create a logarithmic curve that best matches the available dataset (Alkawaz et al., 2022). Additionally, it can be used to find relationships between current and future electricity prices. When dealing with large volumes of data and significant levels of fluctuation, its usefulness in the field of power price forecasting is particularly clear (Liu et al., 2022).Regression modelling is therefore an extensively used tool for predicting power costs. Therefore, incorporating those regression models in other procedures and researching novel ideas will further improve the accuracy and reliability of EPF.

2.5 Factors Affecting EPF

It is crucial to consider the wide range of variables that have the potential to have an impact on energy costs when making estimates for those prices. The dynamics of supply and demand, changes in weather, and adjustments in regulatory rules are a few things that might affect the energy market (Higgs, Lien and Worthington 2015).An important factor in determining power pricing is the interaction between supply and demand for electricity. As is well known, a number of variables, including population growth, economic development, and advancements in energy efficiency, have influenced the demand for electricity (Ziela and Steinert 2018). Additionally, a number of factors, such as fuel costs, the efficiency of power plants, and the use of sustainable energy alternatives, have an impact on the availability of electricity.

In addition, weather has a big impact on how much electricity costs. Due to the higher costs involved with generating energy in such situations (droughts, hurricanes, etc.), the occurrence of extreme weather phenomena might cause an increase in electricity prices (Garca-Martos et al., 2013).

On the other hand, mild weather might lead to lower electricity prices since, in some cases; the costs related to generating electricity are lowered. Additionally, a number of regulatory changes may have an impact on power pricing (Dragasevic et al., 2021). As a result of the increased costs related to complying with laws, such as pollution standards and renewable energy objectives, electricity prices may increase as a result of their adoption. In summary, there are numerous variables that could potentially affect how much electricity costs (Alkawaz et al., 2022). These elements must be taken into account when making forecasts about electricity prices because they have the power to significantly affect price changes.

3. Methodology

3.1 Overview

The dataset utilized for this thesis' analysis of daily demand for electricity production and the weather is thoroughly outlined in this part. The document examines the suggested method for data analysis, defines the importance of the study, looks into possibilities for additional data analysis and data sources, and analyzes any possible problems with ethics. This part also evaluates the research's weaknesses and strengths carefully.

3.2 Data Collection

The procedure for collecting data is an important step in carrying out research on the forecasting of power costs. In order to complete the process, appropriate and reliable information must be gathered that exactly reflects the variables under thought, including electricity use, generation, pricing, and weather. This study involves the use of a data set that extends four years and includes crucial elements specific to Spain. The publicly available dataset provided the data used in this investigation on Kaggle named "Hourly energy generation demand and weather". The primary sources of this data are the ENTSOE (Transmission Service Operator) public portal ("ENTSO-E Transparency Platform" n.d.), Red Electric España (Spanish TSO) for settlement prices ("ESIOS Electricidad", 2023), and the Open Weather API for weather data ("Weather API - OpenWeatherMap" n.d.).

The ENTSOE portal, which provides a storage facility for data supplied by Transmission System Operators (TSOs), is where the information on electrical usage and generation is gathered. These data points provide a thorough examination of the trends in power consumption and generation over a four-year period. Red Electric Espaa, the Spanish Transmission System Operator (TSO) in charge of distributing pricing information, facilitates the acquisition of settlement prices, which is crucial to the characteristics of the electricity market.

Weather data is gathered using the Open Weather API in order to take the influence of the environment on energy consumption, price, and generation capacity into assessment. The dataset consists of measurements taken in Spain's top five cities by population. Weather factors, including temperature, humidity, wind speed, and precipitation, have a great deal of potential to reveal important information about how they affect electrical utilization, price dynamics, and the engagement of various generation sources. Recognizing the complex interaction between environmental factors and the structure of the power market requires the use of weather data.

Two CSV files, "energy_dataset.csv" and "weather_features.csv," containing 26.19 megabytes in compressed space, are used in the study. These files include the following parameters:

1. Generation biomass: Energy produced from biomass energy sources, such as organic materials obtained from plants or animals, is known as biomass generation.
2. generation of fossil brown coal/lignite: Energy produced from low-grade coal varieties such as brown coal or lignite.
3. generating fossil coal-derived gas: Energy produced from gas derived from fossil coal.
4. generation fossil gas: The production of energy utilizing a fossil fuel, such as natural gas.
5. generation fossil hard coal: Production of energy with hard coal, premium coal.
6. generation of fossil oil: The production of fossil energy.
7. Generation fossil oil shale: Energy produced from oil shale, a sedimentary rock that contains organic material.
8. generation fossil peat: Energy produced from peat, an organic substance with high water content.
9. generation geothermal: Energy produced by harnessing heat from the planet's interior.
10. generation hydro pumped storage aggregated: Hydroelectricity, pumped storage energy produced by hydropower facilities with pumped storage, combined.
11. generation hydro-pumped storage consumption: Energy generated through hydropower facilities with pumped storage while they are operating is referred to as generating hydro-pumped storage consumption.
12. generation hydro run-of-river and poundage: Energy generation from run-of-river hydropower plants without water storage.
13. generation hydro water reservoir: Energy produced by hydropower plants with water reservoirs.

14. generation marine: Energy produced by marine energy sources like tides and waves.
15. generation nuclear: Nuclear power facilities that produce energy.
16. generation other: Energy produced using other, unidentified energy sources.
17. generation of other renewable: The electricity from several other renewable energy sources is not specifically specified, yet it is produced.
18. generation solar: The process of producing electricity using solar energy sources, such as solar photovoltaic panels.
19. generation waste: The process of generating power from waste by converting it to energy.
20. generation wind offshore: The electricity produced by offshore wind farms.
21. generation wind onshore: Onshore wind generation is the process by which onshore wind farms produce electricity.
22. forecast solar day ahead: The anticipated solar energy output for the following day.
23. forecast wind offshore day ahead: A forecast for offshore wind energy output.
24. forecast wind onshore day ahead: The expected amount of onshore wind energy production.
25. total load forecast: The anticipated total demand for electricity over a specific time frame.
26. total load actual: The total amount of electricity demand for a specific period.
27. price day ahead: The forecasted electricity price for the following day.
28. price actual: The actual electricity price for a specific period.
29. dt_iso: The timestamp in ISO format.
30. city_name: The name of the city associated with the weather observations.
31. temp: The temperature at the time of observation.
32. temp_min: The minimum temperature measured during the observation period.
33. temp_max: The maximum temperature measured during the observation period.
34. pressure: The atmospheric pressure at the time of observation.
35. humidity: The relative humidity at the time of observation.
36. wind_speed: The wind speed at the time of observation.
37. wind_deg: The wind direction at the time of observation.
38. rain_1h: The amount of rainfall recorded in the past hour.
39. rain_3h: The amount of rainfall recorded in the past three hours.
40. snow_3h: The amount of snowfall observed in the past three hours.

41. clouds_all: The cloud cover percentage at the time of observation.
42. weather_id: A numerical code representing the particular weather condition.
43. weather_main: The main classification of the weather condition.
44. weather_description: A detailed description of the weather condition.
45. weather_icon: An icon that depicts the current weather

The data from the CSV files were cleaned before being used in the ML models in order to be faithful to the thesis topic's objective and produce the best results. Therefore, the next section of this thesis provides details about the data cleansing that was performed.

3.3 Data Cleaning

Machine learning (ML) models are commonly used in research projects, and the quality and consistency of the data used in these projects have an important effect on the ML models' performance. To deliver reliable and exact forecasting results in the framework of the price of energy prediction research, comprehensive data purification is necessary. Numerous essential steps are involved in the procedure of extracting data from large datasets, including the recognition and removal of unnecessary variables, the control of duplicate information, the finding of data that is missing, and the fixing of incorrect information (Lokesh, 2023). The careful data elimination procedure is critical in preserving important information that is required to effectively fix the prediction issue.

Considering it offers capabilities like NumPy and Pandas that make computation and data processing less complicated, Python is utilized in this investigation as the best programming language for cleaning up data. The appendices section's code is used to improve the consistency and clarity of the data-cleaning process. Also, this study makes use of exploratory data analysis (EDA) after importing the CSV files into data frames using Pandas. In order to preserve the consistency and integrity of the data, duplicates, and missing data points are found and eliminated during the Exploratory Data Analysis (EDA) process.

