COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO FIN DE GRADO:

Modelo de previsión de precios en mercados de electricidad de corto plazo

Short-term Forecasting of Electricity Market Prices

Autor: Pía Reneses Rodríguez

Director: María del Mar de los Ríos Calzada

Andrés Felipe Vargas Arboleda

Madrid

Junio de 2023

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

"Modelo de previsión de precios en mercados de electricidad de corto plazo"

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2022/23 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: Pía Reneses Rodríguez Fecha: 12/06/2023

Autorizada la entrega del proyecto

LOS DIRECTORES DEL PROYECTO

Fdo.: Andrés Felipe Vargas Arboleda Fecha: 12/06/2023

Fdo.: María del Mar de los Ríos Calzada Fecha: 12/06/2023

Acknowledgments

I would like to thank all the people from IGNIS that made this project possible. I would also like to thank my family: my dad, who said many years ago that if I ever wanted to truly understand his job, I would have to become an Electric Engineer, finish my master's degree, and read his thesis. Can't say I ever back down from a challenge. My mom, that presented her thesis when I was fourteen and has always made me strive to make her as proud of me as I am of her. My siblings, for the ever-present joy they bring to my life. And to the friends that encouraged me when life got a bit too much.

PREVISIÓN DE PRECIOS EN MERCADOS DE ELECTRICIDAD DE CORTO PLAZO

Autora: Reneses Rodríguez, Pía.

Directores: Ríos Calzada, María del Mar de los; Vargas Arboleda, Andrés Felipe.

Entidad Colaboradora: IGNIS ENERGÍA, S.L.

RESUMEN DEL PROYECTO

En las últimas décadas, la necesidad de pronosticar el precio diario de la electricidad no ha hecho más que aumentar. Hay diferentes beneficios que derivan de buenas predicciones. Estos incluyen, entre otros, mayores márgenes de beneficio para los distintos agentes del mercado y favorecer la integración de fuentes de energía renovables en el mercado eléctrico español (Querol Herrá, Javier, 2019).

Pronosticar los precios de la electricidad es un proceso muy complejo que no tiene una única solución. A lo largo de estas décadas desde que el mercado eléctrico dejó de ser un monopolio, se han investigado y probado muchas técnicas y métodos con este fin en mente. Los modelos ARIMA, Regresión Multi Linear e IA son ejemplos de modelos que se utilizan frecuentemente. Todos ellos dependen de la estacionariedad y normalidad para pronosticar con precisión. En los últimos años, esto es menos realidad con respecto a los precios de la electricidad en España, ya que los comportamientos del mercado dependen de más factores (Electricity in Spain, 2023)

Este proyecto se centrará en la previsión utilizando herramientas estadísticas y no herramientas impulsadas por IA para las predicciones en sí mismas y en el uso de algoritmos de aprendizaje automático para crear un recomendador/selector de modelos.

El proyecto consiste en estudiar y comparar diferentes métodos de modelización y previsión. El primer objetivo ha sido crear dos modelos principales: uno que incluyera términos de estacionalidad y/o autorregresivos y un modelo de regresión.

La regresión modelada fue una regresión multivariante. Las variables añadidas al modelo se eligieron de una lista de variables dentro de las siguientes categorías:

- Precio del gas y del CO2: valores oficiales diarios del precio de cierre del CO2 - solo se dispone de un valor- y del gas -donde hay tres variables posibles y ninguna de ellas es el precio real. Las tres variables se pueden consideran una aproximación muy precisa del precio spot real del gas.
- Precios anteriores del mercado diario de electricidad: información histórica de los precios marginales de electricidad
- Predicciones de generación para fuentes de energía renovables: energía eólica, fotovoltaica (solar FV), energía solar térmica e hidráulica.

- Tecnología de cierre o marginal del precio horario: se refiere a la energía más cara que se vende en el mercado a una hora determinada, que es la responsable del precio de toda la electricidad en ese momento.
- Prediction for hourly energy demand: information that is published daily by *Red Eléctrica.*

Se comprueba la multicolinealidad de las variables utilizando el VIF (Factor de Inflación de la Varianza) como medida. Los conjuntos de variables con valores de VIF lo suficientemente pequeños (< 5) se estudian como modelos. Se comprueba el p-valor de cada una de las variables y se eliminan de los diferentes modelos aquellas que tuvieran p > 0,05. Después de este proceso quedan siete posibles modelos:

Variables/Options	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Gas Price OMIE							
Reference Gas Price							
Auction Gas Price							
CO2 Price							
Market Thermal Gap							
Thermal Solar Energy							
Wind Energy							
Energy Demand							
Nuclear Energy							
FV Solar Energy							
Hydraulic Energy							

Para estudiar la precisión de los modelos, se utilizan los valores RMSE (error cuadrático medio) para comparar. El modelo con el valor RMSE más bajo es la Opción 1, que se convierte automáticamente en uno de los dos modelos principales.

Para crear un pronóstico estacional, se estudian cuatro modelos diferentes: TBATS (regresión trigonométrica, transformación de Box-Cox, errores ARMA, tendencia y componentes estacionales), SARIMAX (promedio móvil integrado autorregresivo estacional con factores exógenos), MSTL (promedio móvil estacional múltiple con descomposición de tendencia estacional utilizando modelos Loess) y Prophet. Los cuatro métodos incluyen estacionalidad, aunque no de la misma manera. SARIMAX es el único modelo que tiene la restricción de que solo permite una estacionalidad. En este caso se implementó la estacionalidad diaria por resultar más relevante. Las variables exógenas añadidas a este modelo son las mismas que las del modelo de Regresión Lineal Múltiple. TBATS, MSTL y Prophet permiten múltiples estacionalidades, pero solo Prophet no requiere que la estacionalidad se identifique de antemano: el propio modelo la estudia y la reconoce.

La siguiente tabla muestra los resultados para los cuatro modelos mencionados. SARIMAX es sin duda el mejor modelo estacional.

Model name	*RMSE*
TBATS	*30,757*
SARIMAX	*24,575*
MSTL	*63,503*
Prophet	*29,023*

La última variable que se añade al modelo fue la Tecnología de Cierre o Tecnología Marginal. Debido a que esta información no está disponible para el día en que se realiza la predicción, se implementa un Modelo de Regresión Logística. Este modelo predice la probabilidad de que una tecnología sea marginal durante una hora determinada. Los resultados de este modelo se introducen tanto en SARIMAX como en el modelo de Regresión Lineal Múltiple (RLM) como variables binarias, creando dos nuevos modelos separados (haciendo un total de 4 modelos).

Los modelos fueron estudiados y comparados, llegando a la conclusión de que los modelos SARIMAX tienen, en general, valores RMSE más bajos que los modelos RLM, pero no siempre es el caso. Para tomar la mejor decisión sobre qué modelo dará los resultados de la predicción, se analizan varias opciones para hacer el mejor sistema de recomendación. Estas opciones incluyen: modelo del día anterior (eligiendo el modelo que habría dado los mejores resultados 24 horas antes), un árbol de decisiones, un bosque aleatorio y un promedio de conjunto (Ensemble Averaging).

Después de ejecutar todos los modelos, los resultados fueron muy concluyentes: Ensemble Averaging (un promedio ponderado de todos los modelos después de optimizar los coeficientes) es el mejor método, reduciendo el sobreajuste (overfitting) y la varianza. Los otros modelos que le siguen en precisión son promedios no ponderados de dos de los modelos RLM y SARIMAX, lo que confirma la conclusión. Finalmente se puede decir que el mejor modelo es aquel creado por la combinación de otros modelos (con buenos resultados de predicción).

Referencias

Electricity in Spain. "Electricity Prices in Spain." *Electricity In Spain*, 6 Mayo 2023, electricityinspain.com/electricity-prices-in-spain/.

Querol Herrá, Javier. "Desarrollo de Un Modelo de Predicción Del Precio de La Energía Eléctrica Para El Mercado a Plazo Mediante Redes Neuronales." *Universitat Politécnica de Valencia*, 2019, riunet.upv.es/bitstream/handle/10251/121891/Querol%20Herr%C3%A1%20-%20Desarrollo%20de%20un%20modelo%20de%20predicci%C3%B3n%20del%20precio%20de%20la%20energ%C3%ADa%20el%C3%A9ctrica%20para%20el%20mercado%20a%20plazo%20mediante%20redes%20neuronales.pdf?sequence=3.

SHORT-TERM FORECASTING OF ELECTRICITY MARKET PRICES

Author: Reneses Rodríguez, Pía.

Supervisors: Ríos Calzada, María del Mar de los; Vargas Arboleda, Andrés Felipe.

Collaborating Entity: IGNIS ENERGÍA, S.L.

ABSTRACT

In the last decades, the relevance of forecasting Day-ahead electricity price has only increased. There are different benefits to this prediction including but not limited to bigger profit margins for individual market agents and assistance in the integration of renewable energy sources to the Spanish Electricity Market (Querol Herrá, Javier, 2019),

Be that as it may, forecasting electricity prices is a very complex process that has no one good solution. Over those decades since the electricity market stopped being a monopoly many techniques and methods have been tried and tested for this specific purpose. ARIMA, Multi Linear Regression and AI models are examples of them. All those depend on stationarity to forecast accurately. In recent years, this has become less of a reality regarding electricity prices in Spain (Electricity in Spain, 2023).

This project will focus on forecasting using statistical tools and not AI powered tools for the predictions themselves and using Machine Learning Algorithms to create a Model Recommender or Model Selector.

The project consisted of the study and comparison of different methods for modeling and forecasting. The first goal was to create two main models, one that would include seasonality or autoregressive terms and a regression model.

The regression modelled was a Multi Linear Regression. The variables added to the model were chosen from a list of variables that fell into the following categories:

- Price of gas and CO_2: these are the official daily values of the closing price of CO_2 – where just one value is available – and gas – where there are three possible variables and none of them are the actual price, but they are considered to be a very accurate approximation of the actual gas spot price.

- Previous prices of the day-ahead market of electricity: historical information of the marginal prices of electricity
- Predictions of generation for renewable energy sources: this applies to wind, photovoltaics (solar PV), solar thermal power (CSP) and hydro energy.

- Closing or marginal technology for the price hourly: this refers to the most expensive energy being sold in the market at a given hour, which would be the one accountable for the price of all electricity at that given time.
- Prediction for hourly energy demand: information that is published daily by *Red Eléctrica.*

The variables were tested for multicollinearity using the VIF (Variance Inflation Factor) as a measure. The assortments of variables that has low enough VIF values (< 5) were tested as models. The p-value of each of the variables was checked and the ones that had p > 0,05 were removed from the different models. After this process there 7 feasible models:

Variables/Options	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Gas Price OMIE							
Reference Gas Price							
Auction Gas Price							
CO2 Price							
Market Thermal Gap							
Thermal Solar Energy							
Wind Energy							
Energy Demand							
Nuclear Energy							
FV Solar Energy							
Hydraulic Energy							

To study the accuracy of the models, RMSE (Root-Mean-Square Deviation) values were the measurement used to compare. The one with lowest RMSE value was Option 1 which was chosen as one of the two main models.

To create a seasonal forecast, four separate models were tested: TBATS (Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend, and Seasonal components), SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors), MSTL (Multiple Seasonal-Trend decomposition using Loess) and Prophet models. All four methods include seasonality although not in the same way. SARIMAX is the only model that has a restriction and only allows one seasonality, in this case daily seasonality was implemented as it was found more relevant. The exogenous variables added to this model wre the same as the Multi Linear Regression model. TBATS, MSTL and Prophet all allow multiple seasonalities but only Prophet does not require the seasonality to be identified beforehand – the model itself studies and recognizes it.

The following table shows the results for all four mentioned models. SARIMAX was shown to outperform all other seasonal models.

Model name	*RMSE*
TBATS	*30,757*
SARIMAX	*24,575*
MSTL	*63,503*
Prophet	*29,023*

The last variable to be added to the model was Closing Technology. Because this information is not available for the day when the forecast is being made, a Logistic Regression Model was created. This model predicted how likely a technology was to be marginal during a certain hour. The results were introduced into both SARIMAX and Multi Linear Regression Models (MLR) as binary variables, creating two new separate models (for a total of four models).

The models were studied and compared, reaching the conclusion that the SARIMAX models have lower RMSE values overall than MLR models, but that is not always the case. To make the best decision several options were analyzed to make the best recommender system. These options included: Day-before model (choosing the model that would have given the best results 24 hours before), a Decision Tree, Random Forest, and Ensemble Averaging.

After testing all of them the results were very conclusive: Ensemble Averaging (a weighted average of all models after optimizing the coefficients) is the best method that reduces overfitting as well as variance. The other models that worked almost as well were unweighted averages of two of the MLR and SARIMAX models, reinforcing the conclusion. Finally, it can be said that the best forecasting model for day-ahead electricity prices is the one created by the combination of other models that already have good results themselves.

References

Electricity in Spain. "Electricity Prices in Spain." *Electricity In Spain*, 6 May 2023, electricityinspain.com/electricity-prices-in-spain/.

Querol Herrá, Javier. "Desarrollo de Un Modelo de Predicción Del Precio de La Energía Eléctrica Para El Mercado a Plazo Mediante Redes Neuronales." *Universitat Politécnica de Valencia*, 2019, riunet.upv.es/bitstream/handle/10251/121891/Querol%20Herr%C3%A1%20-%20Desarrollo%20de%20un%20modelo%20de%20predicci%C3%B3n%20del%20precio%20de%20la%20energ%C3%ADa%20el%C3%A9ctrica%20para%20el%20mercado%20a%20plazo%20mediante%20redes%20neuronales.pdf?sequence=3.

GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO FIN DE GRADO:

Modelo de previsión de precios en mercados de electricidad de corto plazo

Short-term Forecasting of Electricity Market Prices

Autor: Pía Reneses Rodríguez

Directores: María del Mar de los Ríos Calzada

Andrés Felipe Vargas Arboleda

Madrid

Junio de 2023

Index

Contents

CHAPTER 1: INTRODUCTION AND APPROACH TO THE PROJECT

The organization of the Spanish electricity market involves a Day-ahead market, as well as several sessions of Intraday auction markets and an Intraday continuous market. The day-ahead market or single day-ahead coupling (SDAC) is a fundamental component of the electrical energy production market. Its purpose is to facilitate electrical energy transactions by receiving bids to sell or acquire electricity on behalf of market agents for periods of 24 hours. Since 2014, this market has been integrated with the rest of Europe.

