UNIVERSITY OF LJUBLJANA

SCHOOL OF ECONOMICS AND BUSINESS

MASTER'S THESIS

A COMPARISON OF ADVANCED PREDICTIVE MODELS FOR THE FORECASTING OF ELECTRICITY PRICES

Ljubljana, October 2021

HRISTINA TRAJKOVA

AUTHORSHIP STATEMENT

The undersigned Trajkova Hristina, a student at the University of Ljubljana, School of Economics and Business, (hereafter: SEB LU), author of this written final work of studies with the title A comparison of advanced predictive models for the forecasting of electricity prices, prepared under supervision of prof. dr. Igor Lončarski and co-supervision of /

DECLARE

1. this written final work of studies to be based on the results of my own research;
2. the printed form of this written final work of studies to be identical to its electronic form;
3. the text of this written final work of studies to be language-edited and technically in adherence with the SEB LU's Technical Guidelines for Written Works, which means that I cited and / or quoted works and opinions of other authors in this written final work of studies in accordance with the SEB LU's Technical Guidelines for Written Works;
4. to be aware of the fact that plagiarism (in written or graphical form) is a criminal offence and can be prosecuted in accordance with the Criminal Code of the Republic of Slovenia;
5. to be aware of the consequences a proven plagiarism charge based on the this written final work could have for my status at the SEB LU in accordance with the relevant SEB LU Rules;
6. to have obtained all the necessary permits to use the data and works of other authors which are (in written or graphical form) referred to in this written final work of studies and to have clearly marked them;
7. to have acted in accordance with ethical principles during the preparation of this written final work of studies and to have, where necessary, obtained permission of the Ethics Committee;
8. my consent to use the electronic form of this written final work of studies for the detection of content similarity with other written works, using similarity detection software that is connected with the SEB LU Study Information System;
9. to transfer to the University of Ljubljana free of charge, non-exclusively, geographically and time-wise unlimited the right of saving this written final work of studies in the electronic form, the right of its reproduction, as well as the right of making this written final work of studies available to the public on the World Wide Web via the Repository of the University of Ljubljana;
10. my consent to publication of my personal data that are included in this written final work of studies and in this declaration, when this written final work of studies is published.

Ljubljana, October 5th, 2021
(Month in words / Day / Year,
e. g. June 1st, 2012

Author's signature: H. Trajkova

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AR – Autoregressive

ARIMA – Autoregressive integrated moving average

ARMAX – Autoregressive–moving-average with exogenous inputs

ARX – Autoregressive models with exogenous inputs

EEX – European Energy Exchange

ENTSO-E – European network of transmission system operators

EPEX Spot – European Power Exchange

FPC – Forward price curve

HPFC – Hourly price forward curve

LR – Linear regression

MAE – Mean absolute error

MCP – Market clearing price

ME – Mean error

ML – Machine learning

NEMO – National electricity market operator

OLS – Ordinary least squares

OTC – Over the counter

PTR – Power transmission right

RMSE – Root mean squared error

RSE – Residual standard error

RSS – Residual sum of squares

SPV – Solar photo-voltaic

TSO – Transmission system operator

XG boost – Extreme gradient boosting

INTRODUCTION

The deregulation and liberalization of the energy sector that took place in the late 1990s worldwide brought various changes to the way the whole system was organized. Therefore, from a vertically integrated system controlled by the government, we switched to open market systems where the number of energy providers increased and allowed the end customers to choose their energy provider. The incentive for liberalization of the energy sector was born after the liberalization of other sectors which brought increased wealth and various economic benefits. The intention was to improve energy efficiency and consciousness by increasing competition between companies. Another incentive for the liberalization of the energy sector in Europe was expanding the connection between the European energy markets and combining one common market which will provide the producers and distributors with the possibility of participation not only in national but also in international markets (The Economist, 2006).

Speaking of the electricity market specifically, the competitiveness would provide lower costs for electricity production and distribution which translates to lower electricity prices for end customers. Furthermore, with liberalized power systems, new possibilities for trading various electricity products would arise, which provides more flexibility for the market participants.

The main motivation behind energy sector liberalization was that electricity might be like any other product which can be purchased and sold in a market. However, electricity is quite a special type of commodity due to several reasons. First, it cannot be stored, which leads to different electricity demands in different parts of the day, week, or month. This happens because we are not able to store it when the demand is low and use it when the demand is high. Therefore, the electricity demand is sensitive to weather conditions and daily activities, and, therefore, we can categorize the daily demand as the base, peak, and off-peak. Base demand is the 24-hour electricity demand. Peak-hours are 8:00-21:00 and off-peak hours are 21:00-07:00. According to this, the electricity prices are also categorized as the base, peak, and off-peak. Because of this specific quality, the electricity price is characterized by seasonality and abrupt short-lasting spikes which cannot be easily forecasted (Harasheh, 2016). Consequently, during peak hours, the demand is high, which leads to higher prices. Due to this, the electricity market is exposed to market manipulations from the producers, which can intentionally decrease the production to increase the prices and make higher profits. The second reason is that the supply and demand should always match to ensure that the whole system is balanced and the electricity demand is covered. This is the responsibility of the system operator, which makes sure that the system is balanced and everything is executed properly. If that is not the case, certain fees are imposed on the market participants. Third, energy markets have a major role in modern society. Depending on what kind of economies we are currently living in (developed or emerging), the energy markets are crucial if we wanted to maintain our standard of living or if we wanted to make the transition to modernity. The whole world, from food supply to information technology, is dependent on a stable electricity supply. Therefore, a system collapse could be costly. The fact that there are very few substitutes for electricity makes this even harder to ensure. Finally, electricity production emits greenhouse gas. Thus, while combating climate change, there is always pressure on the electricity producers (and users) to switch to greener production which will provide a cleaner future (Thomas, 2004). This puts into place a new trend in the energy sector, increased electricity production from renewable sources. This made the electrical grid and system to be more unstable due to the frequent outages of electricity production from renewable sources. Because of this, the imbalances between electricity demand and supply increase which leads to more volatile prices. Therefore, electricity price forecasting is becoming a very crucial point for the decision-makers to alleviate the negative effects from the price uncertainty, which would provide a stable electric grid and increased economic benefits (Lago, De Ridder & De Schutter, 2018).

These characteristics make electricity production and prices very volatile. For this reason, proper electricity price, demand, and supply forecasting are important to help the decision-makers (power traders, system operators, regulators) in their daily activities. Subsequently, many different researchers are focused on finding appropriate and effective forecasting models using modern and fast algorithms.

Looking at the current technological developments and the improvement of computer power and performance, we can safely say that now, more than ever, the automatization of any mental task is not only feasible but also easy to execute. This is exactly why advanced computational algorithms gained huge popularity and are currently used for different purposes, from heart disease to stock prices prediction. Moreover, a question arises regarding how difficult the construction of such an algorithm is and how does it compare with its simpler counterparts. With this thesis, I try to answer this question and provide an overview of the process of constructing advanced predictive models and their predictive power.

This thesis is mainly focused on modeling the spread between the Austrian and German (AT/DE) electricity prices by using different models. The main aim is to find an efficient method that would predict the AT/DE spread. This model could be later used in constructing the Austrian electricity hourly price forward curves (HPFCs) forecasting the costs for Power transmission rights (PTRs) or other forecasts.

Along with the main aim of the thesis, there are several goals that I would like to fulfill. First, I want to provide a literature review of the existing methods of electricity price forecasting. Then, I would try to find the determinants of the electricity price and examine what mainly influences the volatility of the electricity prices. Furthermore, I will examine the special characteristics of electricity as a commodity which makes the electricity market design certainly different from the common capital markets. In addition to this, I will look into the users of the electricity price forecasts and the importance of electricity price forecasting. This would be analyzed through an overview and comparison of the organization of the German and Austrian electricity markets. Additional to the overview of the price modeling process, different statistical and more advanced predictive models will be compared. The forecasting process, data collection, and parameter estimation will be described as well. Additionally, I will elaborate on the different approaches and steps for constructing the models. Subsequently, I will examine the effectiveness of the standard statistical and the more advanced machine learning models. Finally, with this thesis, I try to answer the question if the complexity of the more advanced models is providing more efficiency in the prediction of electricity price spread or the results provided by the simpler methods are satisfying enough.

This thesis is structured in the following way. Section 1 provides an overview of the energy markets where the energy supply and demand are elaborated. It also consists of a brief explanation of the construction of an HPFC. Afterward, Section 2 consists of a discussion about the importance of electricity price forecasting. Furthermore, Section 3 includes an overview of the models that are frequently used for electricity price prediction. The following section consists of a discussion regarding the electricity spot price as the main subject in electricity forecasts. This section also includes an explanation of the performance measures that are used to test the accuracy of the predictive models. The organization of Austrian and German electricity markets is shown in Section 5. The next section presents the basic predictive model process which is used in this thesis. In Sections 6 and 7, I introduce the forecasting models used in this thesis along with some brief explanation of their characteristics. Section 8 includes the overview of the forecasting process and the presentation of the results. The conclusion of this thesis is elaborated in Section 9. The last two sections contain the reference list and the list of appendices.

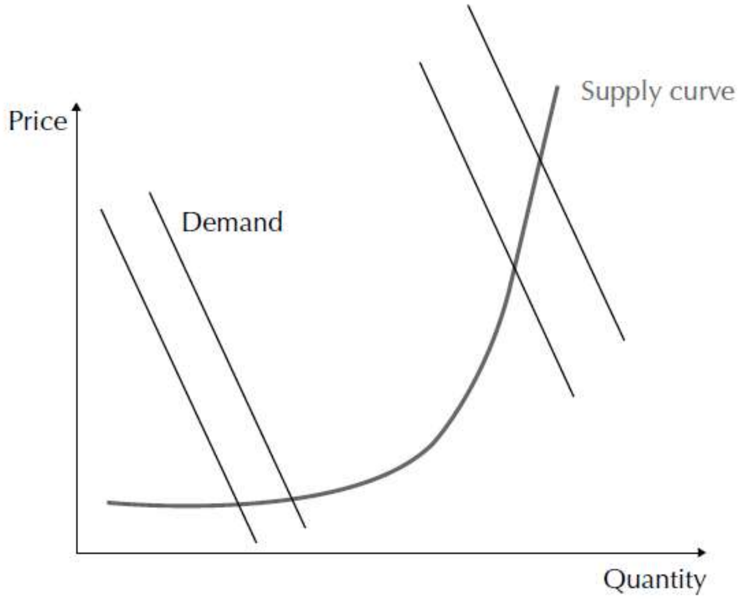
1 ENERGY MARKETS

The energy industry is quite complex with the complicated market design which then provides specific tradable instruments and energy contracts. A stable energy supply is crucial for the modern societies that we live in. Therefore, certain laws and regulations are set up to ensure its stability. The energy markets are hugely correlated with the rest of the global economy. Therefore, every major economic, political, or financial event can easily disturb the energy markets.

Usually, energy users have long-term contracts with the producers and they are based on formulaic prices (based on current or lagged spot and forward prices) which are set in short-term spot and forward markets. In the case of a spike in demand or lower energy supply, due to the inflexible supply sources, the end-users can encounter several energy shortages. On the other hand, in the case of demand decline, the temporary energy surplus can be sold on the spot markets which makes them reliant on the spot markets. However small these transactions seem to be for the whole energy market, their impact should not be overlooked.

That is because the inflexible energy supply chain combined with the necessity for marginal adjustment to the contracts can lead to decreased price sensitivity for both sides (suppliers and consumers), which, therefore, leads to more volatile prices. This is the so-called “tail wagging the dog” quality of markets where the relatively small transactions might have huge impacts on the whole market (Kaminski, 2013).

Figure 1: Energy demand and supply curves



Source: Kaminski (2013).

The price for any energy commodity is formed once the supply curve meets the demand curve. However, their shape is quite different since energy supply and demand have unique characteristics. Given constant energy production, the supply and demand curves can be expressed as in Figure 1. The supply curve is usually horizontal since prices have lower sensitivity to changes in the energy demand. That is because every slight increase of demand is covered with production from similar energy sources, with similar marginal costs. However, if the demand increases substantially, the energy production from more expensive sources has to be increased to satisfy the customers. Thus, the supply curve becomes vertical and it leads to skyrocketing energy prices. This is shown in Figure 1 with the interception of the third parallel line from the left with the supply curve. With each increase in demand, the prices increase even more. On the other side, the demand curve is pretty inelastic because energy is present in every part of our lives. Therefore, our usage is not dependent on the price. This is also because the energy consumers are not informed in time of the price changes since they receive their electricity or gas invoices at the end of the month and do not have the opportunity to decrease the demand in the case of increased prices. It is also important to be noted that the supply and demand of one energy commodity are dependent on the prices of other energy commodities. This makes all energy markets very codependent. Therefore, an energy trader or analyst has to understand the interconnection of all of them to make a useful prediction or decision (Kaminski, 2013).

1.1 Electricity trading

The liberalization of the energy sector revolutionized the way that electricity is exchanged. The traditional market principles were applied for electricity trading just like any other commodity, such as oil, gold, and silver. However, the non-storability of electricity makes the nature of electricity trading significantly different compared to the common capital and commodity markets. This comes from the fact that the electricity demand and supply must always be balanced wherefrom originates the necessity of quite an advanced engineering process for power flow management. That process includes controlling the electricity generation and balancing it with the demand by optimally dispatching electricity to consumers. Nevertheless, the advancements in technology provided means to electrical power engineers for finding optimal solutions for these issues made the power exchanges the main place for exchanging wholesale electricity worldwide (Stephenson & Paun, 2001).

The special characteristics of electricity as an energy commodity, such as its non-storability, bring additional necessary regulations. Therefore, a special transmission operator is needed for managing the security of the electricity system. Basically, it coordinates the supply and demand of electricity and ensures that the system is balanced anytime. In the case some party is not balanced, it should cover its imbalance fees. Other participants in the electricity market are the electricity producers (owners of power plants), trading companies, and electricity providers to end customers. Electricity producers are the suppliers of electricity on the market. The main electricity production sources are extracting and burning fossil fuels (coal, natural gas, oil, etc.), nuclear power plants, hydropower plants, wind, and solar power plants. These electricity sources have different producing power and costs. Renewable sources including hydro, wind, and solar power plants have the smallest marginal costs, which makes them the cheapest power source. However, since they are highly dependent on the weather conditions, they are also the most unreliable. For this reason, even in the countries with huge renewable electricity supply, conventional power plants are still needed to cover the electricity demand when the renewable electricity sources are undersupplying. On the other hand, even though the conventional power plants are more stable, they are costlier and simultaneously extremely bad for the environment whereof comes the incentive of transition to renewable sources for electricity production.