The dataset’s unnecessary fields must be removed in the following stage so that the forecasting of electricity prices is limited to using the correct information. This research tries to uncover connections between variables, a crucial stage in the selection of features, via the method of visualizing and evaluating certain fields. Chosen columns are subjected to the encoding process, making further research and analysis possible. Also, the two datasets, especially "energy_dataset.csv" and "weather_features.csv," are combined in order to gain important knowledge about the relationship between weather conditions and the demand for and production of power. The method of incorporating factors in a machine learning environment assists in lowering worries about co linearity and enhancing the models' overall efficacy.

Feature engineering is an essential part of data purification in this study. The goal of this method is to find any additional connections among areas, making it easier to create brand-new features that could improve the accuracy of predictions. By putting these processes into practice, the data's quality is improved in general, providing a solid base for the creation of reliable and strong machine-learning models.

3.4 Data Transformation

A key component of data purification is data transformation, which guarantees that the data set is properly formatted and ready for analysis and modelling ("Data Transformation in Data Mining - Java point," 2022).Timestamps were transformed into the proper date-time format being the first step in the data transformation process. Timestamps are a helpful resource of time data for hourly power usage, generation, pricing, and weather information. The Pandas package and the Python programming language offer effective tools for handling date-time operations. It is possible to improve the efficiency of the time-series analysis used to find temporal trends and patterns in the data through the transformation of timestamps into the proper date-time objects. The control of outliers in the dataset was another essential component. Outliers are characterized as points of information that vary significantly from the rest of the data, potentially introducing bias into modelling and analytical processes. Several techniques, including the use of Z-score analysis and the installation of IQR (interquartile range) filtering, were used to identify and reduce misfits. So as to assure the accuracy of the data and avoid any errors in the results, it is possible to either discard or transform information points that have been recognized as misfits utilizing these approaches.

The weather dataset provided other weather-related information for various locations. The meteorological features were divided into various categories and categorized according to locations to allow essential comparisons and studies. This change made it possible to investigate how the environment affects the particular patterns of power consumption and generation found in each downtown location.

Also, additional time-related factors were calculated throughout the data transformation procedure in order to analyze the production of electricity. In order properly record the variations in electricity generation trends throughout multiple time intervals, work hours, weekdays, and months were chosen based on the use of timestamps. While weekdays and months showed seasonal and weekly swings in electricity demand, business hours were also crucial for analyzing trends in electricity usage during regular working hours. For it to include categorical variables in machine learning algorithms, it is often necessary to translate them into numerical values. Methodologies like one-hot encoding and label encoding were often used to transform categorical data into a format suitable for analysis. Label encoding is a technique that gives numerical labels to various categories. The use of these encoding techniques made it easier to include data with categories in machine learning methods.

3.5 Feature Selection

The choice of the objective variable is of the greatest importance in the context of forecasting electricity costs because it directly affects statistical modelling and forecasting. Due to the "price actual" field's significance and importance to achieving the research goals, it was decided to use it as the focus variable in the dataset

The electricity costs that have been scientifically observed on the market are shown in the field titled "price actual." The study's choice of this objective variable is meant to make it easier to create machine learning models that are capable of making accurate predictions of the price of electricity under realistic conditions. Also, it is essential for a variety of market participants connected to the energy sector, including power firms, investors, and customers. People who participate in the market can make educated choices, improve their investment strategies, and successfully minimize risks due to their capacity to create precise price forecasts.

Consequently, the price of electricity is an important statistic to assess market dynamics, which fluctuate by a variety of interactions between supply and demand factors, the composition of power-producing sources, weather patterns, and market regulatory frameworks. It will be possible to provide significant information on market trends and potential price movements by the process of conveying energy prices, which will help with the creation of long-term plans and investment decisions.

3.6 Software Tools

Several software programs have been used in this research thesis to assist with data analysis, machine learning modelling, and visualization. These software programs received consideration for their demonstrated ability to handle enormous data sets, substantial machine-learning libraries, and simplicity to use for study.

3.6.1 Programming Platform & Environment

Python, a popular programming language, was used as the main coding language for this study. Python's broad ecosystem, which contains a variety of already constructed software packages, modules, and frameworks, can be used to justify why it was chosen. The research made significant use of Python and also the Jupyter Notebook, a highly interactive computing environment that could be readily integrated into Visual Studio Code (VSCode). With its seamless scripting, debugging, and version tracking capabilities, Visual Studio Code (VSCode) provides the best environment for using Python and Jupiter Notebook.

3.6.2 Data Analysis Tools

In this study, management and analysis of the data required the use of dependable technologies like NumPy and Pandas. It has been impossible to operate multi-dimensional arrays and perform computational tasks without NumPy. On the other hand, Pandas, which offers a wide variety of data structures that may be customized, has developed into a useful tool for efficiently organizing structured data. The inclusion of these libraries significantly aided in the efficient alteration and evaluation of the dataset utilized in the subsequent investigations.

3.6.3 Data Visualisation

Data visualization was a key component of this study, and libraries like Seaborn, Matplotlib, and Plotly Express were used to make it easier. The highly flexible tracking program Matplotlib offers an array of complex customization tools for creating visualizations. Still, Seaborn introduced a more sophisticated tool for creating informative and attractive statistical visuals. The dynamic and engaging visualizations generated by Plotly Express, known for its simplicity of use, improved comprehension of the research findings.

3.6.4 Machine Learning Frameworks

In this research, prediction models were developed and tested using well-known machine learning and deep learning frameworks such as Scikit-learn, TensorFlow, and Keras. The Scikit-learn library, which is known for its durability in the machine-learning community, has been important in offering a variety of approaches for tasks including classification, regression, clustering, and other related domains. The database was the greatest option for bringing machine learning models to use in the field of the price of electricity prediction because of its accessibility and well-documented it was. Because of its ability in managing complex network layouts and big datasets, TensorFlow, a sophisticated deep-learning framework, was used concurrently (Goldsborough, 2016). In this research, prediction models were created and trained using well-known machine learning and deep learning frameworks such as Scikit-learn, TensorFlow, and Keras.

The Scikit-learn library, which is known for its reliability in the machine-learning community, has been essential in offering a variety of approaches for tasks including classification, regression, clustering, and other related domains. The database was the greatest option for bringing machine learning models to use in the area of electricity price prediction because of its simplicity and well-documented it was.

Because of its ability in managing complex network structures and big datasets, TensorFlow, a sophisticated deep-learning framework, was used continuously (Goldsborough, 2016).

3.7 Ethics

The ethical standards for academic honesty and data protection have been recognized in this study effort, and they are followed. The dataset used in this study, "Hourly energy demand, generation, and weather (Electrical demand, generation by type, prices, and weather in Spain)," was obtained from a public database, such as Kaggle. The dataset consists of a comprehensive set of a decade's worth of weather, pricing, and electrical use data for Spain.

The information provided in this investigation has been collected from a variety of accessible sources. The consumption and generating data were gathered from ENTSOE, a site for information supplied by Transmission Service Operators (TSOs) that is open to the general public. The Spanish Transmission System Operator (TSO), Red Electric Espaa, provided the settlement pricing. Also, the Open Weather API has made the weather information that was formerly limited to the bigger public for a personal project accessible to the public. The sources mentioned above are recognized and acknowledged in this research report for the publicly accessible information that they contributed.

There are no sensitive, private, or personal data in the dataset used in this research. The information used in this study relates to an array of energy-related topics, including supply, demand, pricing, and meteorological conditions. It's also important to remember that these kinds of information are not sensitive and cannot be used to identify an individual. The use of the data is strictly restricted to commercial exploitation and only permitted for academic research.

The evaluation of potential forecasts will be performed throughout the framework of this research by comparing them to the current cutting-edge forecasts that are being used within the industry. The key objective of this research is to enhance scientific comprehension regarding energy markets and procedures for price prediction. Also, accepts the inherent risk that exists in forecasting and makes no claims of absolute accuracy.

4. Results and Findings

4.1 Chapter Overview

The analysis and prediction made on the selected dataset indicated in the methodology sections are discussed in this chapter. It provides visualizations of data, exploratory data analysis, and results from a variety of trained machine learning models, including LSTM, XGBOOST, and LightGBM.

4.2 Exploratory Data Analysis

EDA is a kind of statistical approach that is mostly used to look for various unidentified characteristic patterns within the given data collection. Several statistical methods, including autocorrelation, QQ plots; box charts, histograms, and others are used to achieve these results. (Javed et al. 2021) It is significant to mention that comprehending the cost of electricity data plays a crucial role in the study before using ML algorithms for prediction. It enables us to detect hidden patterns, handle excess data, and determine the basic information structures.