The electricity system in Spain is made of various entities or agents involved in the different part of the process. This process is divided into subsystems that have different roles to ensure the delivery of electricity: generation, transmission, distribution, and commercialization of electricity. Electricity generation is carried out by power station owners. They generate electricity and submit bids for each time slot for the following day. OMIE (*Operador del Mercado Ibérico de Energía* or Iberian Electricity Market Operator) is the nominated electricity market operator in Spain and Portugal. The purpose of OMIE is to oversee the management of buyers and sellers on the Day-ahead market – that operates based on supply and demand conditions – and ensure that all operations are efficient and transparent. Power stations can be represented in this market by third party companies, called market agents. The purpose of these companies is to help owners of generation assets maximize profitability by participating in all electricity markets in their stead. IGNIS, the collaboration entity of this project, is one of these companies.

The purpose of this project is to predict the prices of the Day-ahead market before OMIE comes out with the official ones at 12pm. A model will be created with the intention of using it to optimize the operation of the electricity generated by a hydroelectric power plan in the market.

For the creation of the final model, it is expected to create several different models similar in approach and choose two base models. The models will be improved and combined with the purpose of creating a definite model that has the best accuracy and efficiency possible. They will be analyzed and developed using statistical tools, as well as metrics and tools.

To finalize the project, we will choose a methodology to decide a final prediction. This could be one of the models if any of them are the "best" in a statistically significant way, a decision tree, a combination of other predictions or any other recommender system.

To complete this project, the mechanism that will be used is easily understood with the simple diagram in Figure 1.

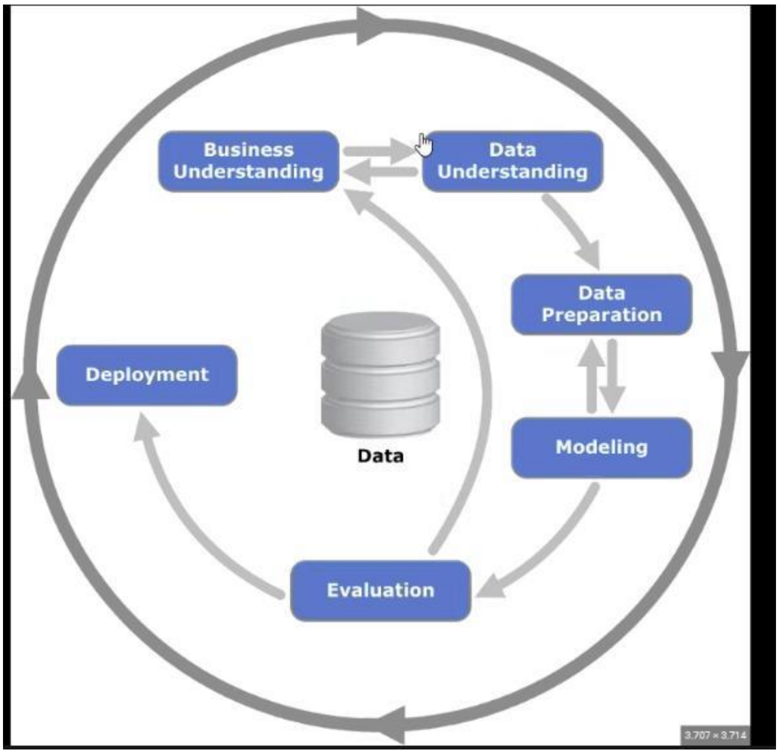

Figure 1. Diagram of Data Science Workflow (Saltz, Jeff, 2022)

Every model created and analyzed will include the steps presented above (Saltz, Jeff, 2022):

- Business Understanding: process expanded upon in the State of the Art, although it is the less focused on stage.
- Data Understanding: understand the data through its attributes (data audit), process that takes place prior to adding any variable to the models.
- Data Preparation: prepare the data for use and analysis. This process will mostly take place in code so it will not be detailed in the text. This fact notwithstanding, this step is essential.
- Modeling: create models using statistical or Artificial Intelligence tools, justifying each step that is taken. This part will be thoroughly explored in the following pages.
- Evaluation: assess the efficacy of a model using different metrics to understand its performance. This step will be necessary after the models are finished and once again after the recommender system is done, to check its accuracy.
- Deployment: not a part of this project.

CHAPTER 2. – STATE OF THE ART

Since the end of government monopolies and the start of competitive markets in the electricity sector over 30 years ago, power exchanges around the world have faced more challenges and uncertainties. These are caused by the growing share of renewables, changing regulations and global financial instability. To cope with these changes, participants and agents need to use methods that can forecast electricity market prices (Marcos Peirotén, Rodrigo Alejandro, 2020).

Spain is one of the countries affected by this phenomenon. The need to forecast electricity prices comes about after years of forecasting electricity demand, which is where all the new research and development of forecasting techniques is based. Agents in the Spanish Electricity Market need the prediction of the electricity prices with the highest possible accuracy for the sake of profit margins in the operations in the market (Querol Herrá, Javier, 2019). Not only that but forecasting day-ahead electricity prices can improve the system stability by avoiding sudden price peaks and grid imbalances. And lastly, it can facilitate the integration of renewable energy sources by accounting for their volatility and variability (this will be further elaborated on in Annex I). The last two decades have brought about increasing interest in the topic and it is a very researched one today.

The most well-known techniques for short-term statistical procedures are based on time-series analysis which range from simple linear regression models to complex seasonal and autoregressive models such as ARIMA. However, such models depend on stationarity in order to forecast short-term price accurately. In recent years, this has become less of a reality regarding electricity prices in Spain (Electricity in Spain, 2023). This situation has resulted in the need for more complex tools, like the Box-Cox transformation, to ensure that the predictions are as effective as possible.

Such statistical techniques are the most appropriate approach when predicting short-term prices, even as AI tools can also be suited for this purpose (Elseidi, Mohammed, 2022). There have been instances of such tools being introduced in this context, but the results are not distinctively positive. Models based on ARIMA are more commonly used for this purpose as they can introduce not only autoregressive and seasonal terms but also other independent variables, holiday effects and other such features (Caro, Eduardo, and Jesús Juan, 2020).

It is of the utmost importance to agents in the Spanish Electricity System to be able to predict prices in the short term, so they can know in advance the best course of action. For these reasons there exist many instances of statistical and AI methodologies used with these intentions. The methods used to forecast short-term prices are not a good representation of the medium or long term, where statistical models are not expected to behave with the same level of accuracy (Caro, Eduardo, and Jesús Juan, 2020).

As mentioned before one of the most commonly used methods to base a forecast around is ARIMA (Cruz García, Alberto Miguel, 2013) but this model can be too complex and time consuming. The balance between computational speed and accuracy of a forecast is an age-old dilemma that has brought about simpler models that rely on the simplification of periods – that are represented by their corresponding respective periods that are expected to behave in a similar way (Marcos Peirotén, Rodrigo Alejandro, 2020).

Taking all the above into consideration, the intention of this project to investigate and test these methodologies and come to conclusions regarding their accuracy and usefulness is unquestionably relevant.

CHAPTER 3. – CREATION OF MODELS

The project consists of the creation of a series of models that use different statistical methods to try and predict the hourly price of electricity for the following day before Red Eléctrica announces the actual prices. In order to use the different models, there will be a methodology with the purpose of choosing which method is most likely to be the best or the most successful prediction and therefore should be the one to be used.

3.1. – OBJECTIVES AND SPECIFICATIONS

The purpose of the model is to be able to have an idea of what the prices of the Spanish electric market will look like one day in advance. To achieve this goal, available data will be used to study and understand the behavior of the price in time and a series of other variables that have been selected as potentially relevant to this behavior.

Different options of models that could potentially be useful will be delved upon and discarded or chosen. They will be compared using statistical measures such as RMSE and AIC between each other. The viability of each of them will be checked using other such measures – VIF and p-values, for example.

After the simpler models have been chosen, we will investigate the possibility of additional variables. Once all possible models have been selected, the arithmetic mean of every two models that are considered relevant will be calculated. These are all the cases that must be studied and compared later to create the best possible recommender system. We will use RMSE values to identify the best method, if there were one, or several methods if more than one method is effective or they work best depending on the circumstances.

3.2. – DATA

Statistical modeling consists of the mathematical representation of one or several variables from previously observed data (Graduate Programs Staff, 2023). In this case, to carry out these representations, a big pool of data exists that can be considered potential variables of the different models. When choosing which ones should be used and which one should be discarded there needed to be a preliminary process. This process consisted of an investigation into past research on the topic.

The conclusions reached were the following: the price of electricity directly depends on the most expensive technology that is being cleared in the market during that hour (Andrés Palacios, Marcos, 2019). This gives significance to the generation each technology is expected to have and will be reflected in the chosen variables (as well as information directly reflecting the marginal or closing technology, although this will be further elaborated on later). It can be inferred that the demand for electricity is equally relevant for the same reasons.

Another factor that has become increasingly relevant in this topic is the price of CO2 emissions. The European Union issues EUA or Europe's emission allowances, which are traded at certain prices. These prices have become increasingly higher in the last years which is why their effect on the price of electricity has also increased. These prices change every weekday. They directly affect the price of electricity because they have caused a significant escalation in the cost of the production of electricity from fossil fuels. Likewise, gas prices also affect production of electricity in Combined Cycle Gas Turbines (CCGTs) because gas is the primary resource that they transform into electricity (Pacce, Matías José, et al., 2021).

The full list of variables that were considered can be found in Table 1. All these variables have been chosen because they are directly related to the price of electricity. They fall in different categories according to their relations to the daily price:

- Price of gas and CO_2: these are the official daily values of the closing price of CO_2 – where just one value is available – and gas – where there are three possible variables and none of them are the actual price. Instead, they are Day Ahead predictions from the day before for the price that will not be known until the information is no longer useful to make predictions. However, these three variables are considered to be a very accurate approximation of the actual gas spot price.

- Previous prices of the day-ahead market of electricity: one of the variables being considered is the price of electricity itself. This would be used in the form of autoregressive terms from a day before, a week before, a month before or any other format that can potentially be relevant. It will be considered a variable that is likely to be useful given that the prices have daily, weekly, monthly, and yearly seasonality.

- Predictions of generation for renewable energy sources: this applies to wind – which can be found in several databases –, photovoltaics (solar PV), solar thermal power (CSP) and hydro energy.

- Closing or marginal technology for the price hourly: this refers to the most expensive energy being sold in the market at a given hour, which would be the one accountable for the price of all electricity at that given time.

- Prediction for hourly energy demand: information that is published daily by *Red Eléctrica.*

Variable name	***Units***
PMD (Daily Market Price)	*€/MWh*
Gas Price OMIE	*€/MWh*
Reference Gas Price	*€/MWh*
Auction Gas Price	*€/MWh*
CO2 Price	*€/ton*
Market Thermal Gap	*MW*
Thermal Solar Energy	*MW*
Wind Energy	*MW*
Energy Demand	*MW*
Nuclear Energy	*MW*
FV Solar Energy	*MW*
Hydraulic Energy	*MW*
Closing Technology	-

Table 1. Potential variables of the model and their units

Each of the variables listed above are measured using different units (euros per certain units for prices, MW for power and such). If the purpose of this project was to understand the effect that they each had on the resulting price it would be necessary to normalize the units so that the coefficients for the variables would have a meaning. However, the focus of this project is not to explain these specific effects but to create and understand models that use the variables and so this step should not be considered compulsory.

One of the variables above is not as self-explanatory as the others and that is because it is a variable created by operating some of the other variables: 'Market Thermal Gap'. This is the name given to the power covered by thermal production (CCGTs, typically). Its relevance is due to it being the part of the power affected by CO_2 and gas prices. It can be calculated by subtracting the expected power from renewable and nuclear energies from the expected demand. As an example of Data Preparation (Figure 1) there is an example of the SQL code bellow, – all variables are imported from SQL data bases –, used to generate the aforementioned 'Market Thermal Gap' variable.

```
SELECT a.Date, a.Period, EnergyDemand-(WindEnergy+
FVSolarEnergy+NuclearEnergy+ThermalSolarEnergy) AS MarketThermalGap
                    FROM(
```

```
                        SELECT Date, Period, Technology, value AS WindEnergy
                        From mercados.tb_datos_1
                        WHERE Date > {Date_i}
                                AND Date < {Date_f}
                        AND Technology = 'wind'
                        AND variable = 'powergeneration'
                        AND tipodato = 'forecast'
                        AND tipoforecast = 'meteologica'
                        AND Period < 25) a
                    INNER JOIN(
                        SELECT Date, CEIL(Period/4) AS Period,
SUM(if(id=460,valor/4,0)) AS EnergyDemand, SUM(if(id=542,valor/4,0)) AS
FVSolarEnergy, SUM(if(id=543,valor/4,0)) AS ThermalSolarEnergy
                                FROM mercados.tb_datos_2
                        WHERE Date> {Date_i}
                                AND Date< {Date_f}
                        AND id IN(460,543,542)
                        AND QH = 'QH'
                        GROUP BY Date, CEIL(Period/4)) b
                        ON a.Date = b.Date
                         AND a.Period = b.Period
                    INNER JOIN(
                        SELECT Date, Period, SUM(value) AS NuclearEnergy
                            FROM mercados.tb_datos_3
                            WHERE Date > {Date_i}
                            AND Date < {Date_f}
                            AND id IN(474)
                            AND QH = 'H'
                            GROUP BY Date, Period)c
                        ON a.Date = c.Date
                         AND a.Period = c.Period
                  INNER JOIN(
                      SELECT Date, CEIL(Period/4) AS Period, AVG(ctd) AS
EnergyDemand
                      FROM mercados.tb_demanda_4
                      WHERE Date > {Date_i}
                      AND Date < {Date_f}
                      AND HOUR(fechahora_prevision) LIKE '10%'
                      AND DATE(fechahora_prevision) = ({Date_f} - 1)
                      GROUP BY Date, CEIL(Period/4) )d
                    ON d.Date = a.Date
                    AND a.Period = d.Period

              ORDER BY Date, Period ASC
```

Before it was possible to analyze the relevance of the selected data when making predictions it was necessary to perform a data audit on them. A data audit consists of a report whose purpose is to evaluate whether data is fit for any given purpose (Butlion, Justin, 2021). In this case an exhaustive analysis of the data was conducted. For it, every potential variable was studied using the following criteria looking at the period between 01/01/2021 and 01/10/2022 (format dd/mm/yy):

- Whether there were any gaps in the data. Some of the variables did not have information for weekends and holidays. This was not a mistake but the way the data exist. For example, the price of CO_2, one of the potential variables, works this way. The price only changes on working days, meaning that during weekends and holidays it is the same as it was the last weekday. A different issue was raised with the switch between daylight saving time and standard time. Every year there is a day with 25 hours and one with 23 hours. Every variable saved the information from these two days in a different way that had to be taken into account when auditing this data.
  If any gaps were found, several measures were taken. For the information that had not been properly loaded in the tables, the person in charge was notified and the issue was promptly solved. For any information that was not available for one reason or another, the choice was made to drop the variable if it was easily replaced with a different one. If it was not, the gaps would be filled with the same value as the last available point in time.