Competing electricity producers depend on the transmission network for dispatching and scheduling their power plants to sell the produced electricity within the organized spot and forward markets. This is concluded through bilateral contracts between electricity producers and end customers or through intermediaries which afterward supply the end-users with electricity (Joskow & Tirole, 2000). These market intermediaries are the electricity trading companies that are participants in the wholesale electricity markets. Trading companies are maximizing their profits through purchasing and selling electricity which benefits the other market participants since they facilitate the exchange of electricity from producers to customers. Numerous electricity trading companies are also providing electricity to the end customers. If this is not the case, they sell it to distribution companies that supply the electricity to the end customers.

The electricity exchange could be performed on the electricity market or directly between counterparties known as "over the counter" (OTC). Most of the developed electricity markets function as a combination of organized exchange with bilateral contracts. If the electricity trading is executed on the electricity market, the buyers and sellers put their bids in the system while another significant electricity market participant called the market operator clears the market and constructs the electricity prices (according to the supply and demand curves) for the next day, which is known as day-ahead trading (Amjady & Hemmati, 2006). Section 5 covers a more detailed overview of the organization of the German and Austrian electricity market which is the main topic of this thesis.

To trade electricity outside national borders, a transmission right is needed for transferring the electricity from one country into another. Generally, cross-border electricity trading and transmission rights allocation are executed on two different markets, except in the case of market coupling. More details regarding the market coupling can be found in Section 5. When these two activities are not combined, one has to acquire the PTR to transfer the electricity across borders. The transmission rights are purchased on auctions organized by special allocation offices (or some transmission system operators in the case of intra-day auctions/reservations). There are short-term (daily, intra-day) and longer-term power transmission rights available for purchasing on the actions. Once acquired, long-term PTRs could be used for cross-border trading or re-sold on the market.

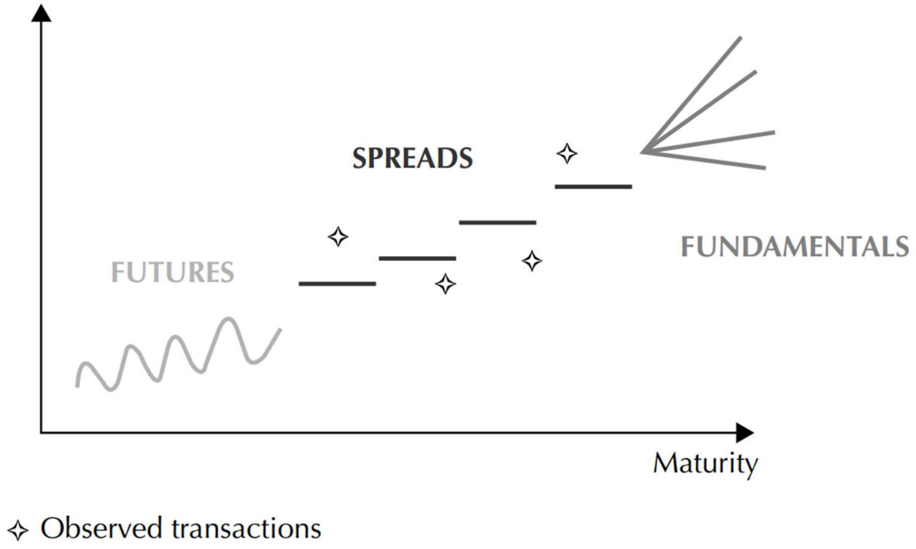
1.2 Spread transactions

In the energy markets, there is a product that could be traded which is related to the locational and calendar difference between the energy prices, the so-called spread. The locational spread is expressed as a difference of prices between two countries where the calendar difference called time spread is a difference between the energy price of one country of different time periods (e.g. Germany 2021 and Germany 2022 product spread). In some cases, the spread products are more actively traded than the absolute price of some commodity (e.g. Germany Base December product) called outright. In this case, the absolute price would be formed indirectly from the prices of the traded spreads (Kaminski, 2013).

1.3 Forward price curve

The collection of energy prices for some future time period is called a forward price curve (FPC). This price curve looks at the energy prices with different maturities when they extend into the future, which means that it does not represent an actual forecast of the prices for that time period. If we look at the FPC today, it only serves as an overview of the future energy prices as it is currently agreed between consumers and producers. In other words, the FPC represents the current forward energy prices. It can be useful for risk managers for their daily analysis of the trading portfolios. However, it can also serve as an indicator for traders which will be helpful in their decision processes. The frequency of the specification of the forward curves depends on the characteristic of each energy market. Usually, for oil and natural gas markets, the forward prices are quoted monthly whereas for electricity the usual frequency is hourly prices. Therefore, for electricity, the hourly FPC are usually used. Figure 2 shows the process of construction of an FPC.

Figure 2: Construction of forward price curve



Source: Kaminski (2013).

The front part of the curve can be constructed from the information observed on the futures market ignoring the potential difference between forwards and futures prices, in the case, if a futures market exists for that energy commodity that we construct the curve for. Then, traders can extract information from the calendar spreads, if present, which should be properly reconstructed to the higher granularity of prices that we would need. Additionally, traders could extract information from the actual transaction and bids on the markets and market observations and communications with counterparties and brokers. Finally, information regarding the fundamentals of energy markets could be used in constructing the price curve. Another solution would be to use advanced algorithms which would model the whole physical system (predicting the future demand and supply curves for that energy commodity) to construct the forward price curve. There are also hybrid models which combine both approaches in constructing the forward price curves (Kaminski, 2013).

After the construction of the HPFC, it should be adjusted for seasonality due to different demand and supply in different seasons. Afterward, it should be further adjusted to make the HPFC arbitrage-free where the constraints ensure that the curves appropriately replicate the futures prices that could be observed on the market. This is an advanced process for constructing and adjusting the HPFC and it will not be elaborated further in this thesis (Sætherø, 2017).

Due to the importance of the HPFCs for the energy trading companies, they usually try to use the best model for constructing the HPFCs. In the case that there is a futures market for a certain energy commodity, models are available for construction of the HPFCs from the futures prices, like the one presented in Sætherø's Doctoral thesis (Sætherø, 2017). However, when there are no reliable futures prices published daily, the task of constructing the HPFCs could be more difficult. This can be solved if there is a very correlated price to the energy commodity's price for which we are interested and through which we can model its HPFCs. For example, if we liked to model the HPFC of the Austrian electricity price because there are only few traded futures products, we would not be able to use the Austrian futures prices. Therefore, a solution is to use the futures price of a highly correlated electricity price of a neighboring country which can be used as an indicator for the HPFC modeling. This can be the German electricity price which is quite correlated to the Austrian due to several reasons that are elaborated in the next chapters. Additionally, the German futures market for electricity products is more liquid which would be appropriate for the HPFC modeling. The next step is to try to model the spread between the German and Austrian electricity prices, which could be then added to the German futures prices to get the Austrian futures prices for constructing the HPFC for Austrian electricity price. This thesis is mainly focused on the AT/DE spread modeling and the review of the process of creating the predictive models and effectiveness of the standard statistical and the more advanced machine learning predictive models.

2 ELECTRICITY PRICE FORECASTING

2.1 Importance of electricity price forecasting

As previously mentioned, deregulation of energy markets brought various changes to the energy sector which was previously controlled by the government by introducing competitive market rules, mainly to reduce the electricity costs through promoting competition. This newly established electricity market with new market rules has different market participants which were described in Section 1.1. Those different participants have different information needs. Therefore, different types of forecasts are available which include electricity demand (load), supply (stack), and price forecasts. Electricity trading companies use the forecasts to adjust their bids and monthly schedules, hedge the volatility of the prices, value the bilateral contracts with their counterparties to maximize their benefits from the transactions, etc. Information from electricity forecasts is quite useful for the producers as well, because they would be able to optimize their production schedules according to the electricity price forecasts to minimize the production costs and maximize the benefits from the trading contracts. Consumers usually have to come up with their decisions if they get their electricity from bilateral contracts or the electricity exchange. Therefore, they can use the price forecasts to choose the most beneficial option. Energy service companies are also users of the price forecasts since they try to efficiently manage their bilateral contracts and contracts on the energy exchange to maximize their benefits and the benefits for their end consumers (Amjady & Hemmati, 2006).

With the increased activity on the electricity markets, reliable forecasting is crucial to ensure a stable electricity system. Therefore, the necessary electricity price forecasting models are developed using advanced technology. Using advanced forecasting models is also important due to the special characteristics of electricity prices which arise from the high volatility, huge impact of uncertain events on the prices, and complicated bidding strategies. Electricity prices in the most competitive markets have the following characteristics:

- Very frequent (usually hourly or quarter-hourly);
- Nonstationary (the mean and variance are not constant);
- Seasonal and calendar effects (different characteristics for weekends, holidays, seasons like summer or winter);
- Huge volatility (due to the high sensitivity from events in the electricity markets);
- Outliers (huge spikes in prices).

As a consequence of these characteristics, electricity prices are quite uncertain and difficult to predict. Since they are very sensitive to events in the electricity markets, once new information is available, the current forecasts become obsolete. Therefore, we need to incorporate the new information as soon as possible in our predictions and to include this uncertainty in as well. Taking this into consideration, electricity market analysts are trying to use advanced models which can be easily adjusted on the newly available information to provide more efficient forecasts (Amjady & Hemmati, 2006).

2.2 Literature review

In the last decade, various forecasting methods for electricity markets have been developed. Most of the analysts are focused on short-term load or short-term price forecasting. Since the prices are more volatile, analysts need more advanced models to incorporate the uncertainty and the huge volatility. Many models that provide accurate load and price forecasts are already available and a short overview of some of them is presented in this section. Because of the fact that predictive models for electricity price spread forecasting are scarce, this section comprises of literature review of electricity price forecasting models.

As classified by Weron (2014) and later elaborated in the paper by Lago, De Ridder, and De Schutter (2018), the electricity price forecasting models can be divided into five areas:

- Game theory models – multi-agent models that are used for simulating the interaction between the heterogeneous agents (producers and companies), for creating the price by matching the electricity supply with the demand;

- Fundamental models – finding the physical and economic determinants of the electricity price;

- Reduced-form models – quantitative models for identifying the statistical characteristics of the electricity prices used mainly for evaluation of derivatives and risk management;

- Statistical models – statistical or econometric techniques for load or electricity price forecasting;

- Computational Intelligence models – non-parametric, non-linear models that could be adapted to complex price dynamics and can learn and improve themselves with the newly provided information.

Additionally, Weron mentions that most of the approaches proposed in the current literature are hybrid models consisting of methods from multiple groups that were mentioned above. Even though there are different alternative models, linear regression (LR) models are still one of the most used models for electricity price forecasting. To deal with their disadvantages, they are usually combined with more sophisticated models to achieve more efficient results (Weron, 2014). Convincing results from linear models can be seen in the paper by Kath and Ziel (2018), who propose two general regression models for quarter-hourly electricity prices for German spot markets. For forecasting the EPEX German 15 min price they propose a multivariate elastic net regression model which is an extension of the ordinary least squares (OLS) optimization with added linear penalty factor whose objective is minimizing the residual sum of squares (RSS) and simplifying the structure of the model (Kath & Ziel, 2018).

As explained above, statistical models are used for price forecasting by mathematical combinations of previous prices and/or exogenous variables, such as production, consumption, or weather. These models are quite attractive since a reasonable interpretation can be derived from them which can help decision-makers, market participants, or system operators to better understand the whole picture of how the variables affect the prices and the whole electricity market. Some of the most commonly used statistical models for electricity price forecasting are multiple regression, autoregressive (AR) models, autoregressive time series models with exogenous inputs (ARX), generalized autoregressive conditional heteroscedasticity model (GARCH) which is developed to treat the non-constant standard deviation of the predicted variable over a period of time and many other models derived from them (Weron, 2014). Amjady & Hemmati (2006) discuss in their paper that various autoregressive models are used for forecasting weekly or daily electricity prices in the Norwegian system. Additionally, they show that the extension of the AR model, autoregressive integrated moving average (ARIMA) model provides more efficient results in the forecasting of electricity prices of Spanish and Californian electricity markets (Amjady & Hemmati, 2006). The extensive paper from Gürtler and Paulsen provides an overview of the efficiency of the time series forecasting models for the German/Austrian electricity spot prices. They analyze multiple versions of ARIMA and GARCH models while applying several data transformations and spike adjustments. The best-performing forecasting model after their conducted study is autoregressive–moving-average with exogenous inputs (ARMAX) model while GARCH models are slightly less accurate. Gürtler and Paulsen add that including the electricity demand and electricity production from renewable sources as an input variable in their models considerably increases their accuracy. They also mention that data for at least 365 days is necessary to observe all seasonal characteristics of electricity prices. Another conclusion is that in their study, log-transformation does not necessarily improve the performance of the models while spike preprocessing does help in having more accurate forecasts (Gürtler & Paulsen, 2018).

According to Lago, De Ridder and De Schutter (2018), statistical and machine learning models provide the most accurate results, which is the reason why they present a broad comparison between 27 electricity price models from both groups. They were studying the models for day-ahead forecasting of the EPEX-Belgian electricity price. From the statistical models, they use AR and ARX, ARIMA, GARCH, dynamic regression, and transfer function models. The authors of this paper describe that the major disadvantage of the statistical models is that they are usually linear forecasters which might not be accurate for high-frequency data with huge volatility. To address these issues better, they propose using some more advanced machine learning models that are part of the computational intelligence (CI) models which were described in the last group of models in the classification from Weron (2014). This group consists of computational techniques developed to increase the efficiency of the traditional models. They are models that can adapt to complex dynamic systems. The main CI models are artificial neural networks, fuzzy systems, and support vector machines (SVM) which are flexible and can handle complex systems and non-linearity (Weron, 2014).