The EDA started by fusing the two datasets into a single dataset to make it more accessible for subsequent analysis. It is well recognized that cleaning the data, which entails locating and addressing outliers and missing data within the dataset, is crucial before putting a model into practice. To give the initial comprehension of the acquired data, a review of the measures of change, central tendency indicators like median, average, standard deviation(SD), and ranges where the data lies is also done.

The patterns of variation and seasonality fluctuations that can be seen in the hourly electricity costs were then examined using histograms, box plots, and line graphs as part of a time-series study. The facts, including how electricity prices are distributed and the connections between pricing and other factors, are better understood thanks to these diagrams.

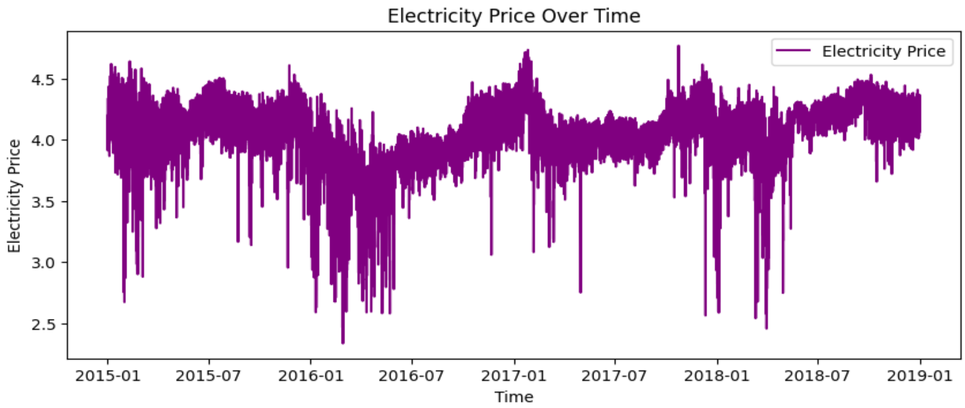


Fig 4.1 Electricity price actual over the years from 2015 to 2019.

Fig 4.1 shows how the "actual price" factor has changed over time within the time period of 2015 to 2019. It is clear that the price trend indicates non-stationery and variation over the specified time frame. Additionally, it shows that costs vary depending on particular time periods. The presence of a regular trends or periodic characteristics in pricing patterns is also notable because it suggests a possible relationship with other regular occurrences like changes in the weather, variations in customer behaviour, or other significant variables. We needed to properly identify the specific causes of the observed price variants, though, in order to gain a deeper understanding of the patterns.

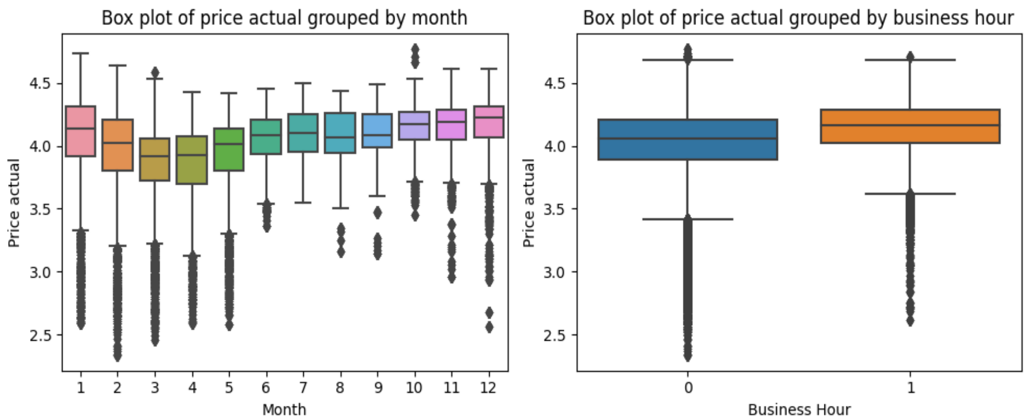


Fig 4.2 Electricity price variation in each of the months and business hours

In Fig. 4.2, a more thorough examination to the "actual price" attribute's analysis provides a more thorough understanding of the observed variances in electricity pricing. Pricing can be categorized by month, and trends between company and non-business hours can be discovered. The box plots, which depicts the price of energy by month (for instance, January is indicated by 1, and in February is indicated by 2), shows that the cost of power typically peaks in the winter. This is likely due to higher demand for heating and shorter days of sunlight.

The lowest prices, on the other hand, appear to be in the latter part of spring and early summers (such as June and May, respectively, indicated by 5 and 6), which could be attributed to warmer temperatures and less energy being required for warming or cooling. Warmer weather conditions over those periods, which resulted in less energy being used for heating or cooling, may help identify this tendency. This increase is most likely the result of increased commercial and manufacturing activity, which might increase energy demand. Figures 4.1 and 4.2, when compared, demonstrate a variety of elements of how prices for energy fluctuate over time. They demonstrate not just long-term trends but also intricate annual patterns and the impact of daily business cycles on prices. They can be used to examine the complexity of energy pricing and the significance of having a thorough grasp of how many temporal elements interact and effect price behaviour.

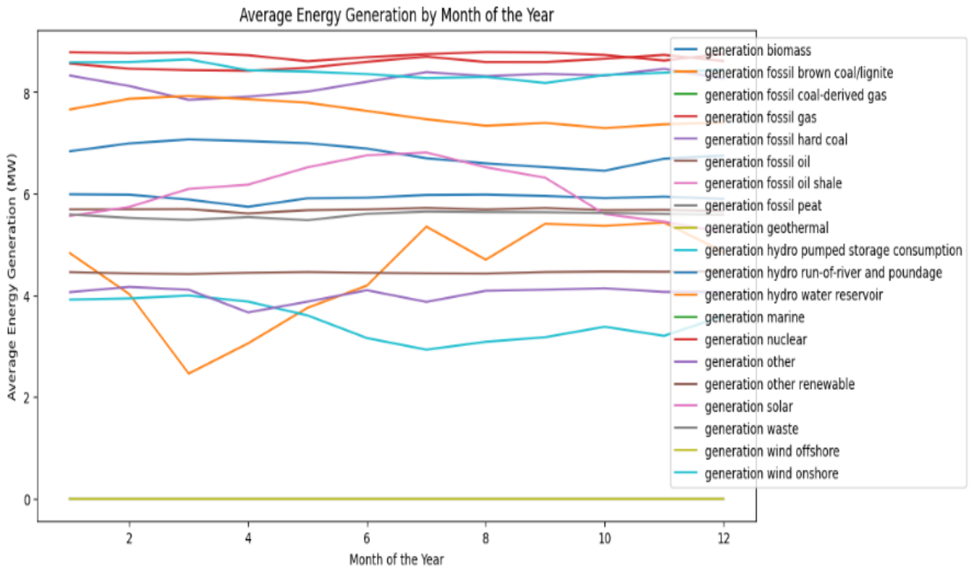


Fig 4.3 Average Energy Generation by month of the year

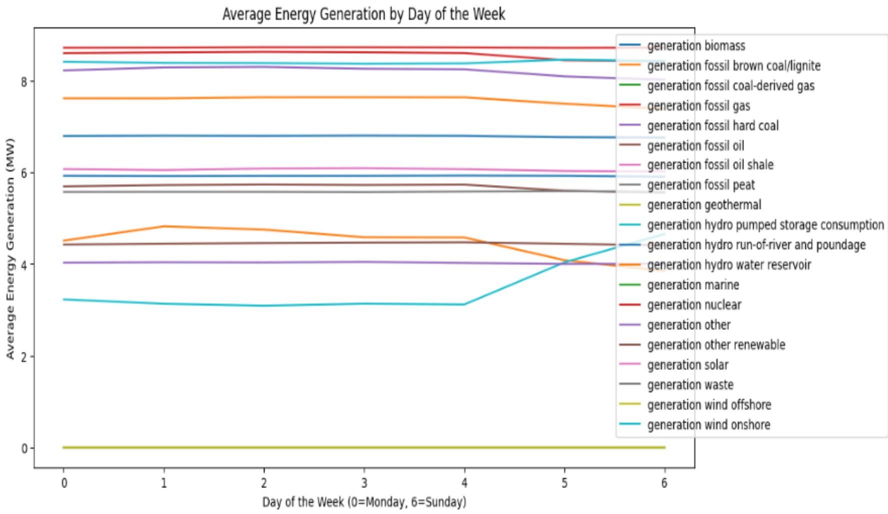


Fig 4.4 Average Energy Generation by day of the week

The thorough analysis performed using Figure 4.3, 4.4, and 4.5 improves understanding of the changing dynamics of the production of energy. The comprehensive examination shown in Figure 4.3 through 4.5 provides a thorough understanding of the many aspects related to energy production and cost over various time periods.