- Whether the distribution of the data made sense and was consistent enough to use as a variable. For this purpose, statistical measures were analyzed. For central tendency, the geometrical mean; for variability, the standard deviation. The quartiles (25^{th}, 50^{th} and the 75^{th} percentiles) were evaluated to identify outliers or anomalies in a dataset and identify obvious skewness or bimodality (Anderson, Daivid, et al., 2023). The maximum value, the minimum value and the mode were also analyzed to finish studying the data.

- Whether one variable was highly correlated to another one. Some of the variables were chosen knowing that there was expected to be one with the purpose of choosing the most relevant out of the two if any at all. Of course, this process was done with a focus on the nature of the variables and later confirmed by checking the numerical correlation between them. An example of variables that got this treatment is the expected value for the different technologies and the 'Market Thermal Gap'– adding all of them would result in severe multicollinearity. In a case like this it would be pertinent to understand the information to be redundant and so it should not be used simultaneously (or otherwise one of the predictions would be considered inaccurate).

Any data that was missing or considered not apt to be used for the purposes of the project will be removed and replaced with the last available data point. An example resulting from this data audit can be found in the tables below. The example in

Table 2, that continues in

Table 3 and Table 4 corresponds to the information on the variable 'Wind Energy'. It is essential to understand the meaning of all features of a variable before using it to make sure that no mistakes are made that could result in wrong or invalid results if the variable is introduced into a model.

Feature: DATE	
Number of rows	*12 382*
Unique values	*516*
% Repeats	*95,83*
Missing Count	*0*
Missing Rate	*0*
Description of feature	*Primary Key: Date*
Observations	*January 2021 – May 2022*
Feature: QH	
Number of rows	*12 382*
Unique values	*1*
% Repeats	*99,99*
Missing Count	*0*
Missing Rate	*0*
Description of feature	*Primary Key: Frequency of data (QH for quarter-hourly data or H for hourly data)*
Observations	*All data is hourly (H)*
Feature: PERIOD	
Number of rows	*12 382*
Unique values	*24*
% Repeats	*99,81*
Missing Count	*0*
Missing Rate	*0*
Description of feature	*Primary Key: Period of the day*
Observations	*Values from 1 till 24*

Table 2. Example of Data Audit of 'Wind Energy' Variable 1/3

Feature: ID	
Number of rows	*12 382*
Unique values	*1*
% Repeats	*99,99*
Missing Count	*0*
Missing Rate	*0*
Description of feature	*Primary Key: Data Identifier*
Observations	*Only has one value (541) that indicates 'Predicted Wind Energy'*
Feature: GEO_ID	
Number of rows	*12 382*
Unique values	*1*
% Repeats	*99,99*
Missing Count	*0*
Missing Rate	*0*
Description of feature	*Primary Key: Geographical Identifier*
Observations	*Only has one value (8741) that indicates 'Spain'*

Table 3. Example of Data Audit of 'Wind Energy' Variable 2/3

In 'Description of feature' it refers to several of the features as primary keys. Primary keys are features that uniquely identify a data point. It is necessary that all primary keys contain some sort of information in SQL, or the table will not function at all. They can contain information such as the one in the tables above: date, time, geographical information, or data identifiers.

Feature: VALUE	
Number of rows	*12 382*
Unique values	*8 122*
% Repeats	*34,40*
Mean	*6 867,69*
Standard Deviation	*3 847,36*
Min. value	*196*
25 Percentile	*3 793,25*
50 Percentile	*6 221,50*
75 Percentile	*9 429,75*
Max. value	*19 757*
Missing Count	*0*
Missing Rate	*0*
Zeros Count	*0*
Zeros Rate	*0*
Description of feature	*Value of 'Predicted Wind Energy' in MW*
Observations	-

Table 4. Example of Data Audit of 'Wind Energy' Variable 3/3

3.3. – DEVELOPMENT OF THE MODELS

The project consists of different models built similarly – with different methods but the same or almost the same variables every time –. All the models chosen and tried will now be explained, in the chronological order that they were studied and implemented, modified, and discarded.

3.3.1 MULTI LINEAR REGRESSION MODEL

Firstly, the simplest and most straightforward method to predict the value of anything given historical data is a linear regression. In this case, a multiple linear regression model was chosen – this does not change the result, given that if the best model had truly been with just one variable that would be seen no matter how the process started.

Multiple linear regression analysis in statistics is used to predict the value of one variable (the price of electricity in this case) using known information about other variables (these will be chosen from the ones presented previously). This method assumes that there is a linear relationship between the dependent variable and all the explanatory variables. For it to be accurate and effective, it is necessary to be certain that the independent variables do not show multicollinearity.

The formula used in this method is broken down here:

$$E.\ 1 \qquad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ \dots + \beta_n X_{in}$$

Y_i = dependent variable

X_i = independent or explanatory variables

β_0 = constant term

β_n = slope coefficients for each independent variable

"Multicollinearity occurs when two or more independent variables in a data frame have a high correlation with one another in a regression model" (Bhandari, Aniruddha, 2023). One of the variables being checked has been created using other variables; the market's thermal gap is the result of subtracting the values of renewable energy sources from the demand. It is hence expected that there will be noticeable collinearity within the variables. It is also foreseeable that the previously mentioned variables that were heavily related to one another will also surely be collinear. The analysis for multicollinearity can be done using resources in Python that provide a metric known as VIF or the variance inflation factor. VIF represents the strength of correlation between independent variables in a regression model, taking values between 1 and positive infinity (Bobbitt, Zach, 2022). A VIF = 1 means that there is no correlation between a variable and the rest; 5 > VIF > 1 means there is a moderate correlation, which in this case will be considered acceptable; VIF > 5 means there is a strong correlation, which will not be deemed acceptable.

In the table below (Table 5) an example of data from the month of February (2023) has been analyzed. As expected, the values of the VIF for the gas prices are above the required 5, as well as the values of the energy predictions that are used in the creation of the variable 'Market Thermal Gap'. Any values in red are too high to be acceptable.

Variable name	*VIF*
Gas Price OMIE	*26,58*
Reference Gas Price	*9,28*
Auction Gas Price	*47,03*
CO2 Price	*3,59*
Market Thermal Gap	*187,79*
Thermal Solar Energy	*3,60*
Wind Energy	*69,09*
Energy Demand	*92,29*
Nuclear Energy	*3,19*
FV Solar Energy	*72,05*
Hydraulic Energy	*12,27*

Table 5. Table with VIF values for all variables together (February 2023)

It is important to take into consideration that this information does not speak to the relevance of each variable or how useful they are when explaining the target variable. This is merely a requirement necessary to use the multi linear regression model. The following tables can be used to prove that all the combinations of variables shown are viable. For the sake of conciseness not all of them are presented. From the knowledge of the variables, it is easy to understand that the variables related to gas prices are correlated between each other and not with the other variables (the ones that represent production of energy of different technologies). It is therefore unnecessary to prove that all the possible solutions do not present multicollinearity, just show that each of them is independent in at least one example.

In Table 6 we can see that although removing the gap significantly reduces all VIF values for the energy related variables, although some of them still surpassed the desired threshold. In Table 7 it is shown that, as expected and assumed before, only one variable of each ‘group’ will result in VIF values that are closer to 1. As seen in Table 8, there exists a combination of the thermal gap and some of the other energy variables that still complies with the parameters. These results have been achieved after an exhaustive study of possible combinations and are presented in this way to facilitate their understanding.

Variable name	*VIF*
Gas Price OMIE	*3,07*
CO2 Price	*3,33*
Thermal Solar Energy	*2,82*
Wind Energy	*1,71*
Energy Demand	*8,56*
Nuclear Energy	*1,85*
FV Solar Energy	*6,22*
Hydraulic Energy	*10,75*

Table 6. Table with VIF values Gas Price OMIE and all energies

Variable name	*VIF*
Reference Gas Price	*2,28*
CO2 Price	*2,28*
Market Thermal Gap	*1,07*

Table 7. Table with VIF values Reference Gas Price and Market Thermal Gap

Variable name	*VIF*
Auction Gas Price	*2,99*
CO2 Price	*3,39*
Market Thermal Gap	*1,76*
Wind Energy	*2,37*
Energy Demand	*1,98*
Nuclear Energy	*1,66*

Table 8. Table with VIF values Auction Gas Price, Thermal Gap, and some energy variables

Once it is certain that the chosen independent variables do not show any multicollinearity, the next step is to study the potential linear relationship between the dependent variables and all other variables. For this purpose, the p-value will be used. The p-value measures the probability that the null hypothesis – in this case, that the variable has indeed a linear relationship with the dependent variable – should be rejected (Frost, Jim, 2023). If the p-value is under 0,05 results will be considered statistically significant. Out of the p-values of variables that are not statistically significant the one that is highest will be removed until all are under 0,05. In the tables, all p-values that are not acceptable are written in red.

In Table 9 as it can be seen below, most p-values indicate that the variables are not actually statistically significant. The result after removing the necessary variables, starting with 'Nuclear Energy' is presented in Table 10.

Variable name	*p-value*
Constant	*0,352*
Reference Gas Price	*0,153*
CO2 Price	*0,371*
Market Thermal Gap	*3,8E-5*
Wind Energy	*0,127*
Energy Demand	*0,011*
Nuclear Energy	*0,797*

Table 9. Table with p-values for all variables first attempt (February 2023)

Variable name	*p-value*
Constant	*0,0049*
Reference Gas Price	*0,045*
CO2 Price	*3,14E-4*
Market Thermal Gap	*9,23E-122*
Wind Energy	*0,01*

Table 10. Table with p-values < 0,05

Using this method, all possible combinations of variables that meet the criteria can be found. They are listed here and shown in the table right below (Table 11).

- Option 1: 'Reference Gas Price', 'CO2 Price', 'Market Thermal Gap', 'Wind Energy'
- Option 2: 'Reference Gas Price', 'CO2 Price', 'Market Thermal Gap'
- Option 3: 'Reference Gas Price', 'CO2 Price', 'Wind Energy', 'Energy Demand', 'FV Solar Energy',
- Option 4: 'Gas Price OMIE', 'Wind Energy', 'Energy Demand', 'FV Solar Energy'
- Option 5: 'Gas Price OMIE', 'Market Thermal Gap', 'Wind Energy', 'Hydraulic Energy'
- Option 6: 'Auction Gas Price', 'CO2 Price', 'Market Thermal Gap'
- Option 7: 'Auction Gas Price', 'CO2 Price', 'Wind Energy', 'Energy Demand', 'FV Solar Energy'

Variables/Options	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Gas Price OMIE							
Reference Gas Price							
Auction Gas Price							
CO2 Price							
Market Thermal Gap							
Thermal Solar Energy							
Wind Energy							
Energy Demand							
Nuclear Energy							
FV Solar Energy							
Hydraulic Energy							

Table 11. Variables of each Option for Multi Linear Regression Model

In all cases, the first value that needs to be removed is 'Nuclear Energy'. This is easily understood if one thinks about the expected values for this variable: it is not supposed to change almost at all, and it has very constant values all throughout the month, which makes it difficult for it to be linearly correlated to the electricity price. However, it is a very significant value when it changes, because it has a very big effect on the price of electricity. This would happen at times when a nuclear energy facility is disconnected from the electricity system for maintenance reasons. Since none of the viable combinations include 'Nuclear Energy' it would be convenient for the 'Market Thermal Gap' variable to be in the final choice, given that the former is included in the latter.

One of the most noticeable things about the daily electricity price is that it is very periodic. A multi linear regression is not an appropriate method to represent this, but to somewhat take it into account three variables will be added to the ones above mentioned: the price of electricity 24 hours before – to consider daily periodicity –, the price of electricity 7 days before - to consider weekly periodicity – and the price of electricity 1 month before – to consider monthly periodicity –. Before we can go ahead with this addition, it is essential that the p-values of the new variables also comply with the requirements. The conclusion from these tests is clear: in all cases the first two variables are relevant but not the third. There is no monthly seasonality to the electricity prices.

The last decision, deciding which model is the best and should consequently be chosen is the only left regarding the multi linear regression model. The statistical measure of accuracy we will use is the RMSE or root mean square error. RMSE is a standard way to measure the error of a model (Moody, James, 2019). In this case, to compare the different options, information of 2 months of predictions (November and December 2022), a total of 1464 data points will be used.

Model name	*RMSE*
Option 1	*18,720*
Option 2	*19,012*
Option 3	*20,681*
Option 4	*20,654*
Option 5	*18,799*
Option 6	*18,976*
Option 7	*19,345*

Table 12. Table with RMSE values for all Multi Linear Regression Models

Table 12 shows the values of the RMSE of all the options – described with their variables in the table above, with the additional two regressive variables. Since all the values are relatively close, it can be concluded that no overfitting has occurred. If that were the case, the options which had that issue would have a significantly smaller RMSE value. This result was to be expected because the sample used was big and the number of variables used to fit the model was not. The information in the table reaches the conclusion that the ideal Multi

Linear Regression Model that should be used is Option 1, the one that included 'Reference Gas Price', 'CO2 Price', 'Market Thermal Gap', 'Wind Energy', 'Price D -1' (price day minus 1) and 'Price W -1' (price week minus 1) as variables.