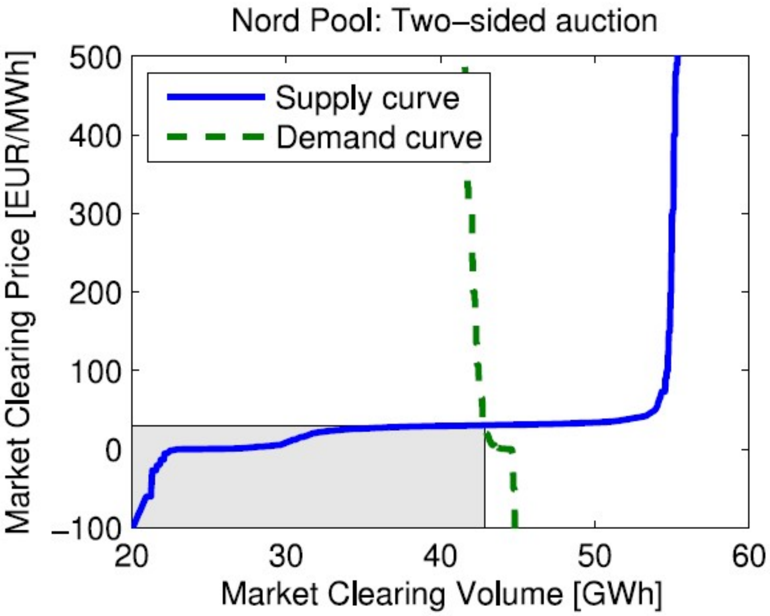
In the study conducted by Lago, De Ridder, and De Schutter (2018), they mention that using a combination of an advanced machine learning model increases the wind speed forecasting accuracy by 30%. Additionally, using the convolutional neural networks (CNNs) model, they obtained better wind power forecasts. Also, they describe that a combination of extreme gradient boosting (XG boost) and deep neural network (DNN) model is very efficient in load forecasting. They conclude that after the overview of all the models, the machine learning models are statistically significantly better than the statistical ones. According to them, an exception to this conclusion is the ARX-based models because even though they are statistical methods, they clearly overperform the other statistical methods and some machine learning ones. Another conclusion of their paper is that moving average models is performing significantly badly while the hybrid methods do not outperform the simpler similar models. The possible reason for these results is that the dynamics of the electricity prices in the last decade due to the increased electricity production of renewable sources leads to higher volatility and higher spikes in the prices, which then makes the traditional models not doing quite a good job in electricity price forecasting (Lago, De Ridder & De Schutter, 2018). A better solution for this non-linear optimization problem is found in machine learning algorithms which became very popular nowadays due to their flexibility and high accuracy in forecasting different topics from medical science and sales to energy markets. The idea for this kind of model that can learn and improve itself was born after the discovery of the statistical methods (Least square methods in the year 1805 and Bayes' Theorem in 1812). The base of machine learning was established in 1950 when Alan Turing proposed a machine that could learn and become artificially intelligent. Later the artificial neural networks were born in 1951 in an attempt to replicate how neurons work in the human brain. This was used to build the first neural network which was improved significantly later with the new research and increased computer power (Wikipedia, 2016). Subsequently, many developments in artificial intelligence were introduced which significantly increased the interest in research in this area. Along with the development of computers and their power, more powerful engines were developed, which is relevant for extracting and using more data in the prediction models. The increased computational power is also necessary for building more complex algorithms, which can require more advanced parameter estimation and available memory power for setting up the models.

This thesis will complement the existing literature by proposing a model for AT/DE electricity price forecasting. Through the construction of the model, I will try to present the forecasting process using advanced algorithms. First, I will make an overview of the organization of the German and Austrian electricity markets and their interconnectedness. Afterward, I will review the data preparation and variable selection process where it is important to ensure that we have the proper data format and variable's granularity to construct the model properly. Furthermore, I will present the key steps for building a predictive model. To sum up, I will present the results and make an inference regarding the efficiency of the models. I am confident that this thesis will complement the literature by providing additional topics for further research and analysis. The AT/DE Spread forecasting model with further adjustments can be used for the construction of the Austrian electricity HPFCs, PTR costs, or some other price predictions.

2.3 Electricity market clearing price

The electricity price which is studied in this thesis is actually the day-ahead price. The electricity spot market does not allow continuous trading because system operators require advanced notice of the electricity demand and supply to make sure that the order is within the constraints of the transmission grids. The organized electricity market should determine the market-clearing price (MCP) as an intersection between the electricity demand and supply curve constructed from the bids entered in the system in the daily auction (Weron, 2014).

Figure 3: MCP for Friday 3/1/2014, 18-19 hours in Nord Pool power exchange



Source: Weron (2014).

The construction of the MCP is presented in Figure 3, which shows the supply and demand curve. They are based on aggregated supply and demand bids in the auction for Friday, 3/1/2014 for 18:00-19:00 hours. For that specific trade day, the MCP is 30,94 EUR/MWh. As previously mentioned, we can see that the electricity demand shown with the green line is almost inelastic to the MCP while the electricity supply is flat until the market-clearing volume of 55 GWh, which could be the max capacity of the power producers. If the electricity demand passes the threshold of 55 GWh, they should activate the more expensive electricity power plants which could drive the price up to 500 EUR/MWh. Each power plant outage, when it cannot produce the planned electricity, could increase the electricity price significantly if there are no power plants with similar costs that would provide the necessary electricity. This is the reason why the electricity price is volatile and very sensitive to fluctuations in electricity supply.

2.4 Predictive modeling process

A predictive model is usually a set of equations with specific adjustable parameters that we create with many observations of data to understand the correlations between certain variables and to infer some conclusions regarding certain output variables. We start the process by observing the real system that we would need to model through analyzing the available data observations. Through this analysis, we try to understand the interdependencies, features, and characteristics of the observations and the whole system. The steps for creating a predictive statistical model are the following:

- **Defining the goals.** Defining what is the outcome of the model, how it will be used, and what could be stated as a good model

- **Getting the data.** Trying to get as much data as possible which could be considered as a determinant of the outcome variable

- **Constructing the model structure.** Defining what kind of model would we construct, linear/logistic regression, non-linear model, etc. Usually, it is best to start with a linear model as a base model and then find more complex models which could overperform the linear one.

- **Data preparation.** Preparation of the gathered data which consists of data quality examination, taking care that the data is an inappropriate format, handling missing values, properly categorizing the data and examining its distribution, visually examining the data, calculating summary statistics for each data set, performing scaling and transformation of the data, and, finally, separating the data into training, testing and validation data sets.

- **Variable selection.** This is a fundamental process for building a good predictive model since we will have to decide which variable we would want to use. In this selection process, some knowledge and expert information for the topic that we are working on will be beneficial. First, we start by choosing a large set of variables that could be possible determinants of our output variable. Then, through our expert knowledge and some statistical techniques, we choose a subset of variables which when included in the model provide high predictive power and stability. Through this process, we should examine the importance of the input variables to the model and eliminate the ones that do not influence our outcome variable. This is an important step before we start with the construction of the model, which ensures that we will not use any redundant input variables which could lead to overfitted model.

The best practice is out-of-sample valuation by separating the data into training (the one used for creating the model) and testing (the one used for testing the performance of the model) data sets.

- As from the variable selection methods, stepwise selection methods are the ones that are mostly used. There are forward and backward selection methods where with each step, the number of selected inputs of our model is increased or decreased by exactly one. For example, in the case of forwarding selection, we start with a model with 0 inputs out of our n possible sets of variables. Then, we calculate n models, each consisting of only one of the inputs. We choose the best model and we continue to the next step. Afterward, we calculate n-1 models including the remaining n-1 sets of variables. In this step, we again choose the best model and continue to the next step until we reach a certain number of variables or until any further step (where we include one more set of variables) does not significantly improve the results of the model. The backward selection model has the same properties as forwarding selection. However, the direction is the opposite. We start with the whole n sets of variables and in each step, we remove one set of variables that is not significant for the model.

- **Comparison of created predictive models.** The best practice is to start with a simple linear model and then build upon the results out of the first model.

- **Determination of the final model.** Selecting the most appropriate and efficient predictive model. The efficiency of the predictive models could be compared with certain error measures.

- **Preparing for the implementation of the chosen model and monitoring.** Preparation for the implementation process and determining the steps for monitoring the performance of the predictive model could serve as indicators for further model optimizations (Wu & Coggeshall, 2012).

2.5 Evaluation of the results of the forecasting model

The accuracy of the forecasting model is usually evaluated based on measures for calculating the difference between the actual price y_i and the predicted price $\hat{y}_t$. One of them is a mean error (ME) expressed in Equation (1.1) as an average of the differences between y_i and $\hat{y}_i$.

$$ME = \frac{\sum_{t=1}^{T} \hat{y}_t - y_t}{T} \tag{1.1}$$

However, this error should be used cautiously because positive and negative errors could cancel out and show incorrect accuracy of the results of the prediction. Therefore, mean absolute error (MAE) is a more commonly used performance measure. It is expressed in Equation (1.2) and calculates the average of the absolute differences between the prediction and the actual values.

$$MAE = \frac{\sum_{t=1}^{T} |\hat{y}_t - y_t|}{T} \qquad (1.2)$$

Another popular performance measure is the root mean squared error (RMSE) which calculates the square root of the average of squared differences between the predicted and actual price. It is expressed with the following equation:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} (\hat{y}_t - y_t)^2}{T}} \qquad (1.3)$$

3 ORGANIZATION OF GERMAN AND AUSTRIAN ELECTRICITY MARKET

Electricity has a very special role in our society, from providing us with light to powering up all of the industrial processes which are crucial for our daily living. With that being said and due to its special characteristics, a real-time balancing of the market participants is necessary. This means that the stability of the system depends on the short-term power markets. The major player in the short-term market is the National electricity market operator (NEMO), which usually is an electricity exchange. Together with the Transmission system operator (TSO), they have their specific tasks:

- Accepting the bids from participants in the short-term power market;
- Responsible for allocating orders according to the results of the day-ahead and intraday markets;
- Publishing the electricity prices after construction of the supply and demand curves;
- Settlement of the trading transactions.

One NEMO could be responsible for one or more countries. For example, three NEMOs are relevant for Germany: European Power Exchange (EPEX Spot), Austrian Energy Exchange (EXXAA), and Norwegian Nord Pool who are responsible for the German bidding zone (Market coupling, 2019). This thesis is focused on the overview of the EPEX Spot power exchange and the determinants of the spread of EPEX Spot German and Austrian prices.

Power exchanges provide a platform where participants could submit their purchase or sell bids. After the submission of the bids until the specified hour when the daily auction ends, the MCP is calculated. It is published and the exchange ensures that the traded quantities are delivered and paid. This published price is the most reliable price for the short-term market. The transactions on EPEX Spot are cleared by ECC, which serves as the European commodity clearinghouse. ECC is also connected with the corresponding TSOs to ensure the physical and financial settlement of the transactions concluded on EPEX Spot.

EPEX Spot was a major contributor to European Market Coupling which allows free flow of electricity across European borders. Currently, 19 European countries are part of this interconnected electricity market which connects the control and market areas to harmonize the systems of the electricity exchanges and to minimize the price differences. With this connection, it was planned to transfer the electricity easily whenever it is needed the most using national and electricity grids from the neighboring countries. This mitigates the supply and demand disturbances that come from renewable energy (the major producer from renewable sources is Germany as the biggest EPEX market) by balancing the energy between the European zones. It also helps further in smoothing the positive and negative electricity price peaks. Market coupling is available in both day-ahead and intraday markets on EPEX Spot where transparent and secure transactions are ensured (Epex Spot, 2020).

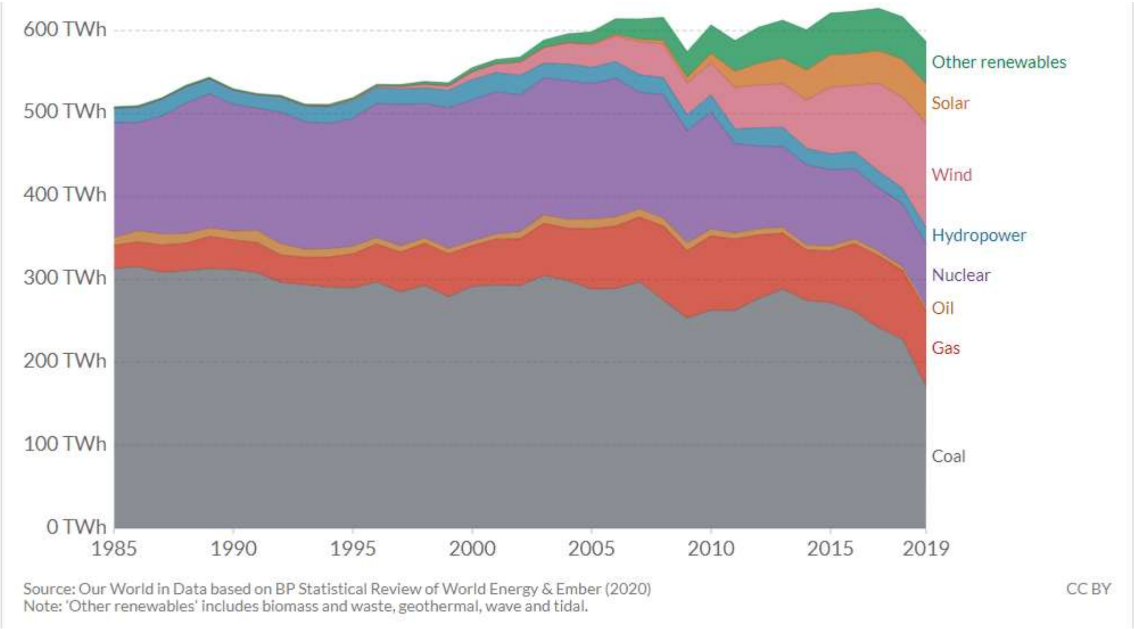
The electricity trading is executed either in the day-ahead/intraday market, OTC, or in the forward markets. For Germany and Austria, the day-ahead and intraday trading is performed on EPEX Spot while the trading in electricity forward products on the European Energy Exchange (EEX). EPEX Spot operates the most liquid day-ahead and intraday markets in Europe: Austria, Belgium, France, Germany, United Kingdom, Luxembourg, the Netherlands, and Switzerland. The day-ahead trading is executed through a daily auction that takes place once per day where all hours of the following day are traded. The auction closes at noon (11:00 in Switzerland) and then the algorithm for calculating the market-clearing price is launched. As previously discussed, the MCP is the intersection of the demand and supply curve and it applies to all buyers and sellers. The intraday market works differently when the trades are executed continuously, 24 hours a day up to 5 minutes before delivery. Hourly, half-hourly, and quarter-hourly contracts are available that allow high flexibility for the market participants to balance their positions. Here, the price is established once the buy and sell bid meet and this represents only the price for that contract. With that in mind, we can note that intraday market price has different dynamics than the day-ahead price. This thesis will be focused on the determinants of the day-ahead electricity prices (Epex Spot, 2020).

Negative prices are possible in the EPEX Spot Market. The electricity prices fall once the demand is low and the supply is very high and the producers cannot stop their power plants with low costs. This is the reason behind the negative electricity prices which show that for producers it is better to pay buyers than to stop the production of their power plants (Epex Spot, 2020).

Germany and Austria were part of the same bidding zone together with Luxemburg before 1/10/2018, which means that electricity could be traded between these countries without limits and for the same price. However, the renewable electricity production in Germany increased drastically which reduced the prices of the electricity produced in Germany. Therefore, the interest in dragging the electricity from Germany to Austria to get the benefits of this cheaper electricity source also increased substantially. This led to congestions in the power grids on the border from Germany and Austria. Since they were in the same bidding zone and the electricity transmission was not restricted when the congestion happened and the German/Austrian power grid exhausted its physical transmission capacities and was not able to transmit the whole electricity that was purchased from Germany. Therefore, the power grids from the neighboring countries (Poland and Czechia) were used to bring the electricity to Austria. This is the reason why the bidding zone was split into two zones: Germany/Luxemburg and Austria. This relieved the pressure on German and neighboring power transmission grids for transmitting the necessary electricity from Germany to Austria. This is an advantage for those countries since they had their power grids burdened with the electricity flow even though they did not benefit from it. From 1/10/2018 onwards, only 4.9 GW could be traded at the German/Austrian border at any time. Because of the market coupling that exists when the trading volume does not exceed the limit of 4.9 GW, the long and short-term trading is still possible and the prices will be identical as before. However, once the trading volume over the border to Austria exceeds the threshold, the German and Austrian electricity prices will differ because the electricity demand has to be covered from other (usually more expensive) power plants in Austria (SMARD, 2018). Usually, electricity trading outside national borders and the allocation of the power transmission rights are two separate markets. However, the European market coupling integrates the electricity markets and the electricity price formed in the markets consists also the costs of transmission rights for transmitting electricity across borders (Market coupling, 2019).