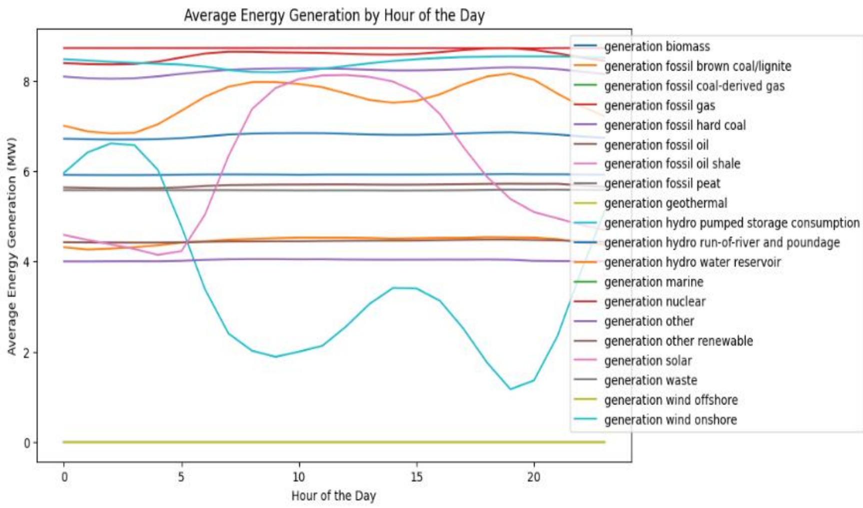


Fig 4.5 Average Energy Generation by hour of the day

The strong link between different sources of production and temporal aspects may be clearly seen in these graphs. Renewable resources, like the sun and wind, regularly exhibit notable changes that are consistent with their availability in nature. Solar production seems to be most active during the day (Fig. 4.3), and seasonal variations have a significant impact on it (Fig. 4.5), showing how reliant they are on the weather. In contrast, there are shifts in the production of wind energy based on seasonal patterns as well as daily wind conditions.

The effects of the price of electricity production are varied and complex. A multitude of factors, including demand cycles, climate patterns, and market trends, may come together to produce electricity output levels that change over time. As depicted on Figure 4.2, there's a direct link between higher electricity spending and more intensive financial transactions during business hours. Peaks seen in particular months, as shown on Figure 4.5, may coincide with the rise of demand for warming or cooling.

These results show the critical role that non-renewable resources, particularly fossil fuels, have played in maintaining sustainable supply of energy across time. In the final analysis, these ongoing inputs demonstrate how dependent we still are on traditional energy sources and how impossible the entire switch to clean power will be. Most importantly, detailed temporal analysis fosters a better knowledge of the major variables influencing the behaviour of energy prices. Additionally, these data's patterns and linkages offer crucial insights that will be applied to price predictions in the future. When the seasonal structures, daily structures, and weekly oscillations of electricity production are taken into account, these simulations become accurate and context aware.

Finally, the knowledge obtained from Figure 4.1 through 4.5 as a whole paints an accurate depiction of the supply of electricity. Thus, by connecting the dots between the past, nature, technological improvements, and human behaviour, this study illustrates how energy is produced as well as how much it costs. Understanding the complex connections not only makes it simpler to comprehend how things function right now, but it also aids in future energy management, pricing estimates, and sustainable efforts.

4.3 Description of the Methods Used

In order to forecast hourly electricity prices, this paper investigates a number of machine learning models that demonstrate important mathematical and computational concepts. Long-term short-term memory (LSTM), regression with linearity, random forest regression, XGBoost, and LightGBM (Light gradient boost Machine) are included in the group of models.

The first model used in this research is the linear regression model, which is the basis of statistical modelling and is straightforward but effective. It operates under the fundamental tenet that there exists a linear relationship between both dependent and independent variables. For the sake of this study, it is presumed that the statistic of interest, "price actual," can be defined as a linear expression of a wide variety of features that are viewed as independent factors within the dataset. Although it makes interpretation and calculation more efficient to assume that electricity price patterns are linear, it is crucial to recognize that the assumption could simplify the complexity of the underlying reality. As a result, even while linear regression is a useful benchmark, it might not accurately reflect the complexity of the dataset.

The second model uses random forest regression, a type of ensemble learning that builds many decision trees and combines their predictions, so making the model more complex. Random forest regression, as opposed to linear regression, may identify non-linear correlation and interactions between features. It is crucial to keep in mind that models based on random forests might result in a sizable computing demand, particularly when working with a lot of trees or dataset. Additionally, although having more computational power than the linear regression models, these models are more difficult to interpret, making it challenging to pinpoint the precise contribution of certain features.

By storing and utilising data from earlier inputs, the Long Short-Term Memory, or LSTM, is a specialised version of recurrent neural networks, or RNNs, that has been specifically created to handle time series data effectively (Moradzadeh *et al.,* 2020). When attempting to estimate energy costs, it is particularly crucial to be able to comprehend and maintain track of long-term relationships in sequential data due to the significance of historical price and time trends. It is crucial to keep in mind that long-short-term memory (LSTM) models have significant processing needs and demand careful hyperparameter adjustment.

Extreme Gradient Boosting, or XGBoost, is an advanced technique that creates a collection of weak models for prediction, typically decision trees, to produce a robust prediction model. The suggested methodology involves incorporating various strategies into the already-existing ensemble in a sequential manner with the intention of correcting any potential faults that may have occurred. Due to its extraordinary qualities, such as its ability to size, its adaptability, and its outstanding performance in several machine learning applications, XGBoost has attracted a lot of attention and is widely used. However, when formalization is not correctly implemented and requires careful tweaking of hyperparameters, it is probable to have overfitting, similar to other ensemble techniques. Additionally, there may be issues with interpretability in XGBoost models, particularly as the ensemble's complexity increases.

In our electrical power price forecasting investigation, LightGBM was selected as the model of choice due to its performance with a sizable amount of data from time series. The algorithm's inherent efficiency—which resulted from histogram-based methodologies—made it possible to learn it quickly even on huge datasets. This made the algorithm especially well-suited for the volume and time precision of data connected to power prices. The model's speed, accuracy, interpretability, robustness against overfitting, and ability to directly handle categorical data made it the perfect choice for price of energy time series prediction.

This demonstrates the differences between complexity and simplicity, interpretability and predictive capacity, as well as computing economy and accuracy, and shows the fact that each model utilized in this study has a distinct set of benefits and drawbacks. The specific requirements and characteristics of the work are taken into consideration when selecting a model, highlighting the significance of having an in-depth comprehension of the models themselves and the data.

4.4 Models Evaluations

This study aims to investigate how well models based on machine learning perform in terms of producing more precise predictions. In order to guarantee that our findings were fair and accurate, a number of experiments were conducted, and the data was analysed to establish the general efficacy of each model. The success or failure of these types of models will depend on how well they perform. It is important to highlight that each model is tested for its ability to anticipate the price of electricity in the future using prior market data as training data.

4.4.1 Model for Linear Regression

The initial quantitative technique in this investigation was linear regression. The two main measures used to evaluate the accuracy of model projections of energy prices were R squared (R2) and the resulting root averaged squared error (RMSE). The model's projection is typically 0.1460 values off from the real cost, according to the calculation of its RMSE, which was determined to be 0.1460. The size of the discrepancy shows that the model may be superior in some aspects even though the level of accuracy is clear.

The R2 value for the model of linear regression was found to be 0.5397. This suggests that 53.97% of the volatility in electricity prices can be accounted for by this model. These results suggest the need for a stronger model to be predicted in order to better understand the intricate patterns in electricity price fluctuations over time, even though they also show that the model's predictive power is reasonable and that a significant amount of the volatility in prices is still unknown.

4.4.2 Random Forest Regressor

The second model in this study was the random forest regressor. Three important metrics, including mean absolute errors (MAE), the root mean square error (RMSE), and the R-squared (R2), were used to assess it. It was calculated that there was an average absolute difference of 0.1552 between the real electricity prices and what the model predicted. The number indicates that the model is capable of making predictions, but it also identifies areas for growth. Better model performance is indicated by lower MAE values.