In Figure 2 there is a graph of an example of a prediction with the final model chosen where the model is working properly as expected. Graphing the results and comparing them with the expected results is a very effective way of checking that the predictions make sense and not only has a small RMSE but also has the correct shape, which is as important as being accurate with the numbers if not more.

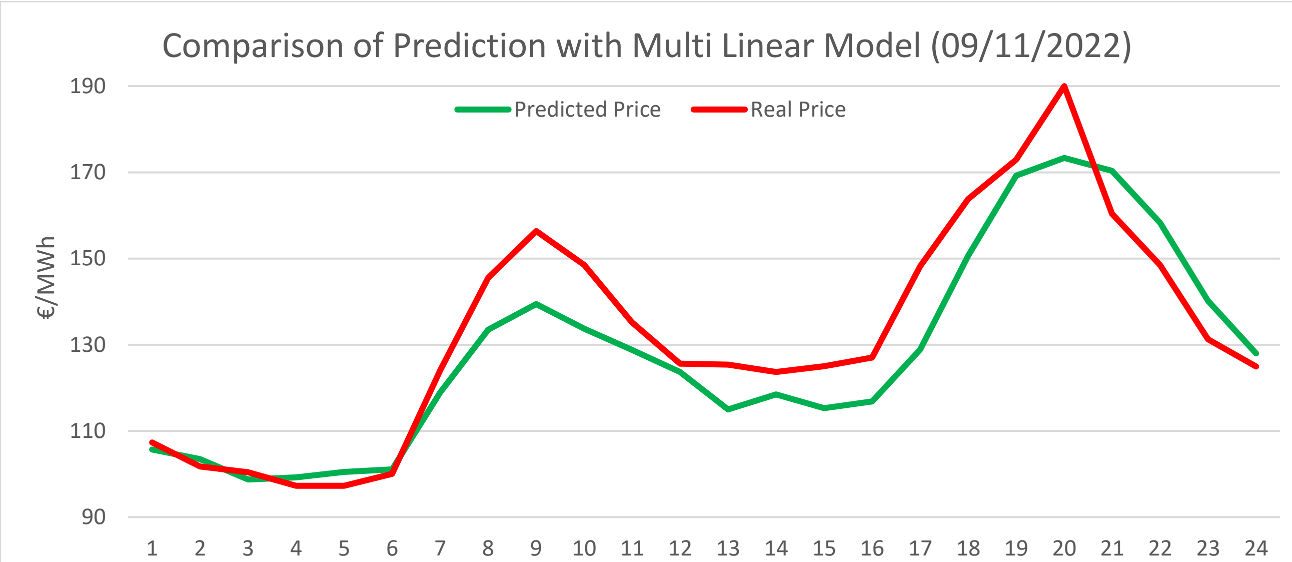

Figure 2. Example of Prediction with Multi Linear Model (09/11/2022)

3.3.2 SEASONAL AND AUTOREGRESSIVE MODEL

As was brought up above, one of the main characteristics of the electricity price of the daily market is that it has periodicity or, in other words, it is seasonal. The focus now will be on finding the best model that implements information about past behavior and/or seasonality. For this purpose, it is relevant to firstly prove that such behavior exists.

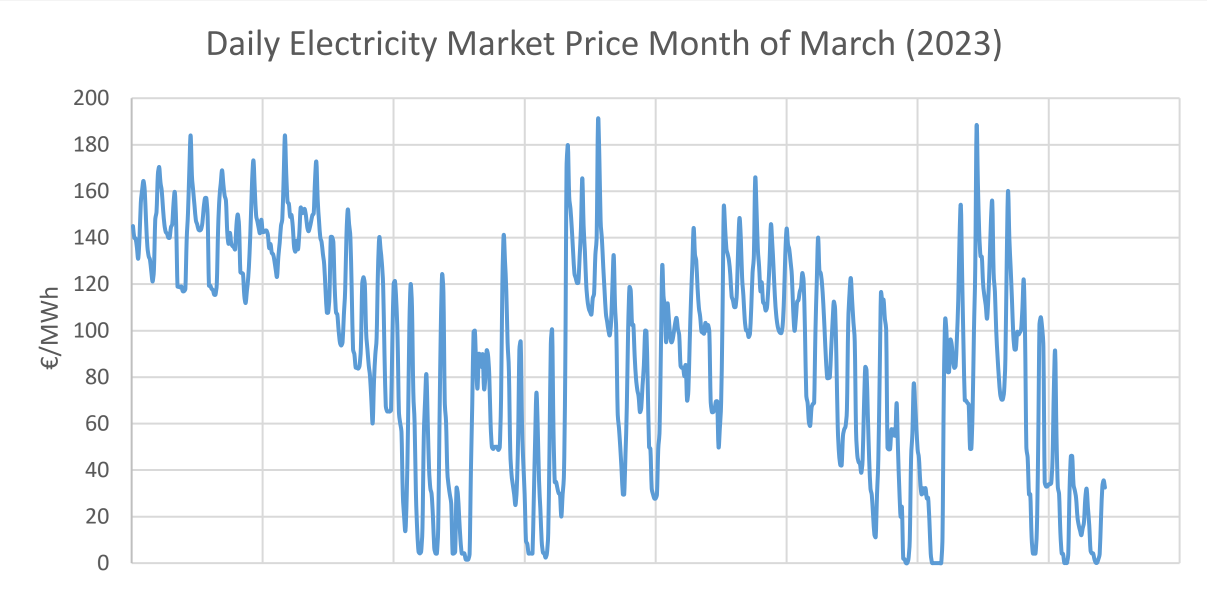

Figure 3. Daily Spanish Electricity Market Prices (03/2023)

To study seasonality, the values of the price of electricity are seen plotted in Figure 3. It seems like there is some trend in the data, but it requires further examination. For example, it is possible to find out whether the seasonality of our data is additive or multiplicative. To understand the difference, we can refer to the equations below:

$$E.2 \qquad Y[t] = T[t] + S[t] + e[t]$$

$$E.3 \qquad Y[t] = T[t] \cdot S[t] \cdot e[t]$$

$Y[t]$ = the series (electricity prices)

$T[t]$ = general trend (could be understood as the moving average)

$S[t]$ = seasonality (pattern that occurs at regular intervals)

$e[t]$ = residual (noise not accounted for by the other two terms)

The equation E.2 explains the form of an additive seasonal model, where the amplitude of the seasonality does not have big changes throughout time. The equation E.3 refers to the form of a multiplicative seasonal model, where the seasonality is proportional to the trend in some way; consequently, it changes over time (Hayes, Spencer, 2022). In this case, the seasonality is additive because the daily (and weekly) periods will surely continue to be the same length no matter what the general trend looks like. Furthermore, multiplicative seasonality cannot be used when the data has zero or negative values.

Before we start the study of seasonality and seasonal and autoregressive forecasting models it is important to explain what a time series is, since this term will recurrently come up in the following explanations. A time series is simply measurements that are tracked over time, meaning that their indexes are chronological time stamps. In this case, the price of electricity is a time series that has a data point for every hour. Therefore, the term time series can and will be used interchangeably with data set in the context.

In the interest of identifying which of these seasonalities the data is a part of, the time series is broken down into its trend, seasonality, and residual components as seen in Figure 4. The information in the graphs confirms the expected daily seasonality, but the same cannot be said of the weekly seasonality, which cannot be appreciated here. Fittingly, there is no pattern in the residuals.

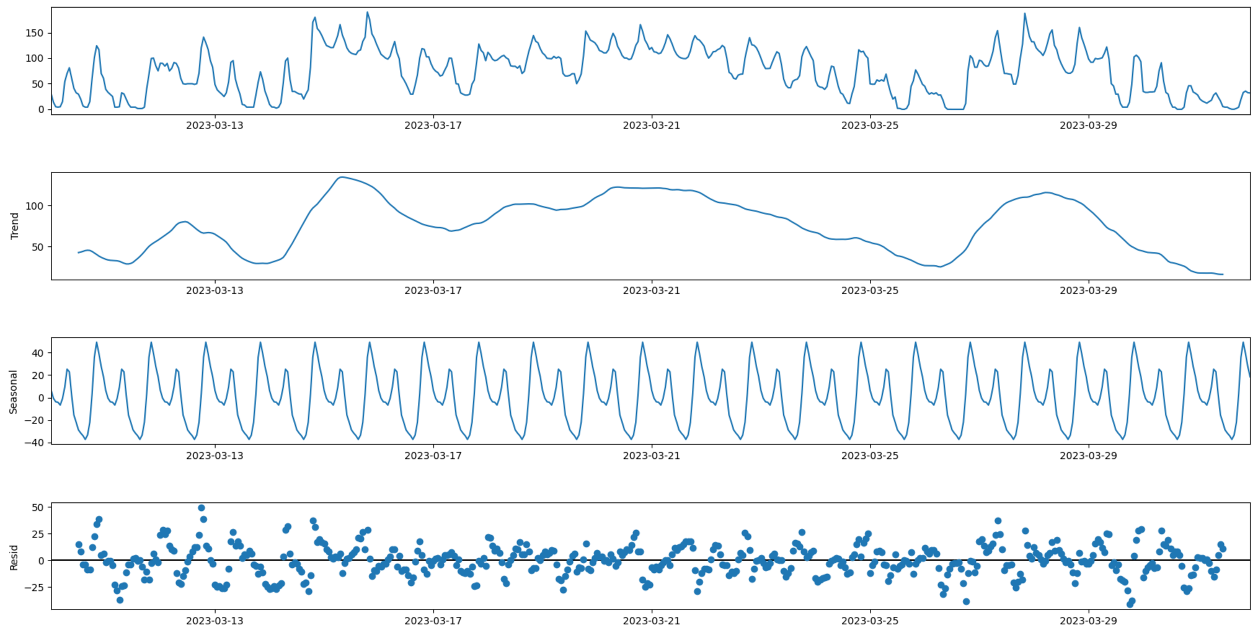

Figure 4. Seasonal decomposition into trend, seasonality, and residuals of electricity prices (03/2023)

The first option for a seasonal model that will be studied is called the TBATS method. The acronym stands for Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend, and Seasonal components. The forecasting method has two parts mentioned and explained previously – trend and seasonal component –, an ARMA model applied on the residual (which will be touched upon later on), a Box-Cox transformation (that makes the series stationary), and trigonometric seasonality used to represent all the seasonal periods based on Fourier series (Azaizah, Nadeem, 2021).

$$\text{E.4} \qquad \boldsymbol{s_t^{(i)} = \sum_{j=1}^{(k_i)} s_{j,t}^{(i)}}$$

$$\boldsymbol{s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cdot cos(2\pi j/m_i) + s_{j,t-1}^{*(i)} \cdot sin(2\pi j/m_i) + \gamma_1^{(i)} \cdot d_t}$$

$$\boldsymbol{s_{j,t}^{*(i)} = -s_{j,t-1}^{(i)} \cdot sin(2\pi j/m_i) + s_{j,t-1}^{*(i)} \cdot cos(2\pi j/m_i) + \gamma_2^{(i)} \cdot d_t}$$

Where:

$s_t^{(i)}$ = ith seasonal component

d_t = ARMA process for residuals

$\gamma_1^{(i)}, \gamma_2^{(i)}$ = seasonal smoothing

m_i = length of ith seasonal period

The equation E.4 explains the “trigonometric representation of seasonal components based on Fourier series”. The basis is that of the Fourier series as we can see in the first two terms of the component (sin and cos terms). It considers the residuals with a seasonal smoothing technique, meaning that they take into account the residual of the same position the previous periods. If one were predicting, for example, for a Monday at 9:00am with a daily and weekly seasonality it would use as most important the information from the Monday of the previous week at that same time.

The Box Cox Transformation is the tool used to transform non-normal datasets into a normal distribution. This makes it easier to find patterns in the data, facilitating the process to find the best model. It also allows the use of statistical techniques that might otherwise lose accuracy. Lastly, thanks to the transformation, the variability of the data does not have a big effect on the results of the prediction (Pannell, Reagan, 2022).

In Figure 5 there is an example of the TBATS model. The examples used to represent each of the models are not enough evidence to conclude that they are better or worse than any other model. However, they are specifically chosen because they accurately show problems that arise with them. In this case, the model is relatively good at following the seasonality and general shape of the price. However, it seems to lack information and cannot perfectly depict the trend. The model’s accuracy will be studied further as it is compared to the other seasonal methods.

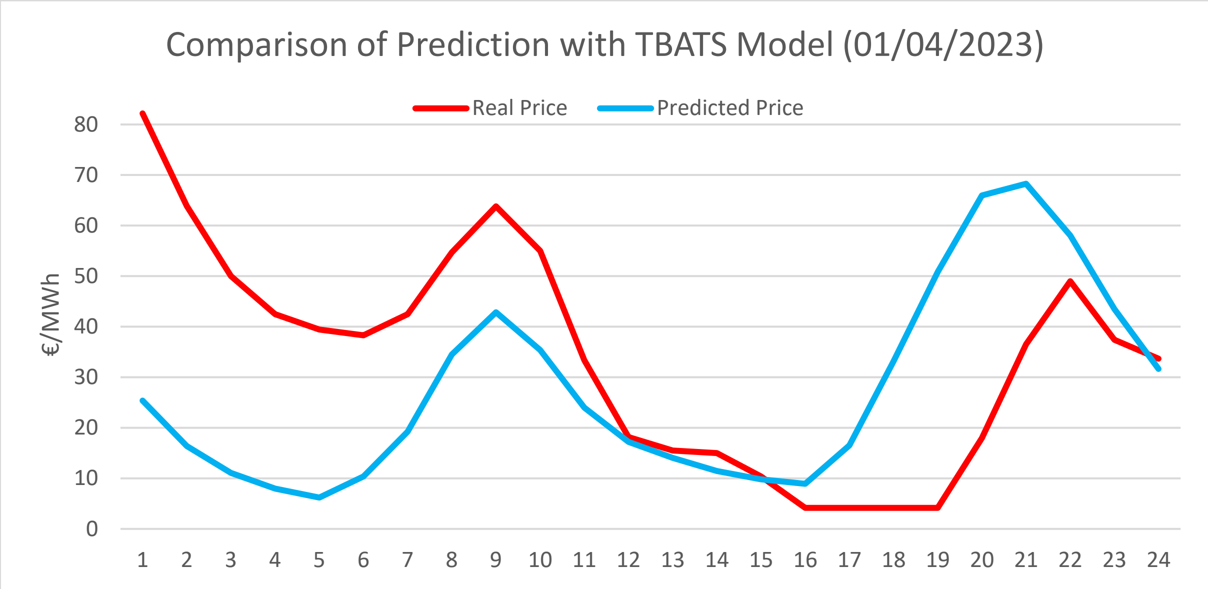

Figure 5. Example of Prediction with TBATS Model (01/04/2023)

The second option for a seasonal model that will be considered is the ARIMA method. The ARIMA model as-is is a non-seasonal model, but there are many options for models classified under the ARIMA umbrella. ARIMA stands for Autoregressive Integrated Moving Average.