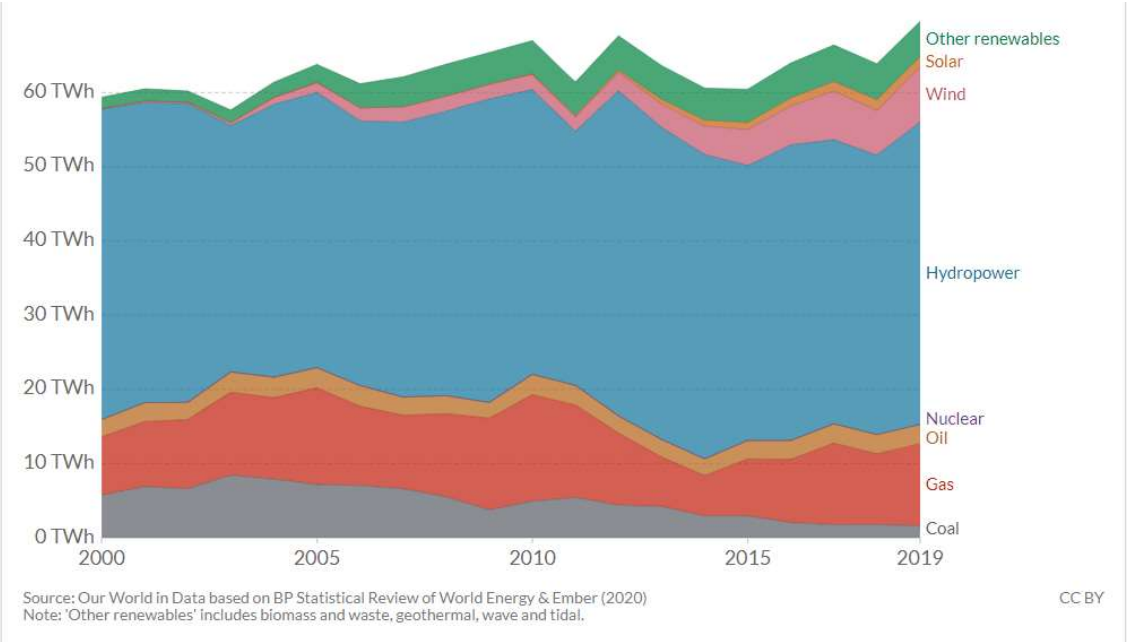
Major sources for the electricity produced in Germany are coal and hydropower plants. However, from the 2000s the increase of electricity production from renewable sources is evident and especially from wind and solar power plants. That is because of the energy transition to renewable sources in Germany that was encouraged with the Renewable Energy Sources Act which was planned to support large-scale construction of renewable power plants by providing expensive feed-in tariff schemes (Pflugmann, Ritzenhofen, Stockhausen & Vahlenkamp, 2019). This was also introduced in many countries to stimulate electricity production sustainably. The feed-in tariffs were introduced and the government provided a guaranteed price to the electricity producers. This would mean that the government compensates the producer for the difference between the guaranteed and current market price (Huisman, Stradnic & Westgaard, 2013).

Figure 4: Electricity production by source, Germany



Source: Ritchie (2020).

Figure 5: Electricity production by source, Austria



Source: Ritchie (2020).

These new policies provided enough incentives and resources for the construction of different wind and solar power plants within Germany. This trend can be observed in Figure 4 where, starting in the late 1990s, electricity production from renewable sources was initiated. Despite the fact that this energy transition took several years, it was executed successfully. This is evident from the increased participation of renewable sources in the total electricity production in Germany. The share of renewable sources in the total production grew from 35% in the 2000s to around 50% in 2019 where the main source of renewable electricity is the wind power plants.

The electricity production in Austria is much smaller than the German production (60 TWh Austria, 600 TWh Germany). Figure 5 shows the electricity production in Austria by sources. We can observe that hydropower is the major source of electricity production in Austria. An increase in renewable power production (wind, solar, and other renewable power plants) is also evident in Austria from 2005.

3.1 Liquidity of German and Austrian electricity financial products

The problem with constructing the HPFC for the Austrian electricity prices came from the low liquidity in futures products that are traded on the EEX. The liquidity of futures contracts can be observed by two technical metrics which are volume and open interest. The metric volume presents the number of contracts of a certain product in a given period while open interest shows the number of active contracts (traded electricity futures contracts) which are still not settled (Nickolas, 2020).

Table 1: Volume and open interest for German and Austrian futures

Date	Contract Name	Volume DE	Open Interest DE	Volume AT	Open Interest AT
2/11/2020	Cal-2021	326	136 733		208
3/11/2020	Cal-2021	420	137 687		208
4/11/2020	Cal-2021	245	137 545		208
5/11/2020	Cal-2021	441	138 156		208
6/11/2020	Cal-2021	292	138 568	5	213
9/11/2020	Cal-2021	545	139 312		213
10/11/2020	Cal-2021	374	139 815		213
11/11/2020	Cal-2021	416	140 363	4	217
12/11/2020	Cal-2021	299	141 056		217
13/11/2020	Cal-2021	152	141 146		217
16/11/2020	Cal-2021	478	141 146	2	217
17/11/2020	Cal-2021	199	141 957	2	221
18/11/2020	Cal-2021	234	142 442		226
19/11/2020	Cal-2021	269	142 831		226
20/11/2020	Cal-2021	243	143 452		251
23/11/2020	Cal-2021	214	143 576		251
24/11/2020	Cal-2021	254	143 576		386
25/11/2020	Cal-2021	764	145 006	3	401

Source: Montel news (2020).

In Table 1, we can see the daily traded volume and the open interest on EEX for both German and Austrian futures products for November 2020. The traded product is Germany Base Year 2021 and Austria Base Year 2021 product.The German futures contract is traded more than the Austrian one. For example, for trade date 11/11/2020, traded number of German contracts was 416 while for Austrian was only 4. On the same trade date, there were 140 363 unsettled futures contracts for the German year 2021 product while there were only 217 for the Austrian product. From this, we can see that the liquidity of the Austrian futures market is way lower than the German one. Therefore, the construction of the HPFC from the Austrian futures price will not be possible.

4 STATISTICAL FORECASTING MODELS

Statistical modeling has a huge impact in various areas of study ranging from sales, medical science to insurance and finance. Its main role is to gain some information from the available data. The usual process is gathering data about the output (which can be quantitative or categorical) and a set of features of some objects that we think are describing the outcome. Using the data, we construct a prediction model also called a learner. Having a good learner means that the model predicts the outcome quite accurately. The different types of outputs lead to two different prediction models: regression and classification. There are also two types of statistical learning supervised and unsupervised learning. Supervised learning means that we already have the outcome data which will help us in the learning process. Unsupervised learning does not have measurements of the outcome and our task would be to get more insight regarding the organization or clustering of the variables that we work with. The statistical learning problem in this thesis is a supervised learning problem since the main goal is the prediction of the AT/DE electricity price spread.

Notation: We usually denote the input variables with X and in the case, if X is a vector, its components are expressed with subscripts for ex. X_i. The output variables are denoted by Y. Using these notations, we can denote the statistical learning problem of this thesis as follows: given the value of the input variables in the vector X, construct a model for accurately predicting the output variables Y denoting it by $\hat{Y}$ (pronouncing it “y-hat”). To construct the prediction rules, we use the available set of measurements of observed input and output variables ((x_i, y_i), where i = 1, …, N) which is known as training data (Hastie, Tibshirani & Friedman, 2008).

4.1 Multiple linear regression

Statistical models are used for price forecasting by mathematical combinations of previous prices and/or exogenous variables. Although there are many alternatives, linear regression models are still popular methods for electricity price forecasting. However, many authors combine them with more advanced models to achieve efficient predictions. Regression is still described as the most widely used statistical model. Mainly, this is because it is simple and can sometimes provide an easy and appropriate description of how the inputs affect the outputs, which can sometimes outperform more complicated nonlinear models. The multiple regression model aims to examine the relationship between the independent and dependent variables. The classical model assumes a linear relationship between the variables and is expressed in Equation (2.1)

$$\hat{y}_t = \beta X_t + \varepsilon_t = \beta_1 X_{1t} + \cdots + \beta_k X_{kt} + \varepsilon_t \tag{2.1}$$

where β_k is a vector of coefficients, X_k a vector of explanatory variables, and ε_t is the error term. The coefficients are estimated by minimizing the sum of squares. The coefficients β_k express how the explanatory variables are correlated with the electricity price $\hat{y}_t$. After the estimation of the coefficients, we can use them to predict the future electricity price (Weron, 2018).

The most popular method for estimating the coefficients in multiple linear regression is the least-squares method. Basically, we find the coefficients β that minimize the residual sum of squares

$$RSS\ (\beta) = \textstyle\sum_{t=1}^{N} (y_t - x_t^T \beta)^2. \tag{2.2}$$

The residual sum of squares is a quadratic function of the parameters and it has always a minimum. However, it might not be a unique minimum. If we change to matrix notation, we have

$$RSS\ (\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)^{\,T} (\boldsymbol{y} - \boldsymbol{X}\beta) \tag{2.3}$$

where each vector with input variables is a row in **X** (which is an $N\ x\ p$ matrix) and $\boldsymbol{y}$ is the N-vector of observed output variables from the training set. If we find the derivative w.r.t β, we get the following equation:

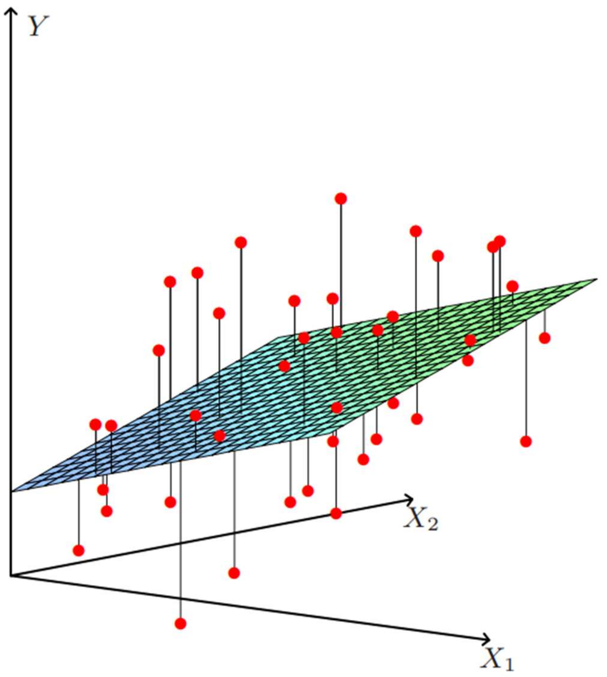
$$\boldsymbol{X^T}(\boldsymbol{y} - \boldsymbol{X}\beta) = 0. \tag{2.4}$$

In the case, if $\boldsymbol{X^T X}$ is nonsingular, the unique solution to our problem is the following equation

$$\hat{\beta} = (\boldsymbol{X^T}\,\boldsymbol{X})^{-1} \boldsymbol{X^T y} \tag{2.5}$$

where the prediction of the t-th input x_t is $\hat{y}_t = \hat{y}(x_t) = x_t^T \hat{\beta}$. The least-squares decision boundary for the regression problem is smooth and stable to fit. This solution relies only on the assumption that the relationship between the input and the output variables is linear which makes a linear decision boundary appropriate. In other words, it has low variance but potentially a high bias (Hastie, Tibshirani & Friedman, 2008).

Figure 6: Linear least squares estimator of Y using X



Source: Hastie, Tibshirani & Friedman (2008).

Figure 6 shows the least-squares optimal solution to our problem in the case if $X \in \mathbb{R}^2$ where the squared distances from the red points to the regression hyperplane are minimal.

4.2 Variable selection in the regression model

As elaborated in Section 2.4, which consists of an explanation of the predictive modeling process, for the variable selection, we can use the forward or backward stepwise model. This section will include an overview of the backward stepwise model for multiple linear regression. We should start by constructing a regression model including all explanatory variables and then, with each step, take out the variable that does not have a significant impact on our output, which would mean that the coefficient β_k from equation (2.1) for that variable X_k is 0. For this reason, hypothesis testing can be used by assigning the null hypothesis as H_0: β_k = 0. For testing this hypothesis, a t-test is used where the test statistic has t-distribution under the null hypothesis:

$$T_k = \frac{\widehat{\beta}_k}{\sqrt{\widehat{\sigma}^2\,(X^TX)^{-1}}} \sim t_{n-p-1} \tag{3.1}$$

The p-value of the test statistic is used to conclude if the test result is non-significant and if the specific variable X_k has any impact on our output variable. If the p-value is greater than the significance level α (the risk that we are willing to take for rejecting the null hypothesis when it is true), we fail to reject the null hypothesis (H_0: β_k = 0), which means that the coefficient of the variable X_k is 0. It does not have any impact on our output variable Y. This test can only be performed on only one variable. Therefore, we would need to perform it in each step where we exclude the variable that has the highest p-value. The process is concluded when we have a good enough predictive model (Hastie, Tibshirani & Friedman, 2008).

Once we created the predictive model, the estimated coefficients can be used together with other observed explanatory variables to predict our output variables $\hat{y}$. Then, we would need to measure how good our predicted variables $\hat{y}$ represent the actual values y. The measure for goodness of fit is called the coefficient of determination R^2. This measure is defined by the following equation:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y}_t)^2} \qquad (3.2)$$

where SSR represents the regression sum of squares and SST the total sum of squares. The total sum of squares can be also expressed as a sum of regression sum of squares SSR and sum of squares for the error SSE:

$$\begin{array}{ccccc} \sum_{t=1}^{n}(y_t - \bar{y}_t)^2 & = & \sum_{t=1}^{n}(\hat{y}_t - \bar{y}_t)^2 & + & \sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \\ \text{SST} & = & \text{SSR} & + & \text{SSE} \end{array} \qquad (3.3)$$

The coefficient of determination R^2 shows how much of the variation in the output variables is explained with our model (Rencher & Schaalje, 2008). However, we have to note that with each new variable added to the model, R^2 increases or stays the same but never decreases. Therefore, this measure should be used with caution while comparing models with a different number of variables. This is the reason why adjusted R^2 is more informative because it is adjusted by the degrees of freedom:

$$R_a^2 = 1 - \frac{SEE/(n-p)}{SST/(n-i)} = 1 - \frac{n-i}{n-p}(1 - R^2) \qquad (3.4)$$

where n is the number of observations, p is the number of free parameters in our model, and i is 1 if our model has an intercept and 0 otherwise. This measure can increase or decrease with the increase of the number of variables used in our model, depending on how much R^2 and $n - p$ change (Wu & Coggeshall, 2012).