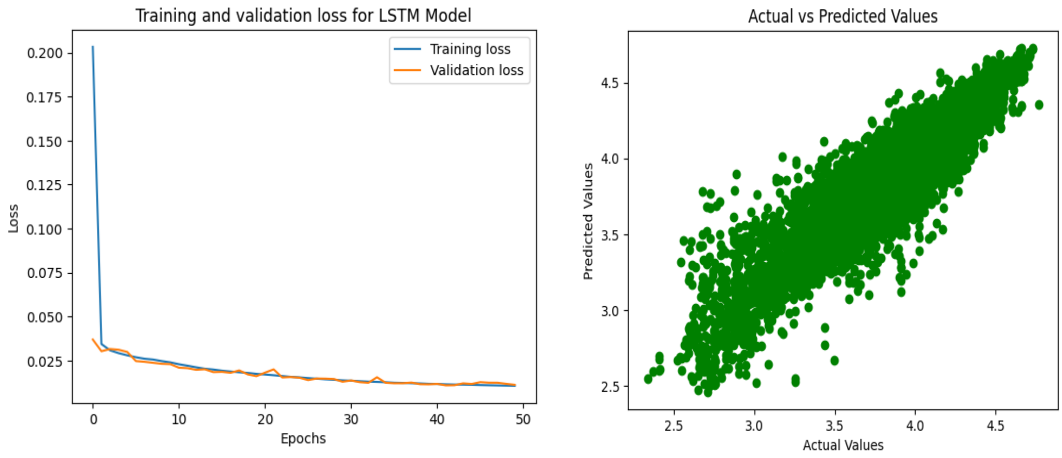
The Random Forest's Regressor's RMSE, which assesses the typical discrepancy between expected and actual values, was discovered to be 0.2002. This indicates that the model of linear regression is more accurate than the random forest model. Additionally, the model only adequately explained 13.52% of the variations in power price, as indicated by the significantly lower R2 value of 0.1352. These findings show that, despite the model's ability to capture non-linear correlations and interactions between features, it had not been very successful in predicting the price of energy. It also demonstrates how crucial it is to enhance models or investigate more complex models in order to increase their capacity for guesswork.

4.4.3 Long Short-Term Memory(LSTM)

The LSTM (Long Short-Term Memory) model, which has the intrinsic ability to manage time-series data well while keeping previous data, is highly suited for a particular task. With a batch limit of 32, the LSTM (Long Short-Term Memory) model finished training over the course of 50 epochs. The performance measurements demonstrate a significant improvement over the preceding models, such as a linear regression and the random forest regressor.

The mean Absolute Errors (MAE) for the LSTM model was 0.0745, indicating that the model's forecasts were on average 0.0745 units off from the actual pricing for energy. Comparing the current models to their earlier iterations reveals a clear reduction in the mean absolute errors (MAE). This decline demonstrates a notable improvement in forecast accuracy. As the model's R2-squared (R2) score is in a range of 0.8565, there is strong evidence to support its excellent forecasting abilities.

According to the score, the LSTM model has the capacity to account for about 85.65% of the changes in power price, which is a significant improvement over the earlier models. The test loss was measured and yielded a numerical rating of 0.0113, showing that the model is significantly flexible in dealing with data that was not previously present in the training set.



(a) **(b)**

Figure 4.6 (a) Training and Validation loss for LSTM Model

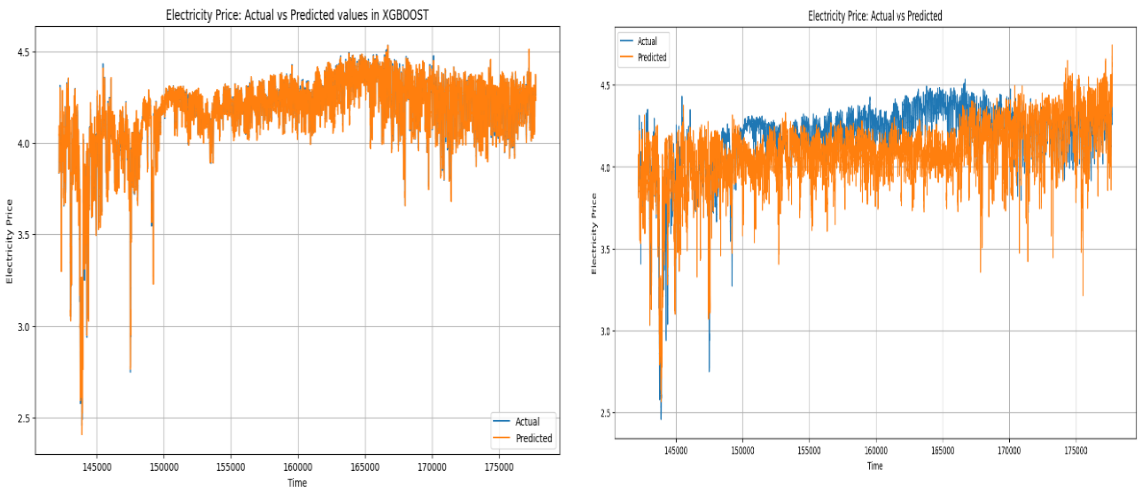
(b) Actual and predicted values correlation

More information was gained from the study of the figure 4.6 (a), it shows the evolution of the validation and training loss pattern over the epochs. It demonstrates that instruction as well as validation losses have been significantly reduced, ensuring a successful process of learning for future price of electricity forecasts. A significant indication that there isn't any over fitting is provided by the subsequent convergence and consistency of these losses. The correctness of the LSTM model was verified using the plot of actual values vs projected values in Figure 4.6 (b). The nodes' strict adherence to the diagonal line, which shows an important correlation between real and anticipated values, can be seen.

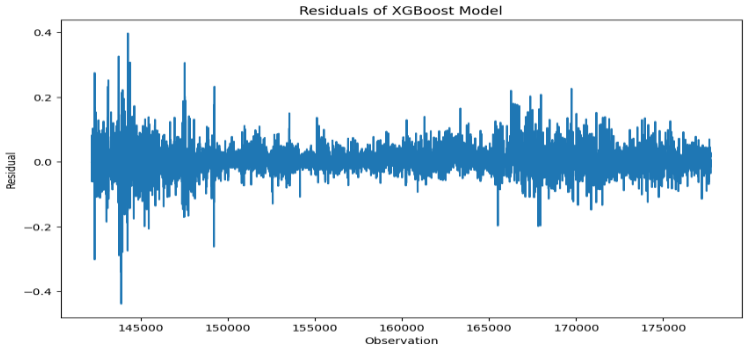
The design and hyperparameters of the model are suitable for this particular implementation. To verify the reliability and validity of these research findings, the study also stresses the need for additional validations utilizing a variety of metrics for assessment and test data.

4.4.4 Extreme Gradient Boosting (XGBoost)

The XGBoost framework was employed in this study to forecast the cost of energy because of its excellent performance and ability to handle both linear and nonlinear data patterns. The model scored well in the first stage, with an MAE and RMSE of 0.01235 and 0.02335, respectively. Furthermore, the model described roughly 98.82% of the variation in electricity costs, as seen by its R-squared score of 0.9882.



(a) (b)



(c)

Figure 4.7 Plots that represents the variations in actual and predicted values

The comparison between actual and anticipated values is shown in the plot above (Fig. 4.7(a)), and it reveals a larger range of data points and residuals, which hint to an increase in the size of prediction errors. Although there are other spots that are farther away from the zero in Figure 4.7(c), it is clear that most of the residuals are grouped there. It suggests that there may be room for more growth. In other words, it also emphasizes the importance of careful hyper parameter selection and the challenging process of fine-tuning machine learning models.

The model's performance was attempted to be improved through hyperparameter adjustment, but this resulted in fewer accurate forecasts. Following the tuning procedure, it was observed that the RMSE score increased to 0.1783, while the MAE increased to 0.1419. Additionally, a drop to 0.3139 was shown by the R squared value. Figure 4.7 (b) illustrates that the model performs poorly due to the greater variability between real and predicted values. The error is significantly higher after the tuning process compared with the error we acquired from the initial model (0.01). It shows that the model's efficiency was not improved by adding the delayed features of "temp" and "wind_speed." In fact, it seems to have made the predictions less reliable.