The model is integrated if it is non-stationary and differencing is applied to normalize it. This part of the ARIMA (I) is represented by the parameter d (d is the number of nonseasonal differences). The Augmented Dickey-Fuller test can be used to check whether the data is stationary or not. The null hypothesis of this test is that the data is stationary, so it will be non-stationary if the p-value is greater than 0,05 even though ADF < critical values (Prabhakaran, Selva, 2022). The Augmented Dickey-Fuller test results are displayed Table 13. The p-value is less than 0,05, so it can be concluded that the data set is non-stationary. In this case the value of d is expected to be different from 0.

There are two methods to find the specific value of d. The first one consists of graphing the original data and two orders of differencing. The one where the data becomes stationary and there is no additional noise will be the one selected. In Figure 6 that means selecting 1st order. In the second method, the autocorrelation plots of the original data and two orders of differencing. The order before the one where the immediate lag has gone on the negative side will be selected (Verma, Yugesh, 2022). In Figure 7 the results would be selecting 1st order as well, so the value of d would be 1.

ADF	*-5,6622*	
p-value	*0,04*	
Num of Lags	*25*	
Num of Observations Used for ADF Regression and Critical Values Calculation	*3622*	
Critical Values	*1%*	*-3,4322*
	2%	*-2,8623*
	3%	*-2,5672*

Table 13. Example Results Augmented Dickey-Fuller Test (04/02/2023)

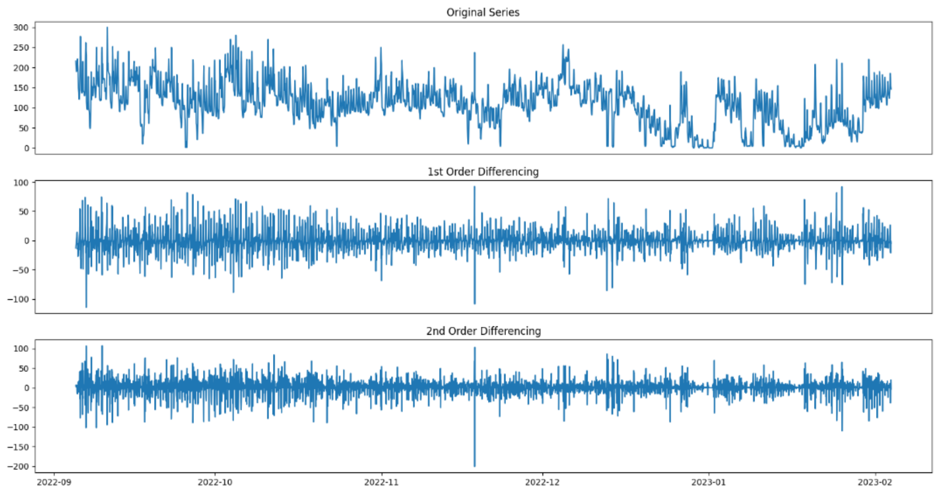

Figure 6. Example Data set, 1st and 2nd Differencing (04/02/2023)

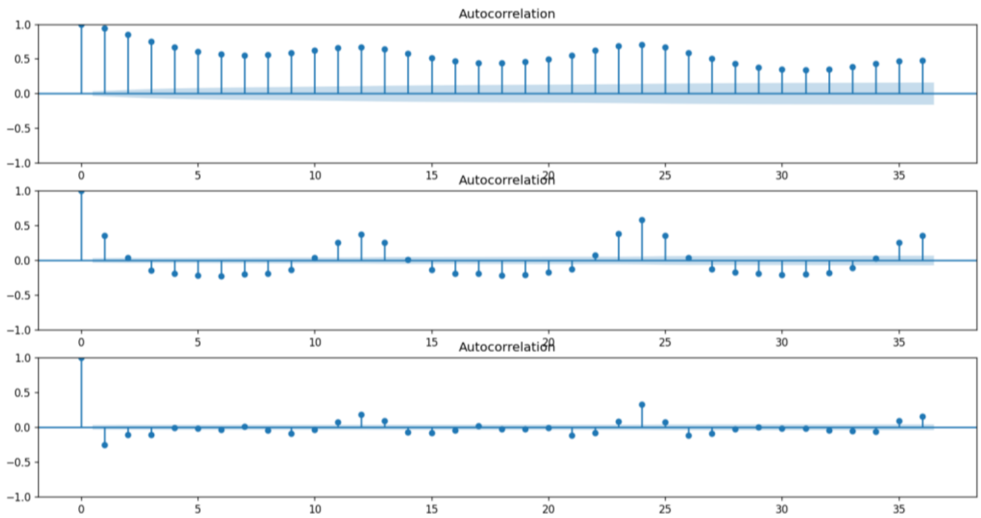

Figure 7. Example Autocorrelation Data set, 1st and 2nd Differencing (04/02/2023)

The model is autoregressive if it uses past results to predict future ones. This means that it assumes that there exists a big correlation between what has taken place and what will happen, which makes sense in the case of electricity prices. This part of the ARIMA (AR) is represented by the parameter p (p is the order or number of terms of the autoregressive model). The plotting of the partial autocorrelation function (PACF) – that gives information on the correlation between a time series and its own lagged values – can be used to identified which lagged values are relevant and which will not bring additional information to the model (Verma, Yugesh, 2022). In Figure 8 it can be seen that the order of the p should be 2, as two of the lagged values are significantly out of the limit. The other significant value that has a value around the same as the first two is the 24th value, which will be considered, if necessary, as part of the seasonal terms.

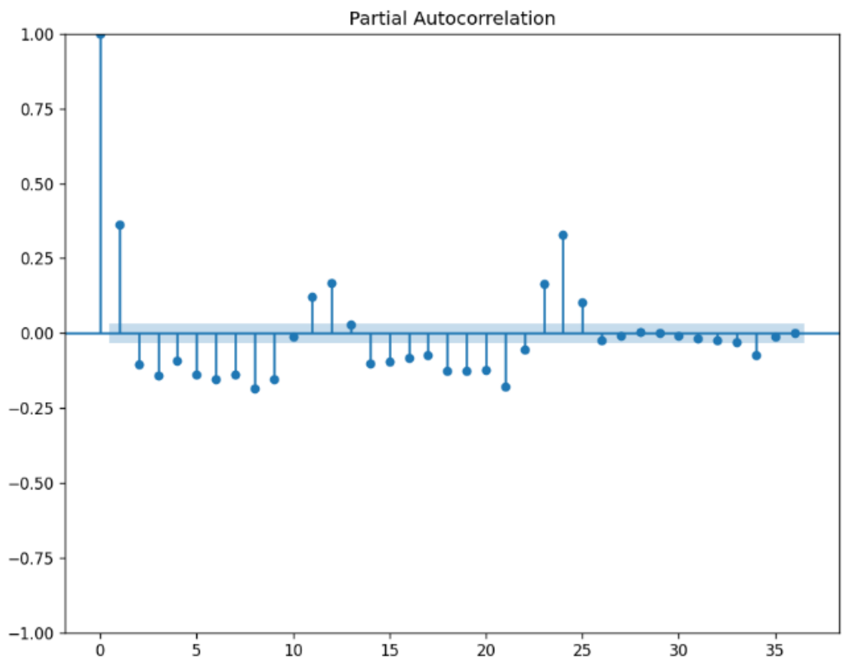

Figure 8. Example Partial Autocorrelation Function Plot (04/02/2023)

The model uses moving-average integrated because it makes use of past error terms (lagged errors). This helps capture short-term fluctuation in a time-series. This part of the ARIMA (MA) is represented by the parameter q (q is the number of lagged error terms in the model). To capture the ideal value of q for a specific data set the ACF (Autocorrelation Function) plot is used; it works in a similar way to the partial autocorrelation function previously touched upon: it gives information on the lagged terms and their significance. The ideal value of q in this example is 2, as can be seen by the number of lagged terms outside of the bounds of the limit.

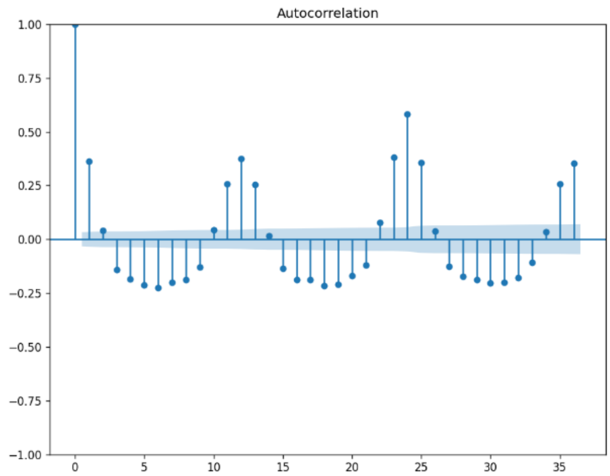

Figure 9. Example Autocorrelation Function Plot (04/02/2023)

The parameters of the model must be written in the order (p, d, q). In this case the model that should be chosen is the ARIMA (2,1,2). Nevertheless, and as mentioned above, this does not represent a seasonal model. For this reason, the model that will be considered next is the SARIMA model, where the additional S stands for Seasonal (seasonal terms are autoregressive values from the previous period).

Compared to ARIMA, SARIMA requires 4 additional orders. The first three orders are exact replicas of the orders of ARIMA but seasonal, so P is the seasonal autoregressive order, Q is the seasonal integration order and D is the seasonal moving average order. Lastly, the last order, s, is the length of the season or period (Duca, Angelica Lo, 2021). Note that SARIMA only allows one seasonality, unlike TBATS which allowed as many as was relevant.

The following process is almost identical to the methods used to get the values of p, d and q. First, it is necessary to check that the seasonal component is stationary using the Augmented Dickey-Fuller Test. The p-value is much smaller than 0,01 for this data set and the ADF is smaller than the critical values, so D can be set as 0. The parameter P is extracted from the PACF of the seasonal component. As can be seen in Figure 10, P can be set as 2. Similarly, the parameter Q can be found out from the ACF of the seasonal component and should be set as 2 (Figure 11). Lastly, the order ‘s’ must be set to 24 (there are 24 hours in a day, the seasonality we want to implement). This would result in a SARIMA (2,1,2) (2,0,2) model.

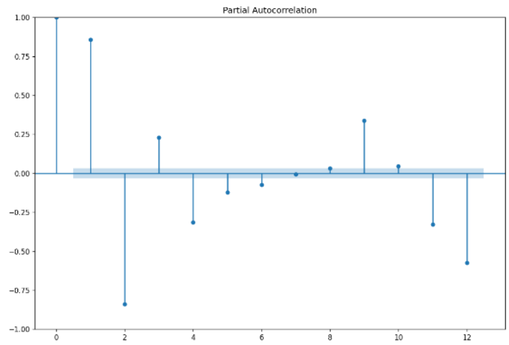

Figure 10. Example Partial Autocorrelation Function Plot of Seasonal Component (04/02/2023)

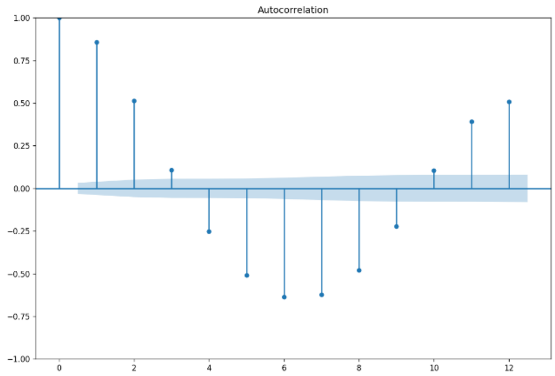

Figure 11. Example Autocorrelation Function Plot of Seasonal Component (04/02/2023)

It must be noted that values of orders higher than 3 will not be practical when the model is trained daily because the parameters are fitted through a decision tree-like optimization process and the upper limit of all parameters will be set to 3. Any value bigger than that would take too long to converge. The statistical measure used to identify the best model will be the AIC or Akaike Information Criterion. The AIC is an indicator of the quality of a statistical model. It measures the amount of information lost when using a specific model and its simplicity, so the lower it is the better.

The last addition that ARIMA allows is that of exogenous variables in the model. This brings us to the final form of the model that will be presented and tested against the TBATS model (and any other seasonal methods). The inclusion of external variables allows the model to react immediately to changes that would only be noticed in a while in lagged terms with just the SARIMA model. The inclusion of variables (indicated with the letter X at the end, making it a SARIMAX model does not modify the order of the model. The variables that will be added to the model are the same ones that worked best for the multi linear regression model, given that the way that the exogenous variables are added to the model is via linear regression.

E.5

$$y_t = c + \epsilon_t + \sum_{n=1}^{p} \alpha_n \cdot y_{t-n} + \sum_{n=1}^{P} \phi_n \cdot y_{t-sn} + \sum_{n=1}^{q} \theta_n \cdot \epsilon_{t-n} + \sum_{n=1}^{Q} \eta_n \cdot \epsilon_{t-sn} + \sum_{n=1}^{r} \beta_n \cdot x_{n_t}$$

The equation E.5 above shows the final form of the formula that creates the SARIMAX model. The c represents a constant and the ϵ_t means difference or noise. The autoregressive terms (AR), third and fourth terms of the equation, are created using the p and P parameters as stated above. It considers as many data points as those orders indicate, and they are each multiplied by a coefficient (α_n and ϕ_n). For the seasonal part, the data points used are picked from previous periods, which is why the equation uses $\cdot\, y_{t-sn}$ and not $\cdot\, y_{t-n}$. The integration part of the model (I) is not represented in the formula because it is a process that only takes place previously on the data points (as many times as d and D indicate). The moving average (MA) in the model are found in the fifth and sixth terms of the equation above. They work the exact same way as their AR counterparts except that in this case the terms are not directly past point, but past error points (ϵ_{t-n} and ϵ_{t-sn}). The rest is the same where q and Q are used instead of p and P and the coefficients are θ_n and η_n. The last term represents the exogenous variables (X) of the model, where r indicates the number of variables (x_{n_t} are those variables) used and β_n each of their coefficients (Artley, Brendan, 2021). It is the same formula that we saw in equation E. 1 for multi linear regression models.