5 MACHINE LEARNING MODELS

Machine learning (ML) which is currently considered a hot topic has various research breakthroughs and different applications in many fields. It has its basis from statistical learning theory. However, it includes also advanced characteristics and methods for building powerful algorithms. ML is concentrated on “teaching” the computer how it can learn specific tasks like recognizing characters, classify data in groups of various types, predicting diseases, and many more tasks. A lot of researchers coming from different areas of science use ML for tackling their daily problems and supporting specialists in their decision processes (Mello & Ponti, 2018). The main aim of ML is understanding the basic principles of learning as a computational process while using tools that come from Statistics and Computer Science while designing better-automated learning methods. As previously mentioned, ML designs algorithms that can learn specific rules from new data, adapt to changing conditions and improve its performance with experience (Blum, 2007).

The distinction between different types of ML is the same as with the statistical learning described in Section 6. Therefore, it is organized into two main types: supervised and non-supervised learning. Supervised learning is focused on finding the best way how the input variables converge to the specified output variables. Therefore, it could predict the outputs with unseen input variables with high accuracy. On the other hand, non-supervised learning is associated with creating models after analyzing the correlation between the input data (Mello & Ponti, 2018). The notation used is the same as with the statistical learning.

5.1 Extreme gradient boosting model

Machine learning models which can make inferences from large data sets are currently gaining huge popularity. Their success and popularity are derived from the statistical models that can capture complex interdependencies between the variables and the scalable learning systems that train the model from large data sets. Gradient tree boosting model is one of the machine learning models which has proved its efficiency in various machine learning and data mining challenges where it provided results for a wide range of problems from web text to high energy physics event classification (Chen & Guestrin, 2016).

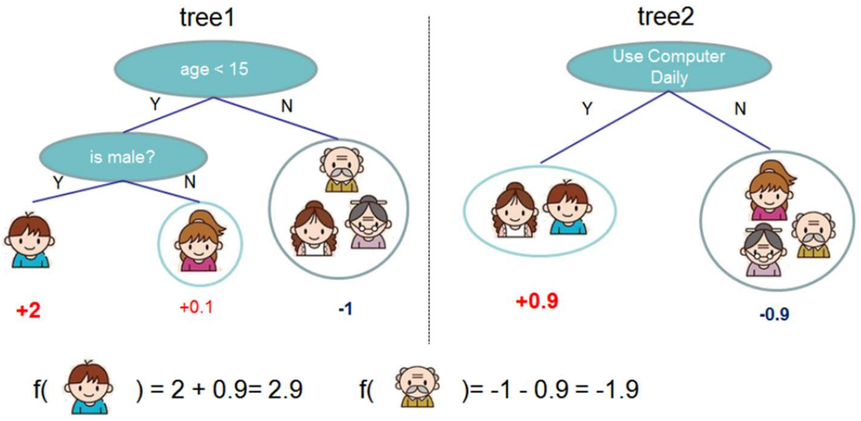
Extreme gradient boosting uses the gradient boosting framework to boost weak learners and the additional features as system optimization and algorithmic improvements makes it one of the better-performing algorithms for various statistical problems. The biggest advantage of this model is the regularization feature which is used for avoiding overfitting of the models which is shown in Equation 4.1 (Saraswat, 2017). This model performs a stage-wise additive process that begins with a weak learner that fits the data and continues with fitting additional weak learners without making any changes to the previous learner to improve the performance of the current model. In other words, we start by estimating $\hat{y}_1$ by fitting the data to one decision tree and continue with fitting the second tree based on the residuals from the previous step. The process is performed to decrease the model error efficiently (Budholiya, Shrivastava & Sharma, 2020). The process of predicting the electricity price with the XG boost model can be summarized as follows. The learning objective of the tree ensemble model for data set with n observation and m features

$D = \{(x_t, y_t)\}$ $(|D| = n, x_t \in \mathbb{R}^m, y_t \in \mathbb{R})$ which uses K additive functions to predict the electricity price $\hat{y}_t$ (which represents the predicted electricity price of the t-th instance at the k-th boost) is expressed in Equation (4.1)

$$\hat{y}_t = \phi(x_t) = \sum_{k=1}^{K} f_k\ (x_t), f_k \in F \tag{4.1}$$

where $F = \{f(x) = w_{q(x)}\}$ $(q : \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T)$ represents the space of regression trees known as Classification and Regression tree analysis (CART). q shows the structure of each tree which maps the example to the corresponding leaf index. T is the number of leaves in each tree. Each one of the functions f_k is connected to corresponding independent tree structure q and leaf weight w. The difference between decision trees, the regression trees contain a continuous score on each leaf. Therefore w_i is used to present the score on the i-th leaf. The decision rules in each tree (q) are used to classify them into leaves and summing up the leaves' scores (w), which means that the final prediction is actually the sum of the predictions from each tree. This can be observed in Figure 7.

Figure 7: Example of a tree ensemble model



Source: Chen & Guestrin (2016).

The learning of the functions in this model is achieved by minimizing the regularized objective expressed in Equation (4.2)

$$\mathcal{L}(\phi) = \sum_i \ell(\hat{y}_t, y_t) + \sum_k \Omega(f_k) \tag{4.2}$$

where $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda||w||^2$ and ℓ is a differentiable convex loss function which measures the difference between predicted $\hat{y}_t$ and observed y_t value. The term Ω is used to penalize the complexity of the regression tree function by smoothing the weights to avoid over-fitting. If the regularization parameter is zero, the objective function is basically the one from the traditional gradient tree boosting model.

The model expressed in Equation 4.2 contains functions as parameters that can not be optimized with the traditional optimization methods. Therefore, the training of the model is performed with an additive approach. If we denote $\hat{y}_t^{(i)}$as the i-th iteration of the prediction of the t-th observation, we would need to add f_i to minimize the objective function

$$L^{(i)} = \sum_{t=1}^{N} \ell(y_t, \hat{y}_t{}^{(i-1)} + f_i(x_t)\,) \;+\; \Omega(f_i\,), \tag{4.3}$$

which means that we greedily add the function f_i which could best improve the performance of our model in Equation 4.2. Afterward, the objective function is optimized with second-order approximation (Chen & Guestrin, 2016).

5.2 Hyperparameters in XG boost model

Since XG boost is a decision tree-based model, there is a certain number of hyperparameters that are used for improving its performance. For instance, max_depth and subsample are used for treating the overfitting problem while eta (the learning rate) is used for managing the weights of trees that are added into the model and also for reducing the adaptation rate of the model to the training data (Budholiya, Shrivastava & Sharma, 2020).

Table 2: XG boost hyperparameters

Parameter	Default	Description
nrounds	100	The maximum number of iterations
min_child_weight	1	The minimum sum of weights needed in a child
max_depth	6	The maximum depth of each tree
subsample	1	The proportion of each sample
colsample_bytree	1	Column's proportion of random samples
gamma	0	The minimum loss reduction for further partition on a leaf node of the tree
eta	0.3	Learning rate (Shrinking weights for each step)
lambda	0	Control for L2 regularization (Ridge regression) and used for avoiding overfitting
alpha	1	Control for L1 regularization (Lasso regression) and used for shrinking and feature selection

Source: Saraswat (2017).

Other hyperparameters are used for selecting the number of trees that would be fit, the proportion of the sample, or the minimum loss reduction (gamma). The hyperparameters lambda and alpha are also included in the regularization term Ω in the objective function

Equation 4.2 (Saraswat, 2017). The description and default values of all hyperparameters are presented in Table 2.

5.3 Hyperparameter tuning and cross-validation

Each hyperparameter has its specific role in the performance of the model. To find the adequate value of each hyperparameter, the process called hyperparameter tuning is executed. Usually, the first model is built with the default hyperparameters which could surprisingly provide impressively accurate results. If the accuracy provided of the first model is not acceptable, the next step would be to amend the eta parameter to 0.1 and with the other default hyperparameters using the k-fold cross-validation to choose the best n-rounds (Saraswat, 2017). It is the most widely used method for measuring the prediction error by splitting the training data set into K roughly equal-sized groups, fitting the model on one part of the data set, and using a different part for testing it. In this way, the efficiency of the machine learning model is estimated by testing it on a data set that was not used in the training process of the model. This procedure usually provides less biased and optimistic results compared to other methods like only train/test data set to split. Using this cross-validation procedure, the optimal hyperparameters would be chosen by looking at the model with the lowest calculated expected error. There are many discussions regarding the optimal value of k and the conclusion is that five- or tenfold cross-validation is recommended and usually used in practice (Hastie, Tibshirani & Friedman, 2008). In this thesis, the fivefold cross-validation procedure will be performed. Once the optimal value of n-rounds is chosen, if it is necessary, a grid search is performed on the other hyperparameters by fixing eta and n-rounds. Once the optimal hyperparameters are chosen, the final model is created and it can be optimized for further use.

6 AT/DE SPREAD FORECASTING USING STATISTICAL AND MACHINE LEARNING MODELS

For this thesis, two forecasting models will be created for predicting the AT/DE spread. The process will be reviewed and their results compared. The forecasting process presented in Section 2.4 will be used.

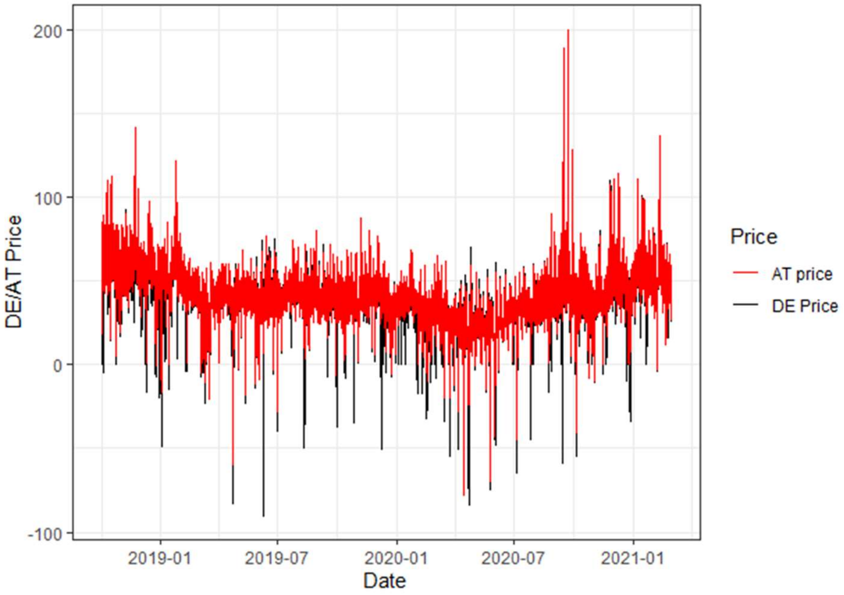
Table 3: Descriptive statistics of German and Austrian spot prices

	Max	Min	Mean	Median	Standard deviation	Skew	Correlation
EPEX Spot DE Price	200.04	-90.01	37.10	37.24	18.33	-0.39	0.897
EPEX Spot AT Price	200.04	-77.68	40.17	39.08	17.04	0.24	

Source: Own work.

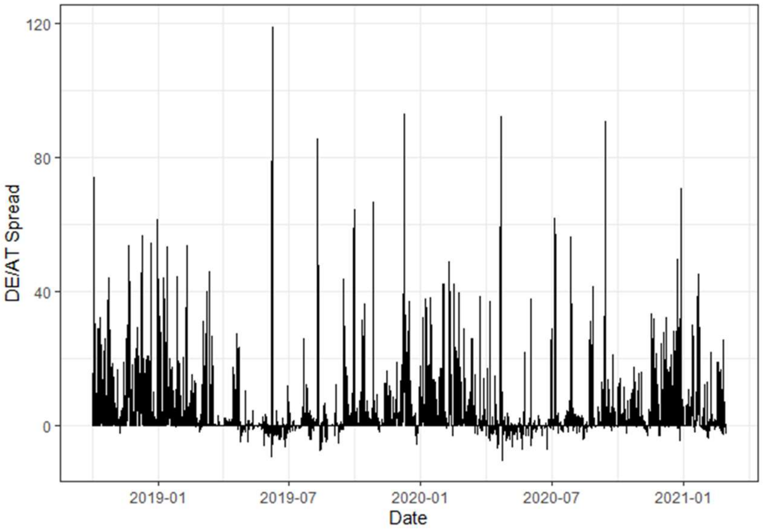
First, we set the goal to predict the outcome and the AT/DE electricity price spread, as best as we can. Then, we observed the dynamics of the German and Austrian electricity prices in the period 1/10/2018-28/02/2021 which is examined in this thesis. The correlation between these two prices is high (0.897) which is also visible from Figure 8 where we can see that both prices tend to move in the same direction

Figure 8: German and Austrian spot prices for period 1/10/2018-28/02/2021



Source: Own work.

Figure 9: German and Austrian spread for period 1/10/2018-28/02/2021



Source: Own work.

From Table 3, we can see that in the period observed, the average German price was 37.10 EUR and the average Austrian price was 40.17 EUR. The average Austrian price is higher due to the phenomenon of negative electricity prices discussed in Section 5, which is more common in Germany. Therefore, German prices have negative skewness (more probable negative prices) and a larger standard deviation than the Austrian prices. This characteristic of German prices is also visible in Figure 8 where the black line has lower negative values.

6.1 The data used in the forecasting models

The data used in this thesis is the electricity production in Germany and Austria from different sources, such as wind, solar photovoltaic (SPV), hydro, etc. The main data sources are Wattsight and the European Network of Transmission System Operators (ENTSO-E). The variables included in the analysis are shown in Table 4 along with their characteristics and their data source. The data source for the German and Austrian electricity spot prices is the European Power Exchange, EPEX. Some of the data have quarter-hourly granularity. However, it is averaged to hourly granularity to use it appropriately in the construction of the predictive models.

The residual load is a power system indicator where it shows the capacity left in the system for the operation of the conventional power plants. In other words, it shows how much electricity is produced after the subtraction of the electricity generation of the plants that have to run (inflexible conventional power plants) and the power plants with low or almost no marginal costs: wind, hydro, and solar. (Energypedia, 2018). On the other hand, the total load presents the total electricity supply produced from all available power plants.

Before constructing a predictive model, the data should be prepared for analysis. We should first check whether there are some missing values (N/As) and decide how we will treat them. From Table 4 we can notice that the number of the missing values is just 89 which is pretty small compared to the total number of data points. The total missing values in our data set are only 0.025% out of the total number of observations. Therefore, we will treat these missing values by just extracting all rows where we have missing values, which means excluding the 89 dates where we have incomplete data set of all variables. This results in extracting 1 513 data points from our total data set, which has 358 343 data points after the treatment.

Table 4: Number of missing values in the data set used in our analysis

Total	Missing values	% of Missing values in total	After treatment
359 856	89	0.0247%	358 343

Source: Own work.