It could happen for a number of motives including the possibility that certain variables are just contributing noise to the algorithms and are not actually helpful to forecast the cost of electricity. Another reason could be that, in contrast to the results that the XGBoost theory predicts, there is no additive or linear connection between these qualities and the cost of power. This indicates that, in compared to the original model, the model's forecast accuracy declined and its ability to account for variations in power costs reduced. As a result, even though the XGBoost algorithm performed particularly well at predicting electricity prices, the procedure of hyperparameter tuning did not significantly improve the accuracy of the model. This finding points to the challenge of continuously enhancing the accuracy of models and it's possible that a simpler model may occasionally exceed simpler models.

4.4.5 Light Gradient Boosting Machine (LightGBM)

The success of LightGBM in foretelling hourly electricity spending was examined in this study. Figure 4.1, which displays the historical pattern of the variable known as the "the cost actual" over a given time period, is the main source of the results

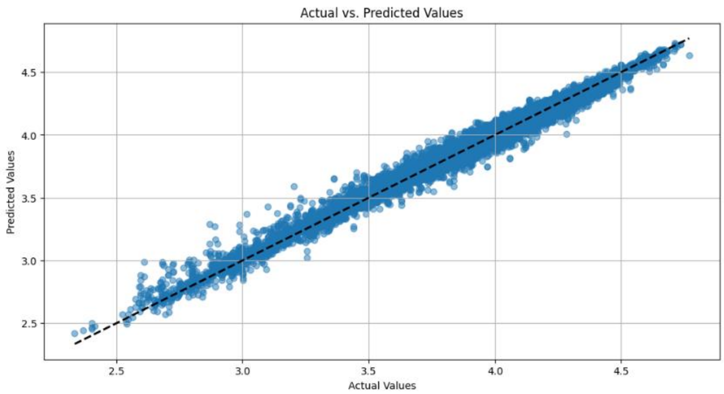


Figure 4.8 Correlation of actual vs predicted values

The figure 4.8 (Actual Vs Forecast LightGBM), which displays a scatter diagram of actual against anticipated values, was essential to this study. The great level of forecast accuracy shown by this graphical differentiation is very clear. A significant correlation existed between the predictions made by the simulation and actual observed values, as shown by the vast majority of data points having a propensity to cluster around the 45-degree line. Random occurrences did, however, explain periodic oscillations, highlighting the value of ongoing model improvement.

The accuracy of this model was assessed using its performance metrics. The observed value of the root average square error (RMSE), which measures how closely the model's predictions matched the actual data, was 0.0356. Additionally, the mean absolute errors (MAE), which provides a numerical indicator of the typical difference between forecasts and actual data, was calculated to be 0.0246. It was highly informative to compare this data to the distribution and instability of the "the cost actual" variable. The rate of determination, or R2, score was the most significant statistic that was observed. The impressive results show that the model performed exceedingly well in identifying within 98.39% of the Changes in cost of electricity.

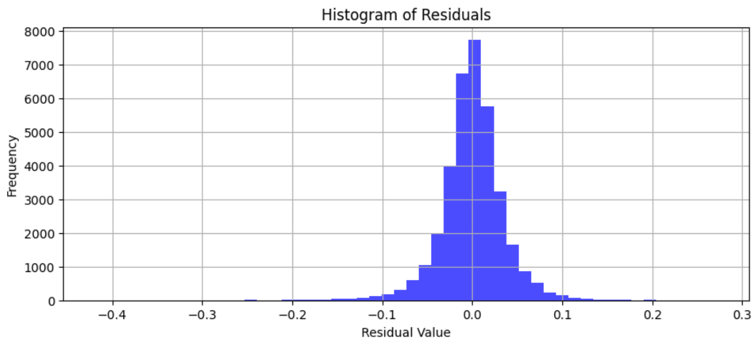


Figure 4.9 Residual histogram representation

Figure 4.9 above, which shows histograms that was represented using the model's distribution, offers a more thorough view of the residuals. The symmetrical residual distribution, which is characterized by a predominance of density near zero, contributes to the model's improved ability to make unbiased predictions. These findings provided further confirmation of the findings of our time-series data cross-validation analysis, which are shown in Figure 4.5. The 5-fold cross-validation results showed a decreasing trend in the root mean squared error (RMSE) values, indicating that the model performs better when more recent data patterns are used.

The excellent performance and dependability of the LightGBM as model must be emphasized. Given the wide range of data and measurements available, the enormous utility of price of energy forecasting is shown by its capacity to combine both precision and accuracy.

4.5 Comparison of models

In order to answer the research question,

“What is the best machine learning model for predicting future hourly electricity costs based on resource and weather features? “

There were evaluations of numerous machine learning models. Finding the model that can most accurately anticipate hourly electricity prices considering the variety and complexity of the data was the goal. Starting with the foundational Linear Regression, it was rapidly shown that its linear assumptions were flawed. Even though it was fundamental, it couldn't keep up with the intricacy of the data on power pricing, thus the results weren't satisfactory for us. Despite its ensemble capabilities, the Random Forest produced similar results. Its performance indicated a difficulty in understanding the intricate relationships and periodical complexity that serves as the pricing of power. LSTMs are particularly adept at spotting long-term dependencies in time series data. In our analysis, the LSTM model worked superbly, achieving a precision (R2 score) of roughly 86%. This demonstrates its value in time series prediction tasks, especially when past trends are significant.

The actual front-runners in our test, however, were gradient enhancing frameworks XGBoost and LightGBM. Due to their intricate algorithms, both of the models have rates of accuracy that are close to 98%. Their performance highlights not only their superior abilities but also how quickly they can adjust to difficult datasets like ours.The results show that XGBoost and LightGBM are equally effective in predicting hourly electricity costs in our sample. As shown by their nearly identical high accuracy, they are adept at capturing the intricate relationships between electricity, temperatures, and electricity pricing. Although LSTM displayed positive outcomes, its outstanding efficiency was largely overshadowed by the boosting models. However, for this specific forecasting problem, Linear Regression and Random Forest showed poorer results. The XGBoost and LightGBM as are the most effective predictive models for projecting future hourly electricity prices based on energy supply and meteorological features in this study, which is in response to our research topic. They have the best accuracy and ability to model complex relationships, making them the ideal options for certain predictive task.

5. Discussions

5.1 Chapter Overview

This chapter focuses on explaining the significance and challenges faced in forecasting electricity price forecasting.

5.2 Significance of data analysis in forecasting

The study mainly aimed to explore the potential of ML algorithms for predicting electricity prices. It was inspired from a research paper that highlighted the exciting possibilities of machine learning in energy markets. The dataset which is used in this study contains 4 years' worth of weather, pricing and energy usage data from Spain, served as the foundation for this analysis.

In this study, feature engineering played an important role which had turned the raw data into meaningful key insights for further analytical purposes. Through the creation and improvement of new features, the study was able to dig deeper into the factors that affect the energy prices, providing insights on historical trends, seasonal changes and all other key factors.

Additionally, the thorough analysis conducted in this study from basic algorithms to complex algorithms, demonstrated the exciting possibilities in data analytics domain. Gradient boosting algorithms like XGBoost and LightGBM, had highlighted the importance of using advanced statistical techniques for accurate forecasting because they are particularly good at managing complex data patterns.

Moreover, it was clear that they were able to efficiently handle complicated data correlations and patterns, resulting in trustworthy and precise forecasts. Additionally, the well performed models' hyperparameter adjustment, particularly for LightGBM, improved their predictive performance, guaranteeing the forecasts' robustness and dependability.

In a nutshell ,this study highlights how powerful is data analytics that can be used to predict energy prices. In addition to providing an outline for precise forecasting, it also emphasizes the necessity for ongoing innovation in analytical approaches to keep up with the dynamics of the electrical market, which are constantly changing.

5.3 Challenges faced during the study

In this study, we came across various constraints as finding a reliable dataset was quite challenging. Initially, this project aimed to build a predictive model for determining the day ahead electricity generation price for residential buildings. However, it was hard to find the required data collection due to data security issues. During this research, even though we were able to find many organisational sites that provides information regarding the energy generation about the different cities, they all provided only the statistical information of the total energy consumed over the years. Ideally, they are not useful for this research. Thus, due to the critical unavailability of data issue and time constraint, the study had diverted into a broader spectrum of this area. With the access of the current dataset spanning the Spanish electricity market, the study aimed from a household price prediction preceptive to a broader forecast of electricity prices. This insight helped to understand that finding the electricity price for a narrower area required many resources and time. Though, this research has identified new features among the available data and helped to analysis with multiple machine learning algorithms .