In Figure 12 the prediction of a period of 24 hours using SARIMAX of the same date that was represented in Figure 5 can be graphed. The values of the RMSE are 14,57 – SARIMAX – and 22,02 – TBATS– respectively. Once again, 24 points are not enough to reach any conclusion. However, the difference between prediction and actual values as well as general shape is better when using the SARIMAX model in this particular instance.

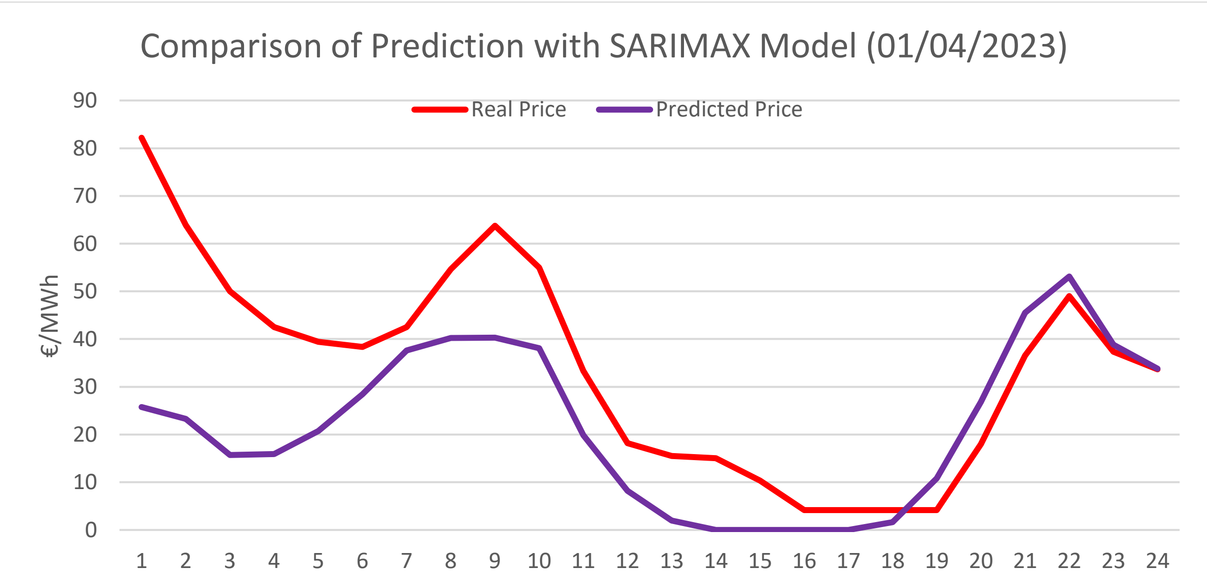

Figure 12. Example of Prediction with SARIMAX Model (01/04/2023)

The third option for a seasonal model that will be analyzed is called the MSTL method. MSTL, or Multiple Seasonal-Trend decomposition using Loess, is very similar to TBATS in the sense that it decomposes a time series into trend, seasonal and residual components (as seen in equations E.2 and E.3 for TBATS). The difference between the methods is that while TBATS' main purpose is to train the forecasting model MSTL, is more focused on decomposing the time series. It is generally recommended due to its accuracy and lower execution time than other similar methods.

Also known as local regression or moving regression, LOESS (Locally Estimated Scatterplot Smoothing) is a method for smoothing data by fitting simple models to localized subsets in the data. It can handle nonlinear relationships between different variables, and it uses a function to assign more weight to nearby observations and less weight to distant ones. This is the method used in MSTL to decompose seasonalities in a time series (Manani, Kishan, 2022). The same way the TBATS needs the user to choose the period of the seasonality or seasonalities the time series has, as does MSTL. If the data being used is non-stationary, then – and only then, because noise would be needlessly added otherwise – it requires a Box Cox transformation and it would have to be indicated to the model.

In the example plot (Figure 13), there are five months of the time series (3623 data points) represented and broken down using an MSTL model that presents weekly and daily seasonalities. The weekly seasonality is clearly visible (the daily one would be as well if there was a zoom in, it is not as apparent due to the scale). For the first time we can empirically prove that both seasonalities are present in the data. Using the MSTL model the forecast of the same day already represented with TBATS and SARIMAX is made, as seen below (Figure 14).

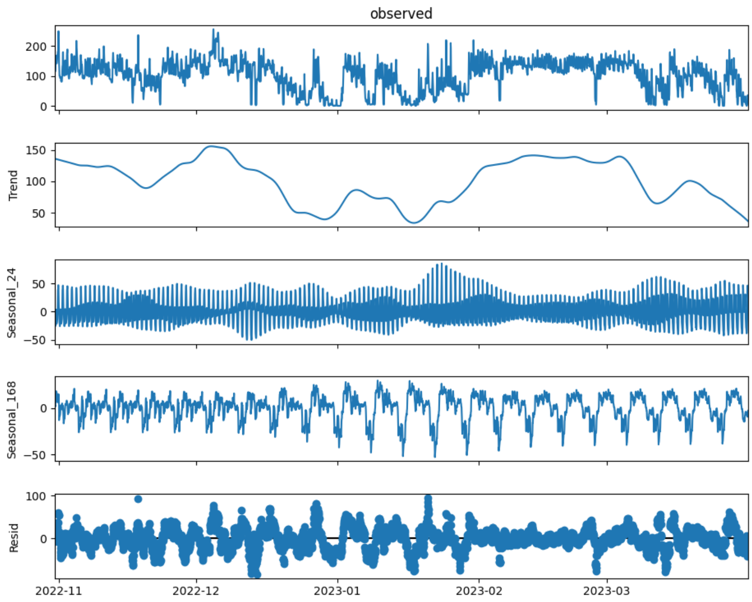

Figure 13. Example of MSTL Model with Daily and Weekly Seasonalities (Nov. 22 – Mar. 23)

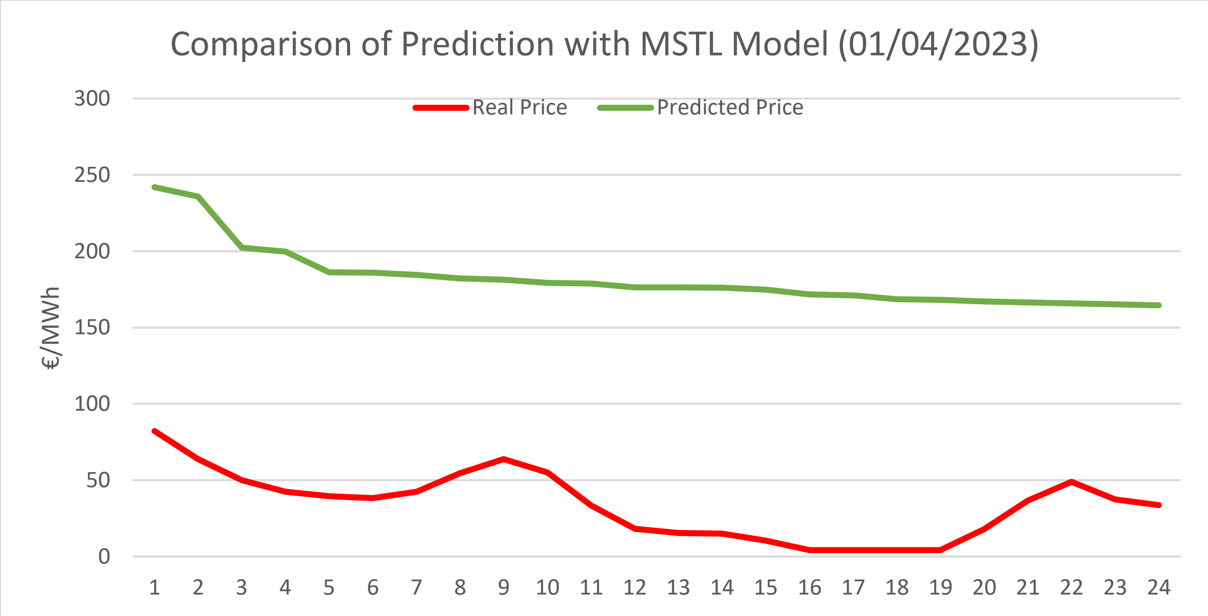

Figure 14. Example of Prediction with MSTL Model (01/04/2023)

The last method we will delve into that deals with seasonality and autoregression is called Prophet. There exist some differences between TBATS and Prophet, although they are similar when used as we are in this project. Prophet only works for additive models, which does not affect anything since we have already established that our time series is additive. TBATS models residuals using ARIMA while Prophet uses normal distribution. There are more small differences – for example, TBATS is more computationally intensive –, but the only one that is relevant to our models is this one: Prophet does not require specific lengths of seasonal periods; they are automatically detected (Elseidi, Mohammed, 2022) (Facebook Open Source, 2022). We will use this as an opportunity to confirm the seasonalities we have observed in the data.

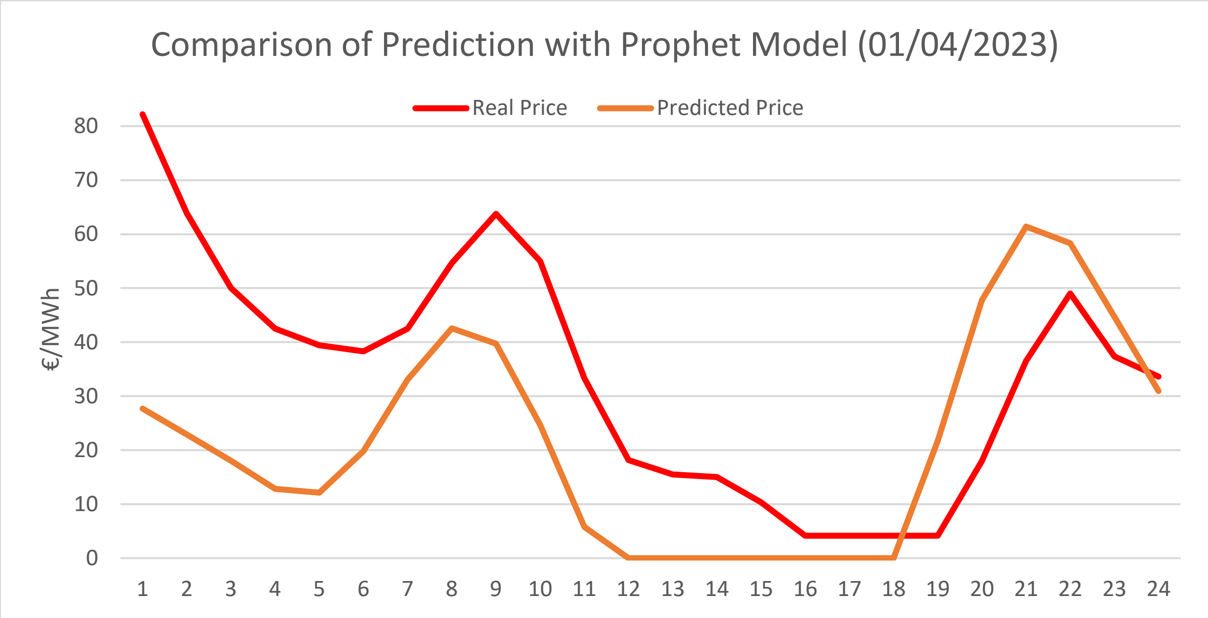

Figure 15. Example of Prediction with Prophet (01/04/2023)

The model does indeed find daily and weekly seasonality as expected. In Figure 15 we can see the final prediction for 01/04/2023 that has already been made thrice with the other seasonal/autoregressive models.

These four are not the only methods that exist, but they are the most used because of their simplicity, execution time and accuracy. At this point, there is enough information to compare them all and make a choice regarding which of them will be selected to be used. It is not surprising to find that to predict the price of electricity, a variable that depends on so many factors, the best model is the one that includes variables to help make those changes faster. In other words: in Table 14, analyzing 720 predictions of data points, we can see that the best seasonal/autoregressive model is the SARIMAX model.

Model name	*RMSE*
TBATS	*30,757*
SARIMAX	*24,575*
MSTL	*63,503*
Prophet	*29,023*

Table 14. Table with RMSE values for all Seasonal/Autoregressive Models (March 2023)

3.3.3 LOGISTIC REGRESSION

We now have our baseline models. As previously outlined, there is a variable that is not quite like the others that has not been included in any model so far. That variable is called 'Closing or Marginal Technology' and it has a major difference with the rest of the variables that have been introduced up to this point. The rest of the variables are predictions or information that is available before the forecast of the daily prices is to take place. However, it would be impossible to know this information in advance. If this is to be used as a variable some work must be done prior to that.

The first thing that needs to be done is to find a way to make the data useful. From the official OMIE page, the information is listed using acronyms that can be seen below (OMIE, 2023).

Acronym	***Technology/Energy***
BG	*Hydraulic Pumping*
HI	*Hydro Energy*
MIE	*Mibel imported from the Spanish electrical system*
MIP	*Mibel imported from the Portuguese electrical system*
RE	*Renewable Energies*
TCC	*CCGTs*
TER	*Solar Thermal Energy*
NU	*Nuclear Energy*

Table 15. Technologies available in Spain and their Acronyms

The information can be processed using as many variables as technologies there are in the table above, making each of them act like a binary variable that is 1 when a technology is the most expensive in each hour and 0 otherwise. This works well because it also allows for more than one closing technology – which is possible.

Once the variables have been defined, we have to decide which ones are: (1) feasible and (2) useful as variables. This is a simpler question than one might think at first glance. To start, both MIP and MIE must be removed because neither of them would have a value of 1 at any point, making them unfeasible as variables. Any technology that has long stretches of time where it does not close will not be a possibility either due to the limitations that regression models have in Python. The information used for each forecast is that of the past month. If at any point, there was over a month of time for a variable to have a value of 1 again (for it to be the closing technology again) the program would not be able to function as the variable would always have the same value and it would be impossible to calculate a coefficient for the variable. This eliminates NU and TER.