The next step is to split the data set into two groups. One of them is known as the training dataset which consists usually of 50-80% of the entire dataset. With this dataset, we estimate the coefficients and the model tries to examine the relationship between the explanatory variables and the variable that we want to predict. In this thesis, the training data set contains data from 1/10/2018-1/9/2020 and consists of 16 824 observations (24 hourly observations for 701 days), which leaves us with 284 597 data points for all 17 variables. The second dataset test is used to assess the accuracy of the predictive model. The testing data set in this thesis consists of data from 1/9/2020-01/03/2021 and 4 344 observations (24 hourly observations for 181 days), which translates to a total number of 73 746 data points in the testing data set.

Table 5: Overview of the variables, their description and data source

Variable	**Unit**	**Granularity**	**Description**	**Data source**
Spot price AT	EUR/MWh	hour	Market clearing price on EPEX Day-ahead auction	EPEX
Spot price DE_Luxemburg	EUR/MWh	hour	Market clearing price on EPEX Day-ahead auction	EPEX
Total load DE	MWh	15 minutes	Total load Germany	ENTSO-E
Total load AT	MWh	15 minutes	Total load Austria	ENTSO-E
Residual load AT	MWh	hour	Residual load Austria	Wattsight
Residual load DE	MWh	hour	Residual load Germany	Wattsight
SPV Production AT	MWh	hour	Actual photovoltaic power production in Austria	Wattsight
Wind Onshore DE	MWh	15 minutes	Actual wind onshore production in Germany	ENTSO-E
Wind Offshore DE	MWh	15 minutes	Actual wind offshore production in Germany	ENTSO-E
Wind Onshore AT	MWh	15 minutes	Actual wind onshore production in Austria	ENTSO-E
Hard coal AT	MWh	15 minutes	Actual hard coal production in Austria	ENTSO-E
Hydro-Run of river AT	MWh	15 minutes	Actual hydro (run-of-river and poundage) production in Austria	ENTSO-E
Biomass AT	MWh	15 minutes	Actual biomass production in Austria	ENTSO-E
Gas AT	MWh	15 minutes	Actual gas production in Austria	ENTSO-E
Hydro-Pumped storage consumption	MWh	15 minutes	Actual hydro pumped storage consu mption in Austria	ENTSO-E
Hydro-Pumped storage production	MWh	15 minutes	Actual hydro pumped storage produc tion in Austria	ENTSO-E
Nuclear generation France	MWh	hour	Actual nuclear production in France	ENTSO-E

Source: Own work.

6.2 Steps for creating statistical and machine learning forecasting models

The first model in this thesis is a multiple linear regression created out of the 17 variables which were presented in Table 5. Out output is the variable *Spread*, the difference between *Spot price AT* and *Spot price DE* while variables *Wind Onshore DE* and *Wind Offshore DE* are summed up in one variable *Wind_DE*. Additionally, the variables *Month*, *Day,* and *Hour* are included to check whether the specific period of the year or month is an important determinant of the electricity price spread. The stepwise backward selection model shown in Section 2.4 would be used in our variable selection process. The results of the full linear regression model are presented in Table 6.

Table 6: Full linear regression model

Dependent variable: AT/DE spread					
Coefficient	**Estimate**	**Std. Error**	**t-value**	**Pr(>|t|)**	
Month	0.275200	0.015860	17.3500	<2.00E-16	***
Day	-0.019370	0.005841	-3.3160	0.000916	***
Hour	0.003104	0.007923	0.3920	0.695250	
Biomass_AT	0.025030	0.001753	14.2830	<2.00E-16	***
Gas_AT	0.000385	0.000113	3.3950	0.000689	***
Hard_coal_AT	0.005087	0.000543	9.3620	<2.00E-16	***
Hydro_Pumped_storage_consumption_AT	-0.000492	0.000141	-3.4820	0.000500	***
Hydro_Pumped_storage_ production_AT	0.000858	0.000107	7.9910	0.000000	***
Hydro_Run_of_river_AT	-0.000763	0.000078	-9.8050	<2.00E-16	***
Wind_Onshore_AT	-0.001606	0.000210	-7.6580	0.000000	***
Nuclear_generation_France	-0.000072	0.000014	-5.1190	0.000000	***
Total_load_AT	0.001144	0.000213	5.3700	0.000000	***
Total_load_DE	0.000017	0.000024	0.6860	0.492679	
SPV_Production_AT	0.000957	0.000484	1.9770	0.048075	*
Residual_Load_AT	-0.000884	0.000202	-4.3740	0.000012	***
Residual_Load_DE	-0.000191	0.000015	-12.7650	<2.00E-16	***
Wind_DE	0.000300	0.000014	21.6360	<2.00E-16	***
Signif. codes	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error	6.732 on 16724 degrees of freedom				
Multiple R-squared	0.4145				
Adjusted R-squared	0.4139				
p-value	< 2.2e-16				

Source: Own work.

The column *Estimate* shows the estimated coefficient $\beta_k$which shows the influence of variable k on the dependent variable *Spread*. As described in Section 6.2, the T-test whose results are shown in the last column (Pr(>|t|)) is used to test whether the β_k is zero, which means that the influence of the variable k on the output is insignificant. The interpretation of the results of the t-test shown in the last column is the following: if the p-value is higher than the significance level (α), we can assume that the coefficient β for that variable k is zero. With this in mind, we would exclude the variables with the highest p-values. The variable *Hour* with the p-value of 0.69525 is the first one that will be excluded from the model. Further variables which will be excluded in the stepwise variable selection are *Total load DE, SPV Production AT, Gas AT, Day, Hydro Pumped storage consumption AT, Residual Load AT,* and *Hydro Pumped storage production AT*. The results of the final reduced model are presented in Table 7.

Table 7: Final reduced linear regression model

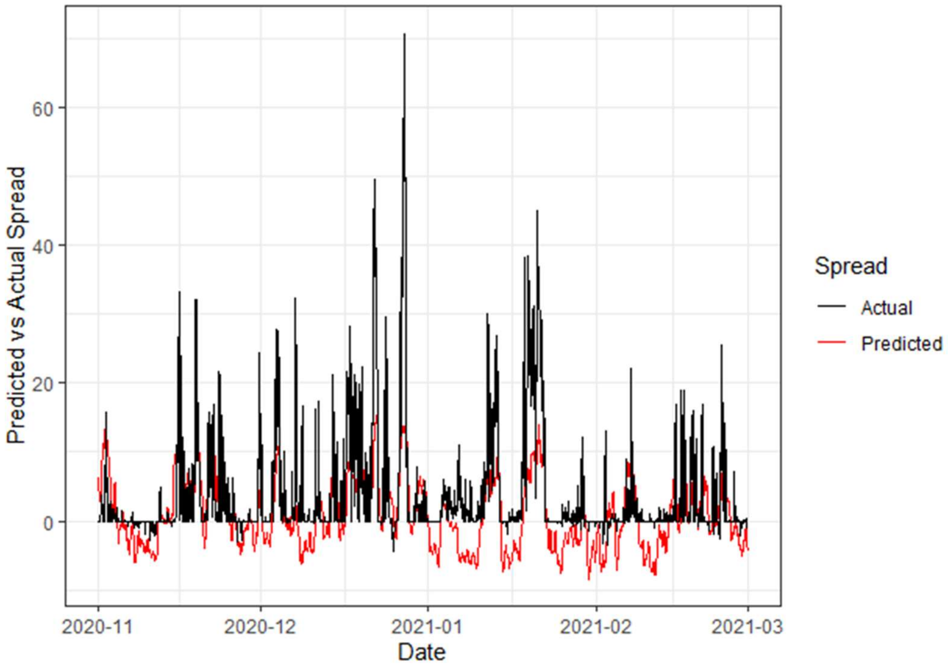
Dependent variable: AT/DE spread					
Coefficient	**Estimate**	**Std. Error**	**t-value**	**Pr(>|t|)**	
Month	0.2304000	0.0146900	15.68	<2e-16	***
Biomass_AT	0.0247400	0.0016690	14.822	<2e-16	***
Hard_coal_AT	0.0051140	0.0005351	9.557	<2e-16	***
Hydro_Run_of_river_AT	-0.0009796	0.0000607	-16.128	<2e-16	***
Wind_Onshore_AT	-0.0008796	0.0000671	-13.106	<2e-16	***
Nuclear_generation_France	-0.0001262	0.0000122	-10.371	<2e-16	***
Total_load_AT	0.0010490	0.0000596	17.592	<2e-16	***
Residual_Load_DE	-0.0001913	0.0000084	-22.748	<2e-16	***
Wind_DE	0.0002753	0.0000086	32.085	<2e-16	***
Signif. codes	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error	6.756 on 16732 degrees of freedom				
Multiple R-squared	0.41				
Adjusted R-squared	0.4097				
p-value	< 2.2e-16				

Source: Own work.

Looking at the goodness of fit of the reduced linear regression model, the coefficient R-squared was not improved a lot and it is 0.41, which means that only 41% of the variability of the output variable is explained from our model. This shows that this model does not have a very high level of explanatory power. Creating the reduced linear regression model with the variable selection process did not help in improving the predictive power of the linear regression model. Observing the estimated coefficients in the column *Estimate*, we can see their relationship with the output variable – AT/DE spread. The sign of the estimated coefficient shows what happens with the spread when that coefficient changes while other coefficients remain constant. Consequently, we can see that the increased Hydro and Wind electricity production in Austria decreases the price spread between AT and DE since it lowers the Austrian electricity price. This is a reasonable result since they do not affect the German price, therefore with a lower Austrian price the spread will decrease. The sign in front of the coefficient for German Residual load is also reasonable since the increased Residual load means that there is higher electricity generation from more expensive power plants. Therefore, the German price will go up and while everything else remains constant, the AT/DE Spread will decrease.

Finally, the comparison of the predicted spread using our estimated coefficients from the final linear regression model with the actual spread is presented in Figure 10. From the plot, it is also evident that the model does not provide precise predictions of the spread.

Figure 10: Predicted vs actual spread with the linear regression model



Source: Own work.

It is important to note, that I have additionally tried using the log transformed independent and dependent variables in my linear regression model, however, it did not provide any significantly different results.

The second model that will be created in this thesis to predict the AT/DE electricity price spread is the XG boost model presented in Section 7.1. The data set is identical as the one in the previous model including the same training and testing data sets and the same explanatory and output variables. The default hyperparameters in Table 2 will be used in creating the first XG boost predictive model. The results from the first XG boost model are presented in Table 8.

Table 8: XG boost model results with default hyperparameters

Feature	**Gain**	**Cover**	**Frequency**
Wind_DE	0.326998	0.141931	0.102738
Residual_Load_DE	0.178624	0.153030	0.095270
Hydro_Run_of_river_AT	0.085671	0.082262	0.087124

(table continues)

(continued)

Table 8: XG boost model results with default hyperparameters

Nuclear_generation_France	0.058307	0.092306	0.078525
Hard_coal_AT	0.053068	0.053797	0.048427
Day	0.050854	0.038564	0.058837
Month	0.050585	0.034958	0.054311
Gas_AT	0.034189	0.064404	0.080335
SPV_Production_AT	0.033886	0.025968	0.031681
Biomass_AT	0.024010	0.052791	0.046843
Total_load_AT	0.019951	0.020634	0.040959
Hydro_Pumped_storage_consumption_AT	0.019738	0.051191	0.040281
Hydro_Pumped_storage_production_AT	0.016707	0.032528	0.031455
Residual_Load_AT	0.015868	0.061928	0.044807
Wind_Onshore_AT	0.014016	0.029075	0.047296
Total_load_DE	0.010428	0.044882	0.042317
Hour	0.007100	0.019750	0.068794

Source: Own work.

Table 8 shows how each variable of the model complements and improves the model. The column *Gain* shows the improvement of the accuracy of the model by including the feature in the branches of the tree. *The cover* is the measure of the number of observations that are concerned by each feature and the column *Frequency* is just a simpler *Gain* measure because it shows how many times one feature is used in all trees that are generated (Xgboost, n.d.).

The features with a higher value of the measure *Gain* are more informative, which means that the main determinants of the AT/DE Spread in this model are German wind electricity production and German Residual load. The wind is one of the main electricity sources in Germany as evident in Figure 4. This means that the wind power plants availability is very important for the German power system because in the cases when wind electricity generation is insufficient, more expensive power plants have to be plugged in to maintain the system stability and to cover the electricity demand which leads to higher prices. Therefore, this shows the importance of Wind production and German Residual load as one of the main determinants of the AT/DE spread. Other important determinants of the AT/DE Spread are Run of River Hydro production in Austria because it is the main electricity source in Austria visible in Figure 5. The fourth important price determinant is the Nuclear generation in France which is quite important for the German power system because it is the main source of the imported electricity in Germany. In that way, the French nuclear generation also impacts the German price and the AT/DE Spread.

Table 9: Performance measures of the first XG boost model

RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
3.67255	0.80188	1.64166	0.28091	0.02696	0.05323

Source: Own work.

The performance measures of our first XG boost model are shown in Table 9. From the table, we can see that the R-squared is 0.801, which is much better than the first linear regression model. This means that the current XG boost model explains 80% of the variability of the AT/DE Spread.

The final XG boost model is created after the hyperparameter tuning explained in Section 7.2 using a fivefold cross-validation procedure. The final set of hyperparameters used are presented in Table 10.

Table 10: Hyperparameters used in the final XG boost model

Hyperparameter	Value
nrounds	15 000
max_depth	5
eta	0.1
gamma	1
colsample_bytree	0.8
min_child_weight	5
subsample	0.75

Source: Own work.

The variable selection in this model is performed by excluding the variables that have a low *Gain* estimate which then improves the performance of the model by lowering the error and increasing the R-squared. In our model, there are no redundant variables with a low *Gain* estimate that substantially improve the performance of the model once they are excluded from the model. That means that the XG boost model can use all the information from the given variables and still give the highest importance only to the ones who are more important for predicting the outcome. The results of the final XG boost are presented in the following tables:

Table 11 includes the measures of the importance of the variables included in the final XG boost model. The most important price determinants stay the same as the previous model, German wind electricity generation, Residual load Germany, Austrian Hydro production, and French Nuclear electricity generation. However, from Table 12, we can observe the improvement of the model since the R-squared increased to 0.836 and the RMSE decreased to 3.34. This means that with the hyperparameter tuning we managed to slightly improve the performance of the final XG boost model.