6. Conclusion and Future Recommendations

The prediction of electricity prices is a significant field of research due to its substantial impact on energy strategies, market dynamics, and the overall economy. This study investigated the area of using machine learning techniques to accurately forecast electricity costs. In order to apply this complex dataset that covers the several aspects of spanish electricity sector over the period of 4 years, this study undertook a thorough analytical process characterized by detailed data processing, the development of relevant variables, and the assessment of algorithms.

Moreover, the dataset has been modified better for predictive analysis by adding graphical representations and feature engineering. In comparison to other models used in this study, XGBoost and LightGBM had shown the highest level of forecasting capability. These models were also tested to evaluate the effectiveness of the hyperparameter tuning to improve their initial results. Ultimately, this study tries to show the basic value of machine learning in predicting power prices by giving a foundation for future research in the field.

6.1 Future Recommendations

Forecasting future electricity prices offers interesting opportunities, especially when focusing on specific segments such as residential properties. Targeting specific locations would provide residents with a nuanced insight into expected price fluctuations, allowing for more efficient use of electricity. Such accurate forecasting can lead to significant cost savings, and when combined with the power of real-time data streams such as IoT devices, dynamic forecasting can adapt to instant changes in resources in use or within the limits of the network

Further research can be carried out using geospatial analysis to understand the impact of geographic climate change on electricity prices. This will be particularly valuable in areas with varying climates and topography. As the field evolves, emphasis will likely shift to more consumer-centric models, changing the way individuals manage their energy consumption

7. Reference

Al Metrik, M., A. Musleh, D. (2022) "Machine Learning Empowered Electricity Consumption Prediction," *Computers, Materials & Continua*, 72(1), 1427–1444, available: http://dx.doi.org/10.32604/cmc.2022.025722.

Alkawaz, A. N., Abdellatif, A., Kanesan, J., Khairuddin, A. S. M., & Gheni, H. M. (2022). Day-Ahead Electricity Price Forecasting Based on Hybrid Regression Model. *IEEE Access*, *10*, https://doi.org/10.1109/access.2022.3213081

Anbazhagan, S. and Kumarappan, N. (2013) Day-Ahead Deregulated Electricity Market Price Forecasting Using Recurrent Neural Network. *IEEE Systems Journal*, Available from http://dx.doi.org/10.1109/jsyst.2012.2225733.

Cerjan, M., Krželj, I., Vidak, M., Delimar, M. (2013) A Literature Review with Statistical Analysis of Electricity Price Forecasting Methods [online], available: https://ieeexplore.ieee.org/document/6625068

Cochell, J.E., Schwarz, P.M., Taylor, T.N. (2012) "Using real-time electricity data to estimate response to time-of-use and flat rates: an application to emissions," *Journal of Regulatory Economics*, 42(2), 135–158, available: http://dx.doi.org/10.1007/s11149-012-9190-7.

Contreras, J., Espinola, R., Nogales, F.J. and Conejo, A.J. (2002) ARIMA Models to Predict Next-Day Electricity Prices. *IEEE Power Engineering Review* ,Available from http://dx.doi.org/10.1109/mper.2002.4312577.

Chan, K.F., Gray, P. and van Campen, B. (2008) A new approach to characterizing and forecasting electricity price volatility. *International Journal of Forecasting* ,Available from http://dx.doi.org/10.1016/j.ijforecast.2008.08.002.

Data Cleaning Techniques in Data Mining and Machine Learning [online] (2023) *ProjectPro*, available: https://www.projectpro.io/article/data-cleaning-techniques/651.

Data Transformation in Data Mining - Javatpoint [online] (n.d.) *www.javatpoint.com*, available: https://www.javatpoint.com/data-transformation-in-data-mining.

Dragasevic, Z., Milovic, N., Djurisic, V., & Backovic, T. (2021, November). Analyzing the factors influencing the formation of the price of electricity in the deregulated markets of developing countries. *Energy Reports*, *7*, 937–949. https://doi.org/10.1016/j.egyr.2021.07.046

ENTSO-E Transparency Platform [online] (n.d.) available: https://transparency.entsoe.eu/.

Esteves, G. R. T., Bastos, B. Q., Cyrino, F. L., Calili, R. F., & Souza, R. C. (2015). Long Term Electricity Forecast: A Systematic Review. *Procedia Computer Science,* https://doi.org/10.1016/j.procs.2015.07.041

Fałdziński, M., Fiszeder, P., Orzeszko, W. (2020) "Forecasting Volatility of Energy Commodities: Comparison of GARCH Models with Support Vector Regression," *Energies*, 14(1), 6, available: http://dx.doi.org/10.3390/en14010006.

Feijoo, F., Silva, W. and Das, T.K. (2016) A computationally efficient electricity price forecasting model for real time energy markets. *Energy Conversion and Management* , Available from http://dx.doi.org/10.1016/j.enconman.2016.01.043.

Ferreira, N. P., Ramos, J. G., & Fernandes, P. O. (2019, August 23). A linear regression pattern for electricity price forecasting in the Iberian electricity market. *Revista Facultad De Ingeniería Universidad De Antioquia*, https://doi.org/10.17533/udea.redin.20190522

Gabrielli, P., Wüthrich, M., Blume, S. and Sansavini, G. (2022) Data-driven modeling for long-term electricity price forecasting. Energy, Available from http://dx.doi.org/10.1016/j.energy.2022.123107.

García-Martos, C., Rodríguez, J., & Sánchez, M. J. (2013, January). Modelling and forecasting fossil fuels, CO2 and electricity prices and their volatilities. *Applied Energy*, *101*, 363–375. https://doi.org/10.1016/j.apenergy.2012.03.046

Gholipour Khajeh, M., Maleki, A., Rosen, M.A. and Ahmadi, M.H. (2017) Electricity price forecasting using neural networks with an improved iterative training algorithm. *International Journal of Ambient Energy* [Post-print], 39 (2), pp. 147–158. Available from http://dx.doi.org/10.1080/01430750.2016.1269674.

Goretti, G. (2020). Forecasting the Short-term Value of Wind Power for Risk-aware Bidding Strategies in Single-imbalance Price Electricity Markets. Technological University Dublin. DOI: 10.21427/VH4H-WR88

Goldsborough, P. (2016). A Tour of TensorFlow. *arXiv:1610.01178 [cs]* [online], 1 October 2016. Available from: http://arxiv.org/abs/1610.01178 [accessed 1 June 2018].

Higgs, H., Lien, G. and Worthington, A.C. (2015) Australian evidence on the role of interregional flows, production capacity, and generation mix in wholesale electricity prices and price volatility. *Economic Analysis and Policy* , Available from http://dx.doi.org/10.1016/j.eap.2015.11.008.

Hryshchuk, A. and Lessmann, S. (2018) Deregulated Day-Ahead Electricity Markets in Southeast Europe: Price Forecasting and Comparative Structural Analysis. *SSRN Electronic Journal* ,Available from http://dx.doi.org/10.2139/ssrn.3122308.

Javed, U., Ijaz, K., Jawad, M., Ansari, E.A., Shabbir, N., Kütt, L., Husev, O. (2021) “Exploratory Data Analysis Based Short-Term Electrical Load Forecasting: A Comprehensive Analysis,” *Energies*, 14(17), 5510, available: http://dx.doi.org/10.3390/en14175510.

Jedrzejewski, A., Lago, J., Marcjasz, G. and Weron, R. (2022) Electricity Price Forecasting: The Dawn of Machine Learning. *IEEE Power and Energy Magazine* , Available from http://dx.doi.org/10.1109/mpe.2022.3150809.

Lago, J., Marcjasz, G., De Schutter, B., Weron, R. (2021) “Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark,” *Applied Energy*, 293, 116983, available: http://dx.doi.org/10.1016/j.apenergy.2021.116983.

Lehna, M., Scheller, F. and Herwartz, H. (2022) Forecasting day-ahead electricity prices: A comparison of time series and neural network models taking external regressors into account. *Energy Economics* ,Available from http://dx.doi.org/10.1016/j.eneco.2021.105742.

Li, Y. and Flynn, P.C. (2004) Deregulated power prices: comparison of volatility. *Energy Policy*,. Available from http://dx.doi.org/10.1016/s0301-4215(03)00130-7.

Liu, L., Bai, F., Su, C., Ma, C., Yan, R., Li, H., Sun, Q., & Wennersten, R. (2022, May). Forecasting the occurrence of extreme electricity prices using a multivariate

Lokesh (2023) "Data Cleaning Techniques in Data Mining and Machine Learning," *ProjectPro*.