The rest of the variables ('BG', 'HI', 'RE' and 'TCC') will be added to the ones from the Multi Linear Regression Model and checked as was done before. Firstly, for multicollinearity (Table 16). Since all the variables have a VIF < 5 and are therefore not collinear, we can test a model with all the variables.

Variable name	*VIF*
Price D -1	*2,03*
Price W -1	*1,78*
Auction Gas Price	*1,27*
CO2 Price	*1,62*
Market Thermal Gap	*2,80*
Wind Energy	*1,46*
BG	*1,48*
HI	*2,25*
RE	*2,58*
TCC	*1,74*

Table 16. Table with VIF values for all variables + additions (Jan. – Apr. 2023)

In Table 17 the p-value clearly indicates that 'HI' is not a good variable to have in the model. It is probably due to the fact that hydro generation can be the closing technology in very different circumstances. For example, hydro energy can be the closing technology when runoff-the-river generation is setting the price, which will be therefore very low. However, hydro generation can also be the marginal technology when dam-based generation is setting the prices, what can be very high because, in this case, the water value will be the cost of substitution of the thermal generation.

It is hence removed from the model. The resulting combination of variables will be applied to create two new models. Why not just replace the old models with these new ones with all the variables? Simply because, as explained above, the variables being added will not be 'information' before the program is run. For them to work, they will have to be somehow inferred. The specific method thorough which that will be accomplished will be expanded on in the following paragraphs. Because of the possible unreliability of the new variables, we will keep both the old and the new models (multi linear regression and SARIMAX with old and new combinations of variables).

Variable name	*p-value*
Price D -1	*1,01e-111*
Price W -1	*1,03e-12*
Auction Gas Price	*0,0055*
CO2 Price	*1,80e-93*
Market Thermal Gap	*7,82e-276*
Wind Energy	*3,08e-63*
BG	*0,0062*
HI	*0,98*
RE	*6,08e-28*
TCC	*0,16*

Table 17. Table with p-values for all variables + additions (Jan. – Apr. 2023)

Logistic regression is a type of statistical model generally used for the purposes of classification or prediction. It estimates the probability of something happening, which is why it can be used in classification. For example, to determine whether a person with certain symptoms is sick with a specific illness. Even though the model can be run to estimate a probability or with even more complex purposes, it is frequently used as a binary variable that opts for a 0 or 1 depending on the probability it predicts. We will try this method – called binary logistic regression – as an option for the prediction of the variables that will not be available when doing the forecast. It is relevant to know that logistic regression can be prone to overfitting, so one should be careful with the number of dependent variables fed to the model.

To begin with, the model can be tried using the same variables as before, excluding 'RE', 'TCC' and 'BG' for obvious reasons. Table 18 has information on the p-values of that attempt with the variable 'RE', where it is apparent that all of them are good additions to the model. Table 19 has the same information but for the variable 'TCC'. In this case some of the variables should be removed ('Auction Gas Price' and 'Market Thermal Gap') as presented in the results in Table 20. The information from these variables is most likely still being taken into account historically as it seems relevant to the variable 'TCC'.

Lastly, Table 21 shows the p-values for variable 'BG', where 'Wind Energy' must be eliminated as well (Table 22).

Variable name	*p-value*
Price D -1	*0,0002*
Price W -1	*0,0007*
Auction Gas Price	*0,0002*
CO2 Price	*0,0013*
Market Thermal Gap	*0,0000*
Wind Energy	*0,0000*

Table 18. Table with p-values for 'RE' Logistic Regression, 1st Attempt (Jan. – Apr. 2023)

Variable name	*p-value*
Price D -1	*0,0000*
Price W -1	*0,0000*
Auction Gas Price	*0,4725*
CO2 Price	*0,000*
Market Thermal Gap	*0,4785*
Wind Energy	*0,0000*

Table 19. Table with p-values for 'TCC' Logistic Regression, 1st Attempt (Jan. – Apr. 2023)

Variable name	*p-value*
Price D -1	*0,0000*
Price W -1	*0,0000*
CO2 Price	*0,000*
Wind Energy	*0,0000*

Table 20. Table with p-values for 'TCC' Logistic Regression, p<0,05 (Jan. – Apr. 2023)

Variable name	*p-value*
Price D -1	*0,0100*
Price W -1	*0,0001*
Auction Gas Price	*0,0040*
CO2 Price	*0,0000*
Market Thermal Gap	*0,0000*
Wind Energy	*0,3296*

Table 21. Table with p-values for 'BG' Logistic Regression, 1st Attempt (Jan. – Apr. 2023)

Variable name	*p-value*
Price D -1	*0,0061*
Price W -1	*0,0000*
Auction Gas Price	*0,0061*
CO2 Price	*0,0000*
Market Thermal Gap	*0,0000*

Table 22. Table with p-values for 'BG' Logistic Regression, p<0,05 (Jan. – Apr. 2023)

The three models discussed above will be implemented to add the information necessary to their corresponding variables before they are introduced into the Multi Linear Regression and SARIMAX models to forecast. Each Logistic Regression Model will predict a probability that will be used as 0 or 1 variables (binary variables). The information from the predictions is only necessary for the 24 hours of the forecast period, the rest is historical information. The variables will be introduced along with all other variables discussed previously into both models (Multi Linear Regression and SARIMAX).

Before moving on to the creation of some sort of recommender system or a method that helps identify the "best" forecasting method, we will calculate the average of the regression and SARIMAX methods. Combining forecasts is a simple way of improving the quality of the results when there is no "best" method. Not only is it not sensitive to extreme values, decreasing the variability of the forecasting accuracy, but resulting in a method that is more accurate than both models it comes from. Furthermore, it is not necessary to average a lot of models together to get these results. On the contrary, "a simple average of the forecasts often does better than more complex combining schemes" (Jose, Victor Richmond R., and Robert L. Winkler, 2007).

The result of this process consists of four independent models (two multi linear regression models and two SARIMAX models) and four "average" models (product of averaging means of each combination of SARIMAX and multi linear regression together. Next, we will study the results of each of the models.

3.4. – NUMERICAL IMPLEMENTATION AND ANALYSIS OF THE RESULTS

The newly developed models have been applied to the processed data to predict a period of 7 months (24 hours at a time). This amounts to 5088 hourly predictions (212 days), which is enough information to reach conclusions on the accuracy and efficiency of the forecasting models.

Model name	***Model description***	***RMSE***
SARIMAX 1	*SARIMAX w/o 'Closing Technology' variables*	*17,164*
SARIMAX 2	*SARIMAX w/ 'Closing Technology' variables*	*17,098*
MLR 1	*Multi Linear Regression w/o 'Closing Technology' variables*	*17,263*
MLR 2	*Multi Linear Regression w/ 'Closing Technology' variables*	*17,423*
AVG 11	*Average of SARIMAX 1 and MLR 1*	*16,067*
AVG 12	*Average of SARIMAX 1 and MLR 2*	*16,058*
AVG 21	*Average of SARIMAX 2 and MLR 1*	*15,892*
AVG 22	*Average of SARIMAX 2 and MLR 2*	*16,048*

Table 23. Comparison RMSE Values All Models (Nov. 2022 - May 2023)

From Table 23 it is difficult to reach any conclusions. The SARIMAX model with the three additional variables is more accurate than the one without them, but the opposite happens in the case of the MLR models. The effect of these variables is therefore not conclusively an improvement. Only two things can be said from just this information:

1) It is confirmed that averaging the models is an effective way of improving the accuracy of the models

2) The SARIMAX models are overall slightly better than the MLR models

We can see that the RMSE is consistent throughout the models, which is a good sign. It means that there were no big mistakes made when choosing variables and orders in the models. This information is not enough to understand the trend of the models nor how well they work. We will be studying different graphs and information from the predicted time series to interpret the results before we attempt to decide regarding the best model and/or the course of action that has to take place to choose it.

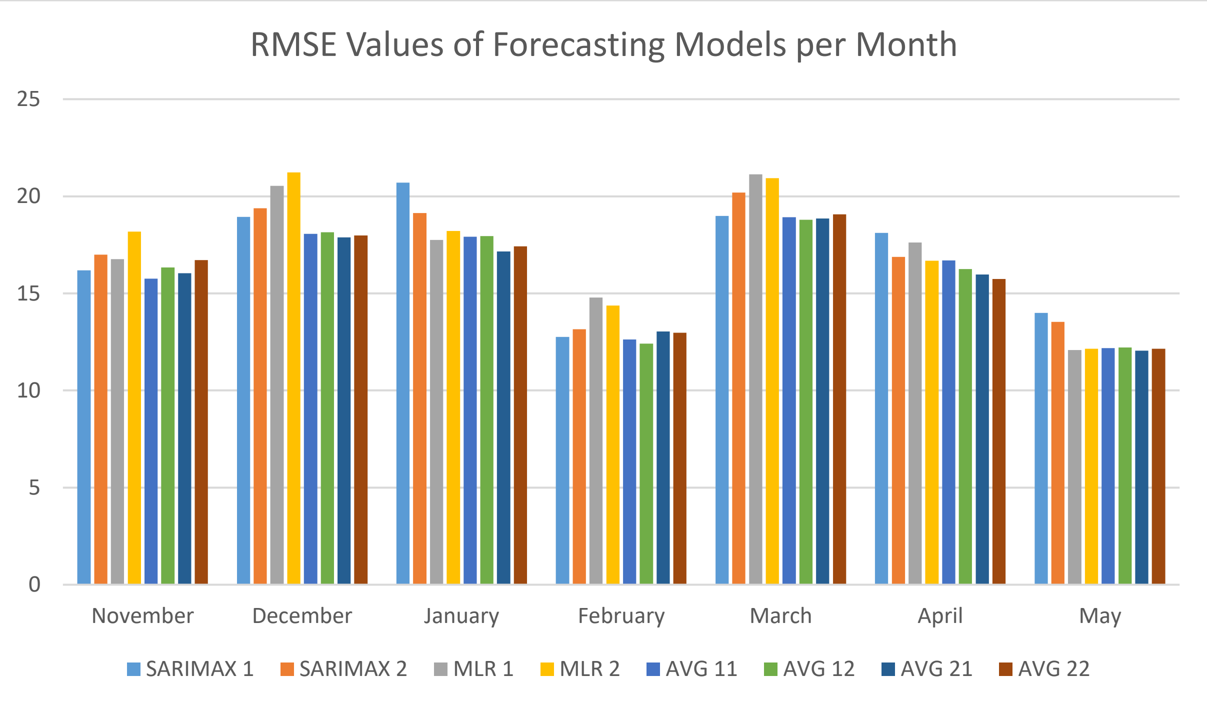

Figure 16. Comparison of RMSE Values of Forecasting Models per Month (Nov. 2022 - May 2023)

To check the consistency of the results it is necessary to break them down and look at them through smaller periods of time. In Figure 16, the predictions are divided by month. The graph shows that the AVG models are consistently more accurate than the other models. Throughout the months, all AVG models are equally "good" at predicting, meaning that there is no clear better model. However, SARIMAX 1 and both MLR models are the worst models several months in a row: SARIMAX 1 January, April, and May; MLR 1 February and March; and MLR 2 November and December. If it were necessary to choose one model outside of the AVG models, SARIMAX 2 is the best choice. Out of all the models, the best option is AVG 12. Not only is it the one with the lowest overall RMSE (Table 23), but it is also consistently in the top 2 best models: best model December, January, and May; second best November, March, and April.

As mentioned before, the RMSE by itself is not enough to tell a good prediction from a bad one. The shape of the graph must be consistent with the expected shape one as well for it to be considered acceptable. To check this, the average price per period of each model as well as the real price are represented diving the data once again by month: November in Figure 17, December in Figure 18, January in Figure 19, February in Figure 20, March in Figure 21, April in Figure 22 and May in Figure 23. Although there is a slight tendency to predict the values higher than the actual ones, especially visible in Figure 19 and Figure 21, the general tendency is good. The shape of all the graphs is the same in all predictions and very similar to the expected shape. All months have a daily shape with two peaks, one in the morning and another one in the afternoon, as well as a valley in between them that becomes more prevalent as the months pass by.

The results from the graphs are consistent with the RMSE results in Figure 16, as it is possible to see that as the RMSE values of the forecasting models decrease, so does the difference between the predictions and the real values in the graphs (Figure 22 and Figure 23). The graphs with the biggest differences between the models and the actual prices show more clearly which are the models that are more accurate, and which are routinely farther from what they should be. The more we can see of a specific model, the worse it is. For example, and as expected MLR 2 is very distinguishable during the months of December and March (Figure 18 and Figure 21) where it is the model with the highest RMSE.

In the graphs it is noticeable that the smallest differences occur during the times between peaks and valleys where the price is changing more drastically. The biggest difference on the other hand happens in those peaks and valleys. The months that have the shortest times spent in peaks and valleys, where the price changes most (like April and May) are the ones where the predictions are the best. This suggests that there might exist some variable that could be added that would explain the values of peaks and valleys better than the ones present.