Table 11: XG boost model results with tuned hyperparameters

Feature	Gain	Cover	Frequency
Wind_DE	0.283992	0.140881	0.112356
Residual_Load_DE	0.180656	0.129601	0.100657
Hydro_Run_of_river_AT	0.091501	0.102554	0.092402
Nuclear_generation_France	0.058398	0.086640	0.081716
Day	0.051523	0.024237	0.036995
Month	0.048384	0.019589	0.020499
Hard_coal_AT	0.047282	0.043639	0.046481
Biomass_AT	0.036555	0.032081	0.036652
Total_load_AT	0.031848	0.051219	0.060905
Gas_AT	0.031257	0.068791	0.071217
Hydro_Pumped_storage_consumption_AT	0.027257	0.038985	0.039176
Total_load_DE	0.026046	0.052206	0.063351
Residual_Load_AT	0.021727	0.066772	0.067681
SPV_Production_AT	0.020394	0.031754	0.034581
Wind_Onshore_AT	0.017112	0.049669	0.061637
Hydro_Pumped_storage_production_AT	0.015491	0.041543	0.041512
Hour	0.010578	0.019839	0.032182

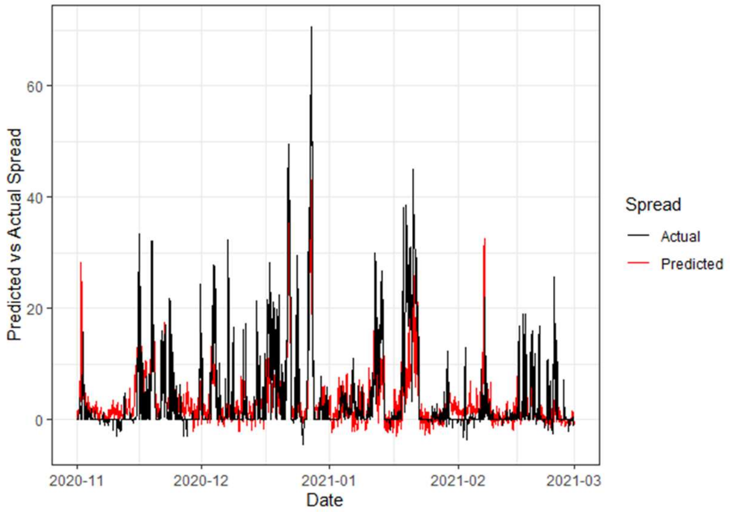
Source: Own work.

Table 12: Performance measures of the final XG boost model

RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
3.34018	0.83670	1.56710	0.06392	0.00595	0.02284

Source: Own work.

Figure 11: Predicted vs actual spread with the XG boost model



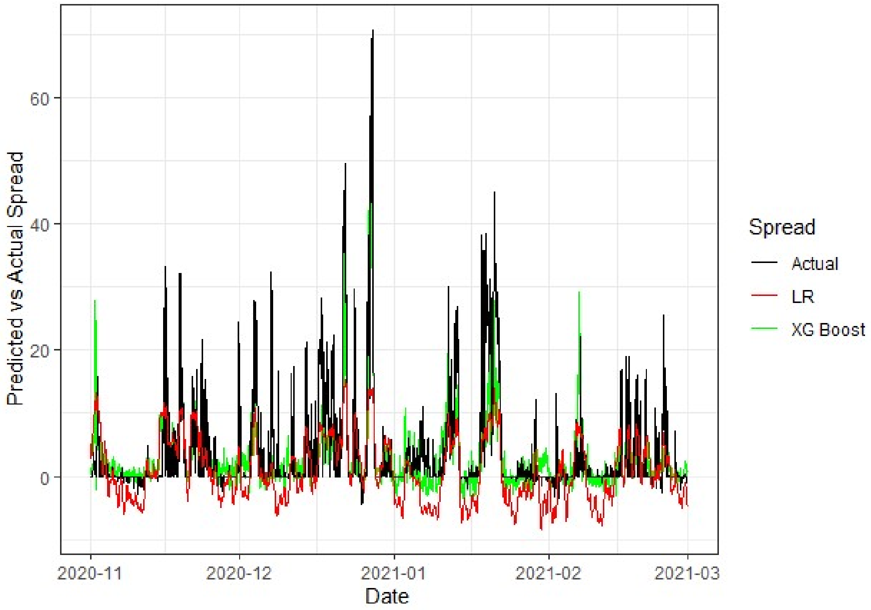
Source: Own work.

Finally, Figure 11 shows the comparison between the predicted spread using the final XG boost model with the actual spread. From the plot observation, we can see the improvement of the predicted spread comparing to the linear regression model. We can conclude that the XG boost model is more powerful and it can extract more information from the same data set. This is a very important conclusion because this model can be used for various other predictions because of its great performance.

6.3 Comparison of the forecasting models

From Figure 12, it is evident that the XG boost predicts the AT/DE spread more efficiently. Multiple linear regression is not as precise and it predicts more negative spreads. The negative spread would happen in the case where German electricity price is higher than the Austrian. However, this is not very common since as seen in Figure 4 the wind electricity production is the main renewable source of electricity in Germany and it was nearly 20% of the total electricity production in 2019. Wind electricity production is cheaper. Therefore, most of the time, the German electricity price would be lower than the Austrian. The desired increase of the electricity production from renewable sources in Germany was successful and was providing great results in 2020 when the wind power plants produced more electricity than the fossil fuels plants (Wettengel, 2021). This trend will certainly provide even lower but more volatile prices.

Figure 12: Linear regression and the XG boost model results



Source: Own work.

The performance measures presented in Section 4.2 were calculated for both predictive models and shown in Table 13. By looking at the results, we can have the same conclusion that the XG boost model has a lower error and that it provides more efficient predictions. The Mean error measure shows us the average of all the residuals calculated as a difference between the actual AT/DE spread in our testing set and the predicted spread from the models. The XG boost model has a lower mean error. However, this measure is not very informative since the positive and negative residuals can be canceled out and lower error will be shown which could present misleading accuracy of the predictive models. Therefore, other measures are used additionally where the absolute value of the residuals is considered. The estimated MAE also shows that the XG boost provides better results. The final performance measure is RMSE, the square root of the squared residuals, and it is the most reliable and frequently used measure for the estimation of the efficiency of predictive models. The linear regression has an RMSE of 6.51 while XG boost shows two times lower RMSE (3.34) which again confirms our conclusion that the latter provides better results.

Table 13: Performance measures of both predictive models

Model	**ME**	**MAE**	**RMSE**
Linear regression	2.049	4.197	6.510
XG boost	0.700	1.567	3.340

Source: Own work.

The linear regression model examines the linear dependencies between the dependent and independent variables. Furthermore, the implementation is straightforward as it is the interpretation of the results. However, due to its limitation of only looking at the linear dependencies and its sensitivity to outliers, it might not be able to make very accurate predictions. The XG boost model outperforms the Linear regression model in predicting the AT/DE electricity price spread. It can analyze non-linear correlations between the input and output variables. The boosting process reduces the variance and bias in the machine learning model while also improving the model predictions. With this process, the weak learners from one boosting tree are corrected by the previous tree and then transformed into strong learners. The XG boost model also contains a regularization parameter to avoid overfitting and it can also work quite well with outliers in the data set. The main drawback of this model is that it is relatively slow to implement and its efficiency and speed are highly dependent on the computing power of the machine used for training the model. In addition, the machine learning algorithm has many hyperparameters that must be set up, using the process hyperparameter tuning, to achieve better predictions which makes the training of the model more complicated. Despite its disadvantages, it is worth noting that its high predictive power makes it especially useful for predictions in many different fields.

6.4 Analysis of main electricity price determinants

Linear regression predicts a lot of negative spreads which means that German price is higher than Austrian which rarely happens. Therefore, I conducted additional analysis to find out what are the most important price determinants when the linear regression predicts positive and negative spreads. For this analysis, I split the test data set into two groups: the one where linear regression predicts positive spread and another where it predicts a negative spread, positive, and negative group respectively. The main differences between the two groups are presented in Table 14. From the table, it can be observed that the Wind generation is higher and has a higher standard deviation in the positive group. Additionally, the French Nuclear generation is higher and Hard Coal production in Austria is zero in the negative group.

Table 14: Summary statistics of variables in the positive and negative group

	Negative group	**Positive group**	**Comment**
Hard Coal AT	0	0 and >0	0 for the negative group
Max Wind DE	26 076	46 064	Overall higher Wind generation in Germany for positive group
SD Wind DE	4 859	10 454	Higher volatility of Wind generation in Germany for positive group
Min Res Load DE	28 633	2 747	Overall higher Residual load for negative group
Min Nuclear Generation France	37 084	19 964	Overall higher French Nuclear generation in the negative group
SD Biomass AT	4.9	46.3	Higher volatility of Biomass power generation in Austria for positive group

Source: Own work.

Moreover, I have checked the main price determinants in the two groups by training and testing again the two models. From the output of our XG boost model, I was examining the feature that has the highest Gain parameter. Furthermore, I calculated the correlation of that specific variable with the predicted spread from the linear regression. The results are presented in Table 15.

Table 15: Main price determinants for positive and negative group

	Negative group	
Feature	**XG boost – Gain**	**LR - Correlation**
Nuclear generation France	0.14	0.8941
	Positive group	
Feature	**XG boost – Gain**	**LR - Correlation**
Wind generation Germany	0.348	0.8151

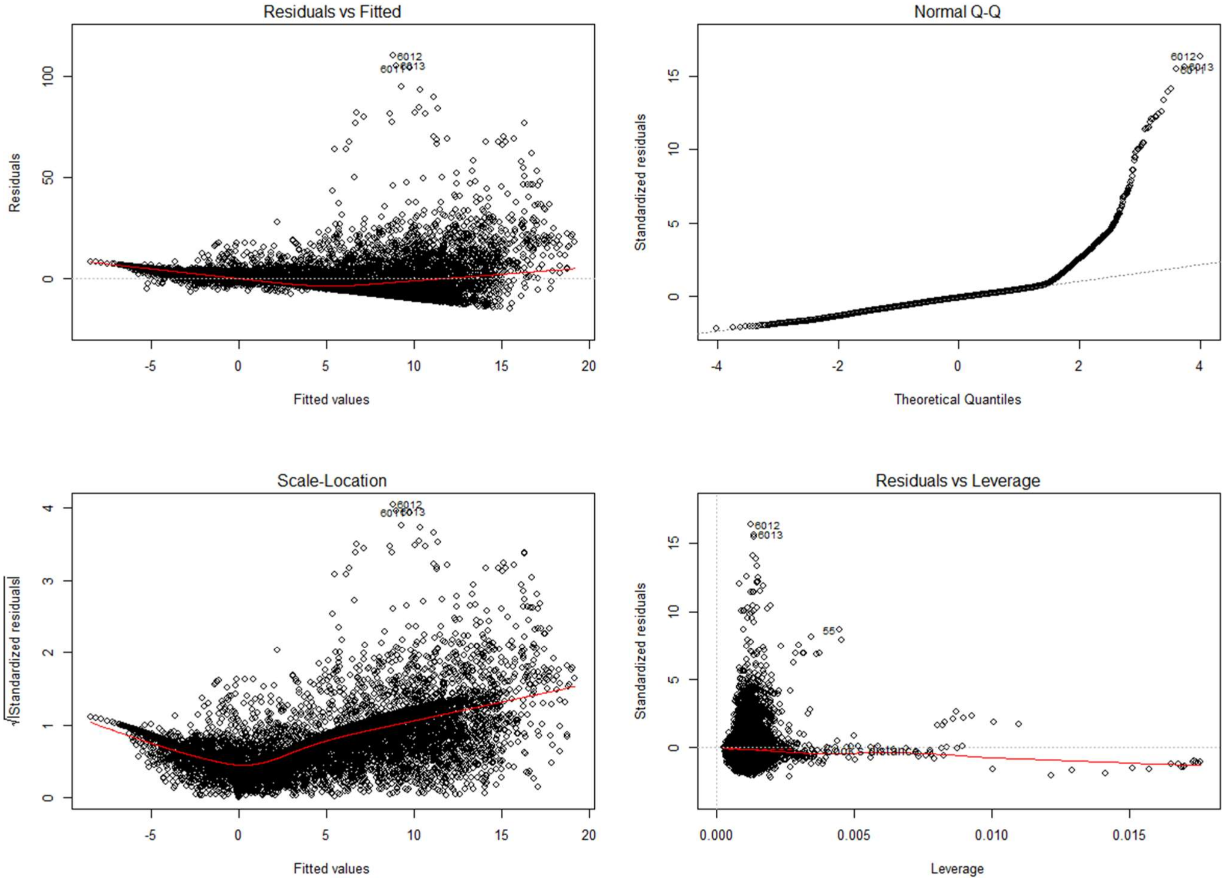
Source: Own work.

In the results, we can observe that the French Nuclear generation greatly influences the prediction in the negative group. The correlation between Nuclear generation and the predicted values is quite high, 0.8941 which justifies my conclusion. On the other hand, the main determinant in the positive group is the Wind generation in Germany where the correlation with the predicted values is also high and 0.8151. This concludes that the two main determinants of the variability of the AT/DE electricity price spread are French nuclear generation and German wind generation which was also seen in the previous results in Section 8.2.

6.5 Robustness checks

Because the linear regression model is appropriate only if certain assumptions are fulfilled, robustness checks are necessary. Therefore, I used some descriptive tools and plots to test whether the necessary assumptions are correct.

Figure 13: Diagnostic plots for the linear regression model



Source: Own work.

The following assumptions were reviewed in Figure 13:

- Linearity – The first plot shows whether there is a pattern in the residuals which would imply a non-linear relationship between the dependent and independent variables. In our case, no clear pattern is visible. Therefore, we can assume a linear relationship. However, some points are way farther than the horizontal line on the plot, which is an indicator for further check for the presence of outliers in the data set.

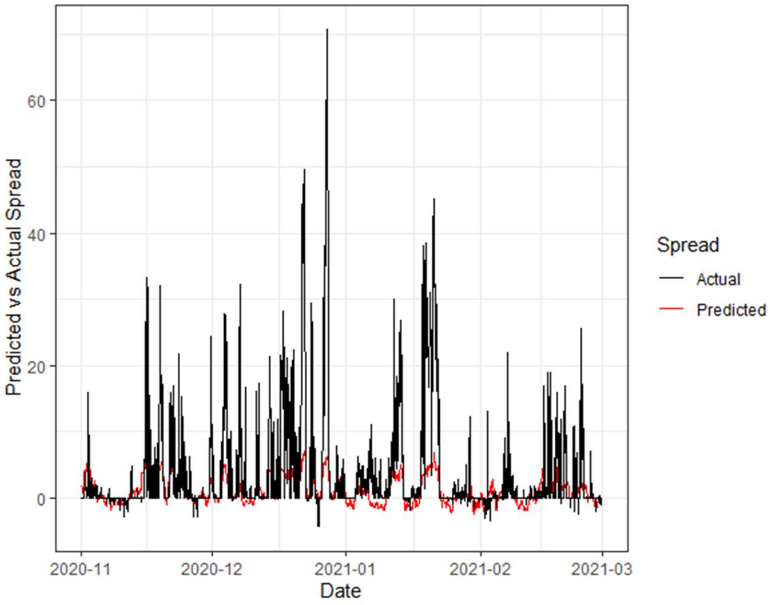
- Independent and normal residuals – The second and fourth plots should be analyzed to check the normality of residuals. The normal Q-Q plot shows a straight line in the case of normal residuals. From the plots, it is evident that the residuals are not normally distributed since they have more heavy-tailed distribution.