Pallonetto, F., Jin, C. and Mangina, E. (2022) Forecast electricity demand in commercial building with machine learning models to enable demand response programs. *Energy and AI* , Available from http://dx.doi.org/10.1016/j.egyai.2021.100121.

Pandey, Nivedita and K. Upadhyay. "Different price forecasting techniques and their application in deregulated electricity market: A comprehensive study." 2016 International Conference on Emerging Trends in Electrical Electronics & Sustainable Energy Systems (ICETEESES) (2016): 1-4.

Pinhão, M., Fonseca, M., & Covas, R. (2022, June 11). Electricity Spot Price Forecast by Modelling Supply and Demand Curve. *Mathematics*, *10*(12), 2012. https://doi.org/10.3390/math10122012

Maciejowska, K., Nowotarski, J., & Weron, R. (2016, July). Probabilistic forecasting of electricity spot prices using Factor Quantile Regression Averaging. *International Journal of Forecasting*, *32*(3), 957–965. https://doi.org/10.1016/j.ijforecast.2014.12.004

Mohsenian-Rad, A.-H. and Leon-Garcia, A. (2010) Optimal Residential Load Control With Price Prediction in Real-Time Electricity Pricing Environments. *IEEE*

Transactions on Smart Grid , Available from http://dx.doi.org/10.1109/tsg.2010.2055903.

Moradzadeh, A., Zakeri, S., Shoaran, M., Mohammadi-Ivatloo, B., Mohammadi, F. (2020) "Short-Term Load Forecasting of Microgrid via Hybrid Support Vector Regression and Long Short-Term Memory Algorithms," *Sustainability*, 12(17), 7076, available: http://dx.doi.org/10.3390/su12177076.

Quillbot (no date). Available at: https://quillbot.com/.

Rodriguez, C.P., Anders, G.J. (2004) "Energy Price Forecasting in the Ontario Competitive Power System Market," *IEEE Transactions on Power Systems*, 19(1), 366–374, available: http://dx.doi.org/10.1109/tpwrs.2003.821470.

Sajjad, M., Khan, Z. A., Ullah, A., Hussain, T., Ullah, W., Lee, M. Y., & Baik, S. W. (2020). A Novel CNN-GRU-Based Hybrid Approach for Short-Term Residential Load Forecasting. *IEEE Access*, https://doi.org/10.1109/access.2020.3009537

Sridharan, V., Tuo, M. and Li, X. (2022) Wholesale Electricity Price Forecasting Using Integrated Long-Term Recurrent Convolutional Network Model. *Energies* ,Available from http://dx.doi.org/10.3390/en15207606.

Stefenon, S.F., Seman, L.O., Mariani, V.C. and Coelho, L. dos S. (2023) Aggregating Prophet and Seasonal Trend Decomposition for Time Series Forecasting of Italian Electricity Spot Prices, Available from http://dx.doi.org/10.3390/en16031371.

Tschora, L., Pierre, E., Plantevit, M., & Robardet, C. (2022, May). Electricity price forecasting on the day-ahead market using machine learning. *Applied Energy*, https://doi.org/10.1016/j.apenergy.2022.118752

Wang, Z., Li, Y., Shen, Y., Zhou, L., Wang, C. (2017) "Virtual electricity retailer for residents under single electricity pricing environment," *Journal of Modern Power Systems and Clean Energy*, 5(2), 248–261, available: http://dx.doi.org/10.1007/s40565-017-0269-0.

Weron, R., et al., 2014. Electricity price forecasting: A look at where we are now and where we might be going. International Journal of Forecasting, Volume 30, Issue 4.

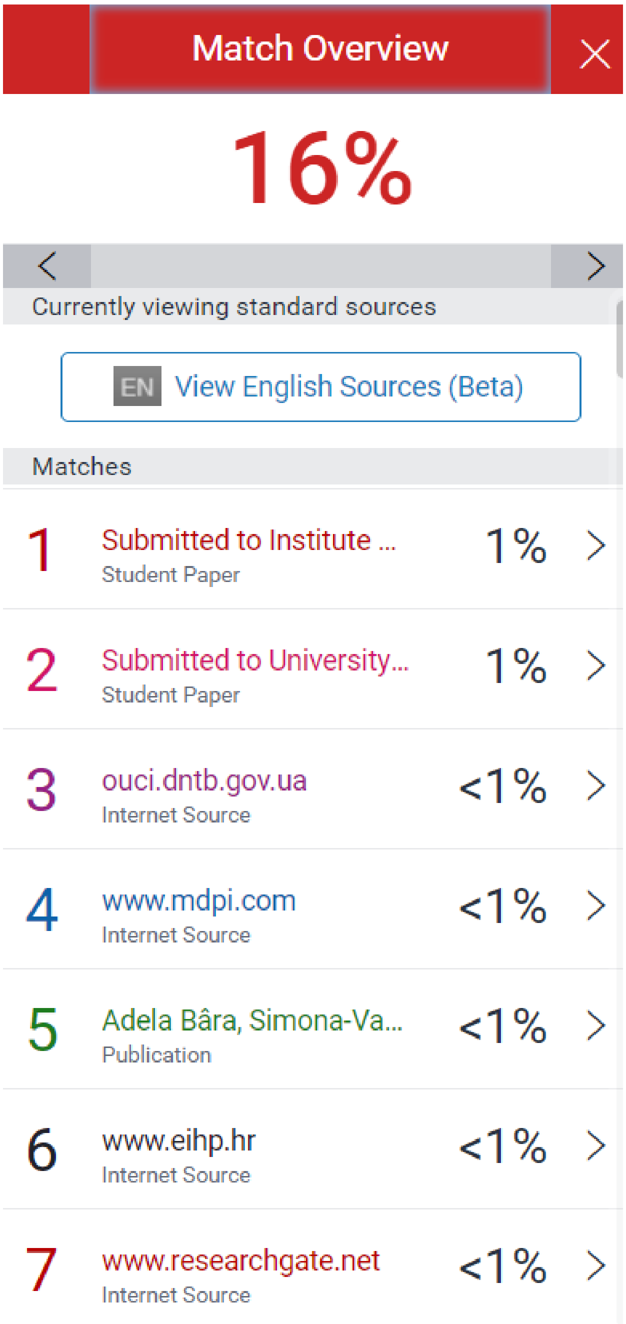
Yamashita, D., Isa, A. M., Yokoyama, R., & Niimura, T. (2008). Forecasting of Electricity Price and Demand Using Autoregressive Neural Networks. *IFAC Proceedings Volumes*, https://doi.org/10.3182/20080706-5-kr-1001.02527

Ziela, F., Steinert, R. (2018) "Probabilistic mid- and long-term electricity price forecasting ," *Renewable and Sustainable Energy Reviews*, 28(1364–0321), available: https://www.sciencedirect.com/science/article/pii/S1364032118303885.

Zhang, H., Chen, B., Li, Y., Geng, J., Li, C., Zhao, W. and Yan, H. (2022) Research on medium- and long-term electricity demand forecasting under climate change. *Energy Reports* , Available from http://dx.doi.org/10.1016/j.egyr.2022.02.210.

APPENDIX

Similarity Report and Digital Receipt

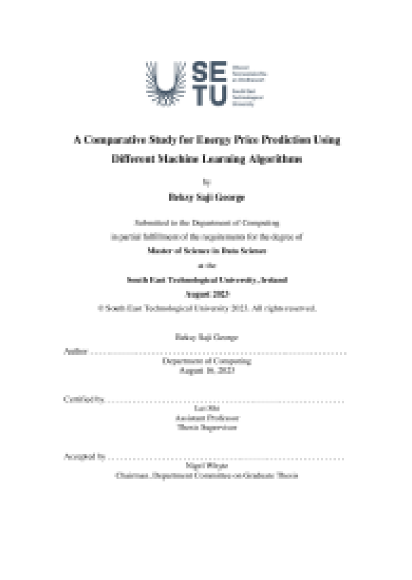


Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Beksy Saji George
Assignment title:	Turnitin FINAL Dissertation Submission
Submission title:	C00290800_Beksy_Saji_George_Dissertation.pdf
File name:	C00290800_Beksy_Saji_George_Dissertation.pdf
File size:	1.1M
Page count:	55
Word count:	14,478
Character count:	86,723
Submission date:	14-Aug-2023 10:23AM (UTC+0100)
Submission ID:	2145651877


A Comparative Study for Energy Price Prediction Using Different Machine Learning Algorithms

by

Beksy Saji George

Submitted to the Department of Computing

in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

South East Technological University, Ireland

August 2023

© South East Technological University 2023. All rights reserved.