- Homogenic variance – The third plot shows whether the residuals are equally dispersed across the ranges of predicted values. When the variance is homogenous, the plot will show a horizontal line with equally spread points. In our case, we can observe heteroscedasticity.

Linear regression provides valid results only when the assumptions are valid. Otherwise, its results might be misleading. That makes linear regression inappropriate for AT/DE spread forecasting without proper outliers treatment.

Robust linear regressions were introduced in order to provide results that will not be affected if the linear regression assumptions are violated (STHDA, 2018). For the analysis that follows, I used the robust linear regression introduced by Huber in 1964. This robust linear regression uses M-estimation which stands for "maximum likelihood type" and it is robust to outliers in the dependent variable. It is, however, not appropriate for outliers in the explanatory variables.

Figure 14: Predicted vs actual spread with the robust linear regression model



Source: Own work.

The predicted values with the robust linear regression are presented in Figure 14. From the plot, it is evident that this robust model provides better results since there are few predicted negative spreads. Because this robust linear regression uses the maximum likelihood estimation, the performance measure R squared is not appropriate for examining the goodness of fit of the model predictions. For this reason, residual standard error (RSE) will be used to compare both models since it is a measure of the standard deviation of the residuals (Wikipedia, 2021).

Table 16: Residual standard error

	Multiple linear regression OLS	**Robust linear regression MLE**
RSE	6.7318	1.7006

Source: Own work.

The residual standard errors are presented in Table 16. In the table, we can see that the robust linear regression has a lower RSE. With these results, we can conclude that the robust linear regression provides better predictions of AT/DE price spread than the multiple linear regression.

Although the robust linear regression provides better results than the multiple linear regression, it still does not overperform the XG boost model. XG boost has better predictions even without any outliers treatments. Therefore, I can conclude that it is more appropriate model for forecasting the AT/DE electricity price spread.

7 CONCLUSION

The main focus of this thesis was finding an appropriate predictive model for the AT/DE electricity spread by comparing the effectiveness and the modeling process between two advanced predictive models. First of all, I have discussed the characteristics of electricity that make it a special kind of commodity in detail. Therefore, special regulations were implemented in the electricity market for ensuring stability in the system. Furthermore, the importance of electricity forecasting was highlighted by reviewing the research done so far while focusing on several forecasting models used for electricity price forecasting. Finally, two models for forecasting the AT/DE electricity spread were described and the results were reviewed.

Based on the analysis in this thesis, it can be concluded that the highly volatile electricity price is extremely difficult to predict due to its special characteristics. Additionally, the results indicate that the main determinants of the AT/DE spread are German wind electricity generation, Austrian hydro electricity generation, German residual load, and French nuclear generation. Wind is the main renewable source of electricity in Germany, same as Hydro generation for Austria. Therefore, these are the main determinants of electricity prices. The

German residual load shows how much of the electricity is generated from the non-renewable and conventional power plants. When the Residual load increases, the German electricity price goes up and the AT/DE spread decreases. One of the main determinants of the AT/DE spread is the French nuclear generation as well. In the case of peak demand in Germany, the main foreign electricity provider is France, which has nuclear generation as the main electricity source, hence making it an important determinant of the German electricity price.

For this thesis, two forecasting models were examined: multiple linear regression and the XG boost model. I have provided a theoretical overview for both models and an overview of the modeling process. The results from the comparison indicate that the machine learning model provides more efficient predictions compared to the linear regression model. The performance measures in Section 8.3 just confirm this conclusion, which is evident from the substantial decrease of the error measures for the XG boost model. The main disadvantage of linear regression is the fact that it looks into the linear relationship between the independent and dependent variables and it is very sensitive to outliers. Looking at the electricity price, which has many unexpected spikes, we can see that linear regression is not quite effective in forecasting the electricity price. On the other hand, the XG boost model is efficient in finding non-linear dependencies between the input and output variables. However, it has a more complicated setup and multiple hyperparameters have to be adjusted to achieve better predictions. Another disadvantage of this model is that its performance is dependent on the computing power of the computer used for training the model and, therefore, the implementation can be slow. Even though it has some disadvantages, XG boost provides great results for a wide range of real-world statistical problems.

Robustness tests were conducted in Section 8.5, which showed that the assumptions for multiple linear regression are not fulfilled. Therefore, a robust linear regression was used for further analysis. This model does not rely on the assumptions of linear regression and it is effective in treating outliers in the dependent variable which is why it provided better results than the multiple linear regression model. However, even with the improvements, the XG boost model still overperforms, which again confirms its high predictive power.

The XG boost predictive model designed for the Austrian/German electricity price spread, can be further modified and used for creating the Austrian HPFC, or predicting the AT/DE PTRs costs, all of which is important for electricity trading companies, regulators, or market operators in electricity markets for their decision processes.

The results obtained from this thesis, go in line with the research done so far, by pointing to the conclusion that the machine learning models are powerful and can provide better inferences from the given data sets. Together with the constant technological improvements and increasing computer power, we can expect further developments in the machine learning area, which can provide even more powerful predictive models that can be useful for various studies from different fields. This highlights the need for further research in this field to develop new advanced forecasting models and assess their performance, strengths, and limitations to be able to provide new insights into electricity price forecasting.

REFERENCE LIST

1. Amjady, N., & Hemmati, M. (2006). Energy price forecasting - problems and proposals for such predictions. *IEEE Power and Energy Magazine, 4(2), 20–29.* https://doi.org/10.1109/mpae.2006.1597990

2. Blum, A. (2007). Machine learning theory. (No. 26). *Carnegie Melon University, School of Computer Science.* http://www.cs.cmu.edu/afs/cs/user/avrim/www/Talks/mlt.pdf

3. Budholiya, K., Shrivastava, S. K., & Sharma, V. (2020). An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2020.10.013

4. Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.* https://doi.org/10.1145/2939672.2939785

5. Energypedia. (2018). *Residual load*. Retrieved March 8, 2021 from https://energypedia.info/wiki/Residual_Load#:%7E:text=Residual%20load%20is%20an%20indicator%20in%20a%20power%20system.&text=A%20common%20definition%20for%20residual,%2C%20solar%20and%20hydro)%22

6. Epex Spot. (2020). *Basics of the power market.* Retrieved November 11, 2020 from https://www.epexspot.com/en/basicspowermarket

7. Gürtler, M., & Paulsen, T. (2018). Forecasting performance of time series models on electricity spot markets: a quasi-meta-analysis. *International Journal of Energy Sector Management, 12(1)*, 103–129. https://doi.org/10.1108/ijesm-06-2017-0004

8. Harasheh, M. (2016). Forecasting Wholesale electricity prices with Artificial intelligence models: The Italian case. *Preprints.Org, 1–20.* https://doi.org/10.20944/preprints201607.0001.v1

9. Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* New York: Springer.

10. Huisman, R., Stradnic, V., & Westgaard, S. (2013). Renewable energy and electricity prices: Indirect empirical evidence from hydro power. *Institut d'Economia de Barcelona (IEB), 1–16.* http://ieb.ub.edu/wp-content/uploads/2018/04/2013-IEB-WorkingPaper-24.pdf

11. Joskow, P. L., & Tirole, J. (2000). Transmission rights and market power on electric power networks. *The RAND Journal of Economics, 31(3), 450–487.* https://doi.org/10.2307/2600996

12. Kaminski, V. (2013). *Energy markets*. London: Risk Books.

13. Kath, C., & Ziel, F. (2018). The value of forecasts: Quantifying the economic gains of accurate quarter-hourly electricity price forecasts. *Energy Economics, 76, 411–423.* https://doi.org/10.1016/j.eneco.2018.10.005

14. Lago, J., De Ridder, F., & De Schutter, B. (2018). Erratum to "Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms" *Applied Energy*, *229*, 1286. https://doi.org/10.1016/j.apeenergy.2018.06.131

15. Mello, F. R., & Ponti, A. M. (2018). Machine learning: A practical approach on the statistical learning theory. *Springer International Publishing AG, part of Springer Nature 2018.* https://doi.org/10.1007/978-3-319-94989-5

16. Montel news. (2020). *Market data*. Retrieved November 12, 2020 from https://www.montelnews.com/marketdata

17. Next Kraftwerke. (2019, November 12). *Market coupling*. Retrieved November 20, 2020 from https://www.next-kraftwerke.com/knowledge/market-coupling

18. Nickolas, S. (2020, April 22). *Open interest vs. Volume: What's the difference?* Investopedia. Retrieved November 27, 2020 from https://www.investopedia.com/ask/answers/050615/what-difference-between-open-interest-and-volume.asp#:%7E:text=Volume%20and%20open%20interest%20are,are%20active%2C%20or%20not%20settled.

19. Pflugmann, F., Ritzenhofen, I., Stockhausen, F., & Vahlenkamp, T. (2019, November 21). Germany's energy transition at a crossroads. *McKinsey & Company.* Retrieved November 25, 2020 from https://www.mckinsey.com/industries/electric-power-and-natural-gas/our-insights/germanys-energy-transition-at-a-crossroads#

20. Rencher, A. C., & Schaalje, B. G. (2008). *Linear models in statistics* (2nd ed.). Hoboken: Wiley-Interscience.

21. Ritchie, H. (2020, July 10). *Austria: Energy country profile*. Our World in Data. Retrieved November 25, 2020 from https://ourworldindata.org/energy/country/austria?country=AUT%7EDEU

22. Sætherø, A. S. (2017, November). *Hourly Price Forward Curves for Electricity Markets. Construction, Dynamics and Stochastics (Doctoral thesis).* Universitat Duisburg- Essen. https://duepublico2.unidue.de/servlets/MCRFileNodeServlet/duepublico_derivate_00044584/DissASSaethroe.pdf

23. Saraswat, M. (2017, April 13). *Beginners tutorial on XGBoost and parameter tuning in r tutorials & notes | machine learning*. HackerEarth. Retrieved February 26, 2021 from https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial/

24. Schapire, R. E., & Freund, Y. (2012). *Boosting: Foundations and algorithms (adaptive computation and machine learning series).* Cambridge: The MIT Press.

25. Shahidehpour M., Yamin H. & Li Z. (2002). *Market operations in electric power systems: forecasting, scheduling, and risk management*. Hoboken: Wiley-IEEE Press.

26. SMARD (2018, October 1). *Germany and Austria introduce congestion management.* Retrieved November 24, 2020 from https://www.smard.de/page/en/topic-article/5892/9846

27. Stephenson, P., & Paun, M. (2001). Electricity market trading. *Power Engineering Journal, 15(6), 277–288.*

28. STHDA (2018). *Linear regression assumptions and diagnostics in R: Essentials*. (2018, March 11). Articles - STHDA. Retrieved July 12, 2021 from http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/

29. The Economist (2006, February 16). *The politics of power*. Retrieved September 12, 2020 from https://www.economist.com/special-report/2006/02/09/the-politics-of-power

30. Thomas, S. (2004, September). *Electricity liberalization: The beginning of the end.* Public Services International Research Unit (PSIRU), Business School, University of Greenwich. https://core.ac.uk/download/pdf/67063.pdf?repositoryId=51

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

32. Wettengel, J. (2021, January 5). Renewables produce more power than fossil fuels in Germany for the first time. *Clean Energy Wire.* Retrieved March 19, 2021 from https://www.cleanenergywire.org/news/renewables-produce-more-power-fossil-

fuels-germany-first-time#:%7E:text=According%20to%20the%20data%2C%20wind,%E2%80%9D%2C%20the%20think%20tank%20stated

33. Wikipedia (2016, June 16) *Timeline of machine learning*. Retrieved on October 26, 2020 from https://en.wikipedia.org/wiki/Timeline_of_machine_learning

34. Wikipedia (2021, April 29). *Robust Regression*. Retrieved July 16, 2021 from https://en.wikipedia.org/wiki/Robust_regression

35. Wu, J., & Coggeshall, S. (2012). *Foundations of Predictive Analytics (Chapman & Hall/CRC Data Mining and Knowledge Discovery) (1st ed.).* Boca Raton: Chapman and Hall

36. Xgboost (n.d.). *Understand your dataset with XGBoost.* Retrieved March 03, 2021 from https://xgboost.readthedocs.io/en/latest/R-package/discoverYourData.html

APPENDICES

Appendix 1: Povzetek (Summary in Slovene)

Glavna naloga magistrskega dela je bila iskanje primernega napovednega modela za ceno električne energije s primerjavo učinkovitosti in procesa modeliranja med dvema naprednima napovednima modeloma. Na podlagi analize v tej nalogi je mogoče sklepati, da je ceno električne energije zaradi njenih posebnih lastnosti zelo težko napovedati. Za namen te naloge sta bila preučena dva modela napovedovanja: večkratna linearna regresija in model XG boost. Pokazala sem teoretični pregled obeh modelov in pregled procesa modeliranja. Rezultati primerjave kažejo, da model strojnega učenja ponuja učinkovitejše napovedi v primerjavi z linearnim regresijskim modelom. Glavna pomanjkljivost linearne regresije je to, da preučuje linearno razmerje med neodvisnimi in odvisnimi spremenljivkami in je zelo občutljiva na izstopajoče vrednosti. Če pogledamo ceno električne energije, ki ima veliko nepričakovanih skokov, lahko ugotovimo, da linearna regresija ni povsem učinkovita pri napovedovanju cene električne energije. Po drugi strani je model XG boost učinkovit pri iskanju nelinearnih odvisnosti med neodvisnimi in odvisnimi spremenljivkami. Poleg tega je model bolj kompliciran in za doseganje boljših napovedi je treba prilagoditi več hiperparametrov. Druga pomanjkljivost tega modela je ta, da je njegova zmogljivost odvisna od računalniške moči računalnika, ki se uporablja za usposabljanje modela, zato je lahko izvedba počasna. Čeprav ima nekaj pomanjkljivosti, XG boost zagotavlja odlične rezultate za širok spekter resničnih statističnih problemov.

Izvedeni so bili testi robustnosti, ki so pokazali, da predpostavke za večkratno linearno regresijo niso izpolnjene. Zato je bila za nadaljnjo analizo uporabljena robustna linearna regresija. Ta model se ne opira na predpostavke linearne regresije in je učinkovit pri obravnavi odstopanj v odvisni spremenljivki, zato je zagotovil boljše rezultate kot model večkratne linearne regresije. Toda tudi z izboljšavami model XG boost še vedno bolje deluje, kar znova potrjuje njegovo visoko napovedno moč.

Rezultati, pridobljeni s to analizo, se ujemajo z dosedanjo raziskavo in kažejo, da so modeli strojnega učenja močni in lahko dajejo boljše napovedi iz danih podatkovnih nizov. Skupaj s nenehnimi tehnološkimi izboljšavami in naraščajočo računalniško močjo lahko pričakujemo nadaljnji razvoj na področju strojnega učenja, ki lahko zagotovi še močnejše napovedne modele, ki so lahko uporabni za različne študije z različnih področij.