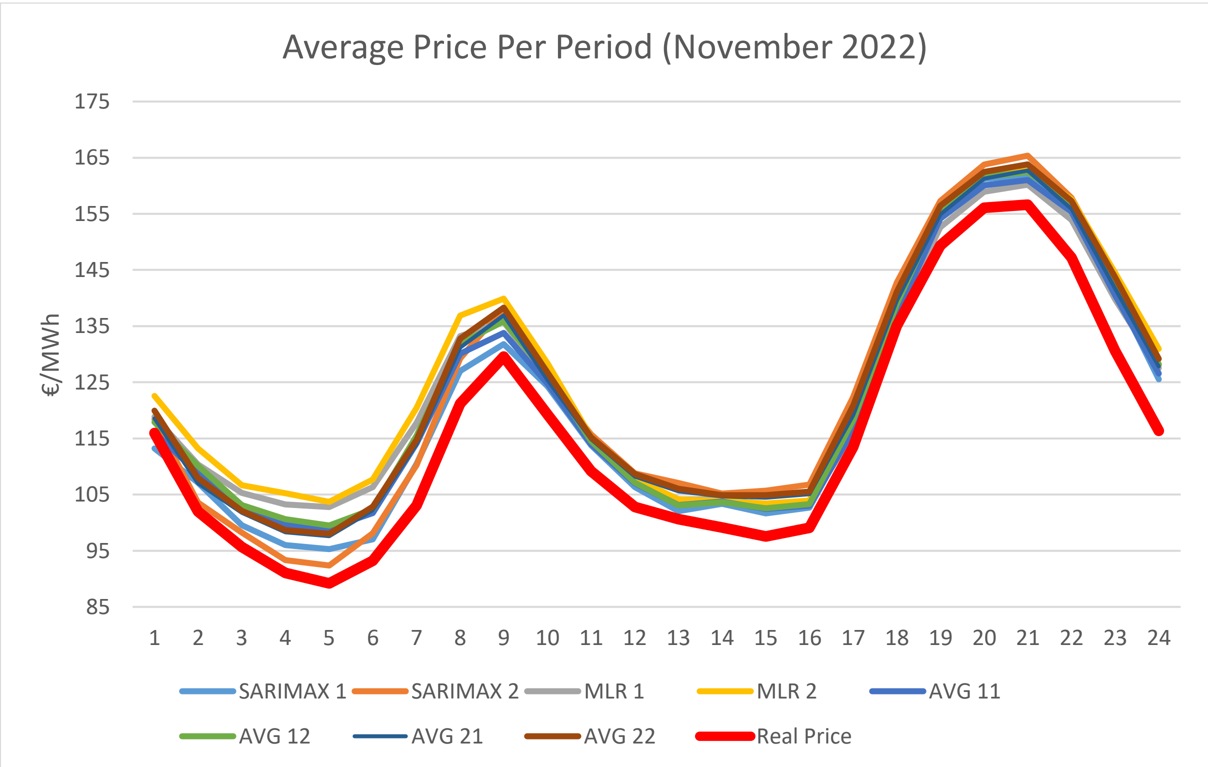

Figure 17. Comparison Average Hourly Price Forecasting Models (November)

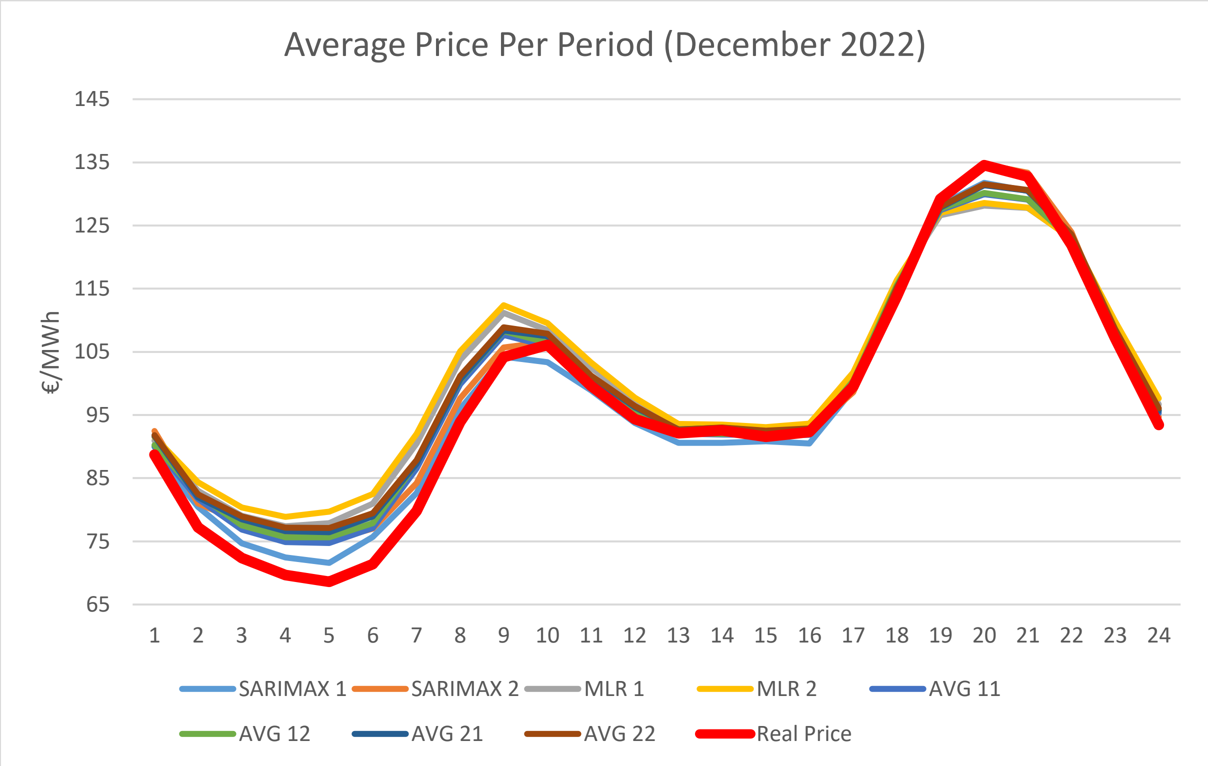

Figure 18. Comparison Average Hourly Price Forecasting Models (December)

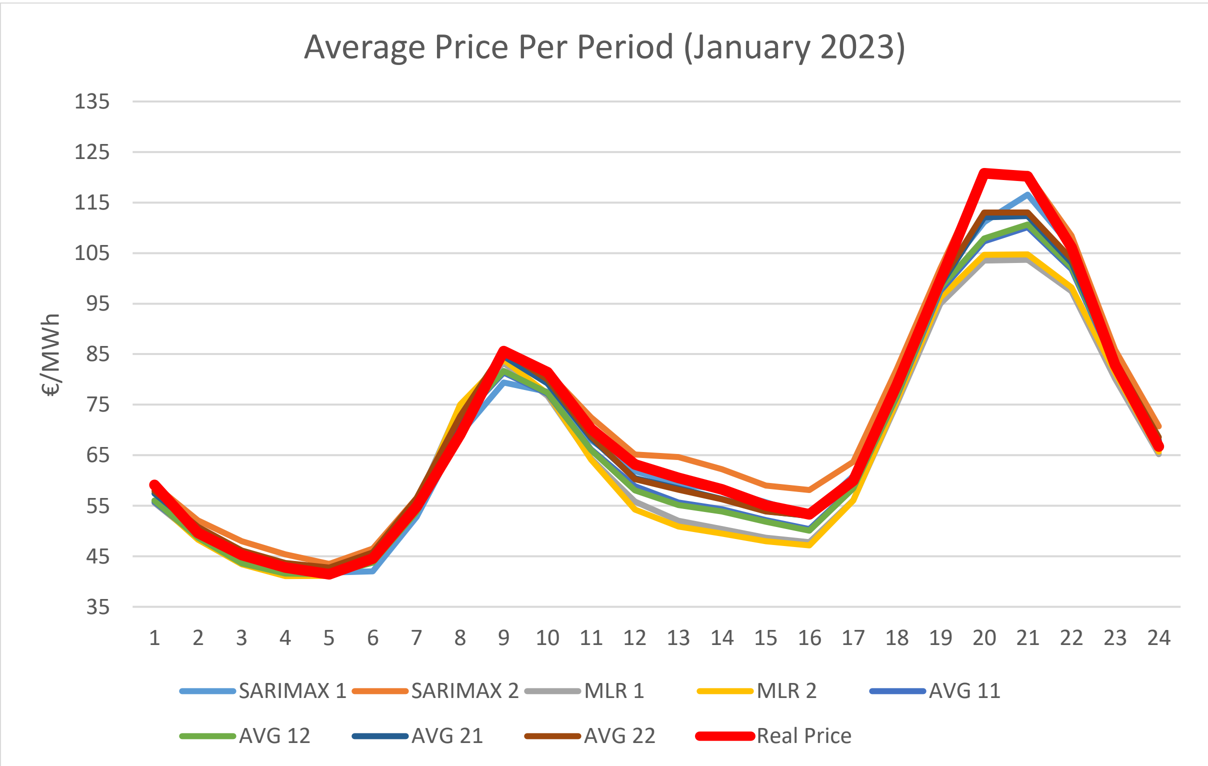

Figure 19. Comparison Average Hourly Price Forecasting Models (January)

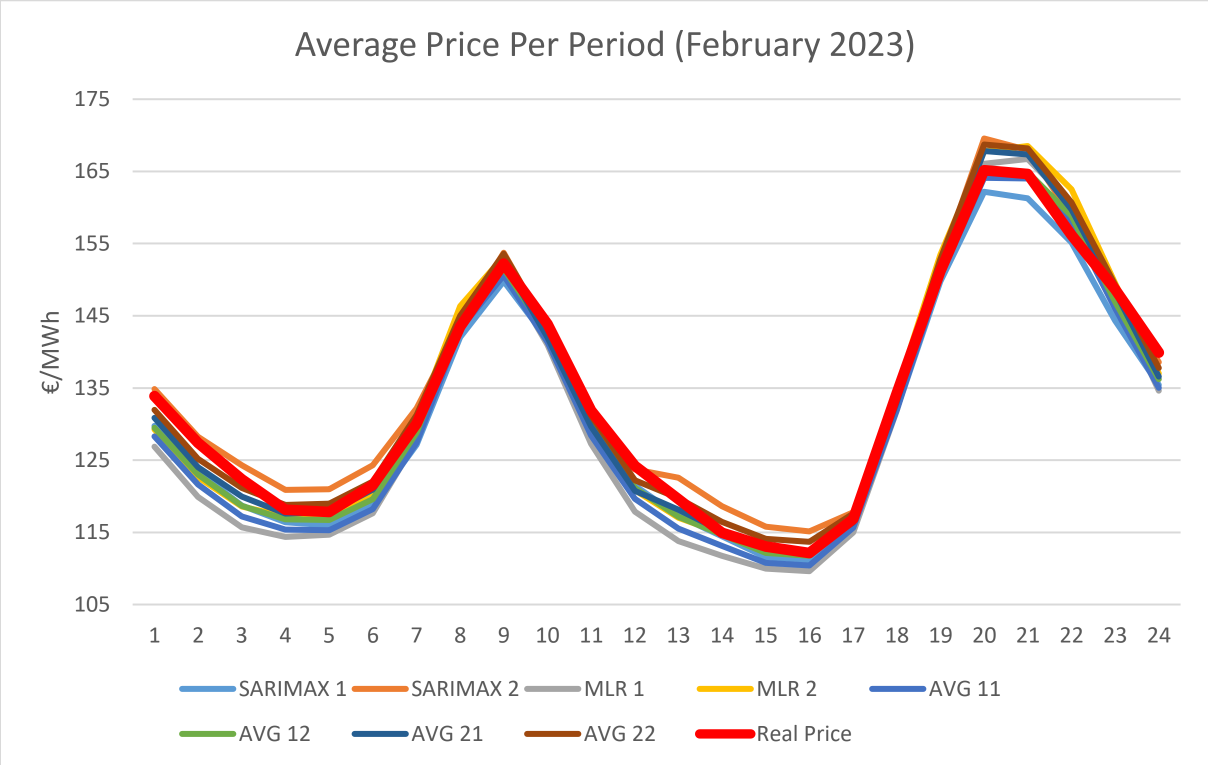

Figure 20. Comparison Average Hourly Price Forecasting Models (February)

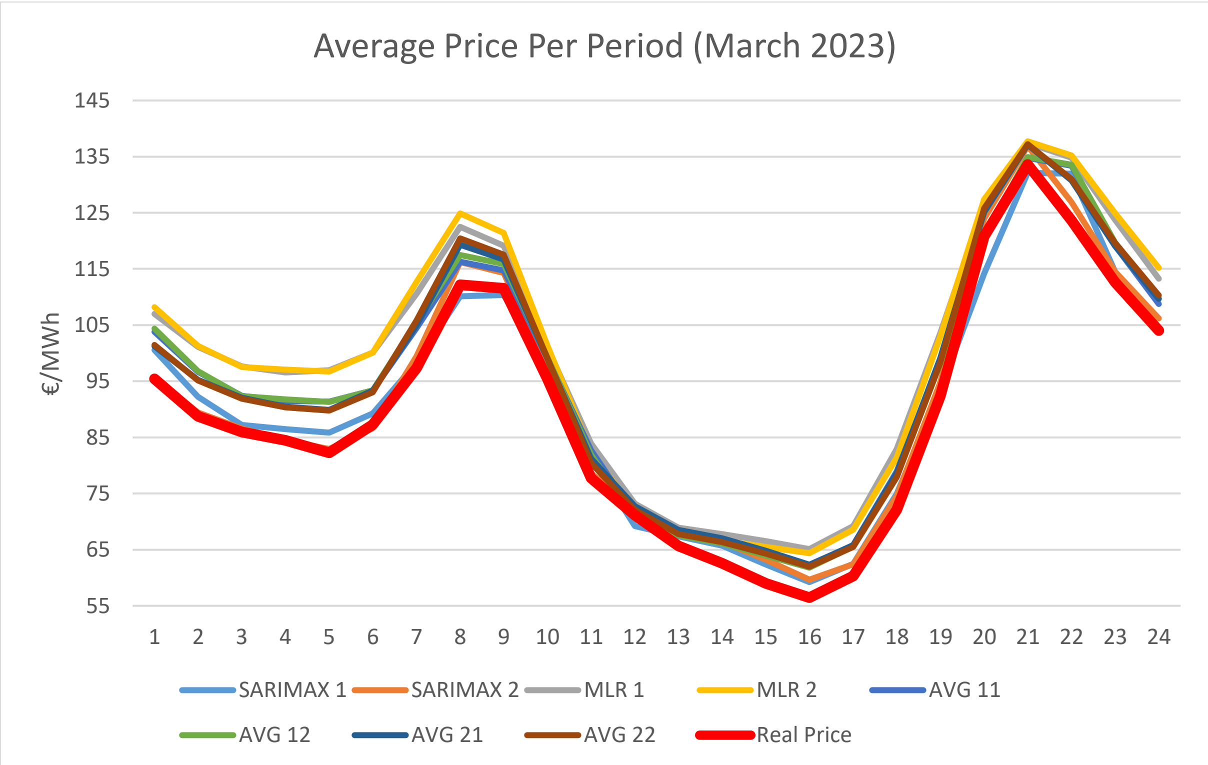

Figure 21. Comparison Average Hourly Price Forecasting Models (March)

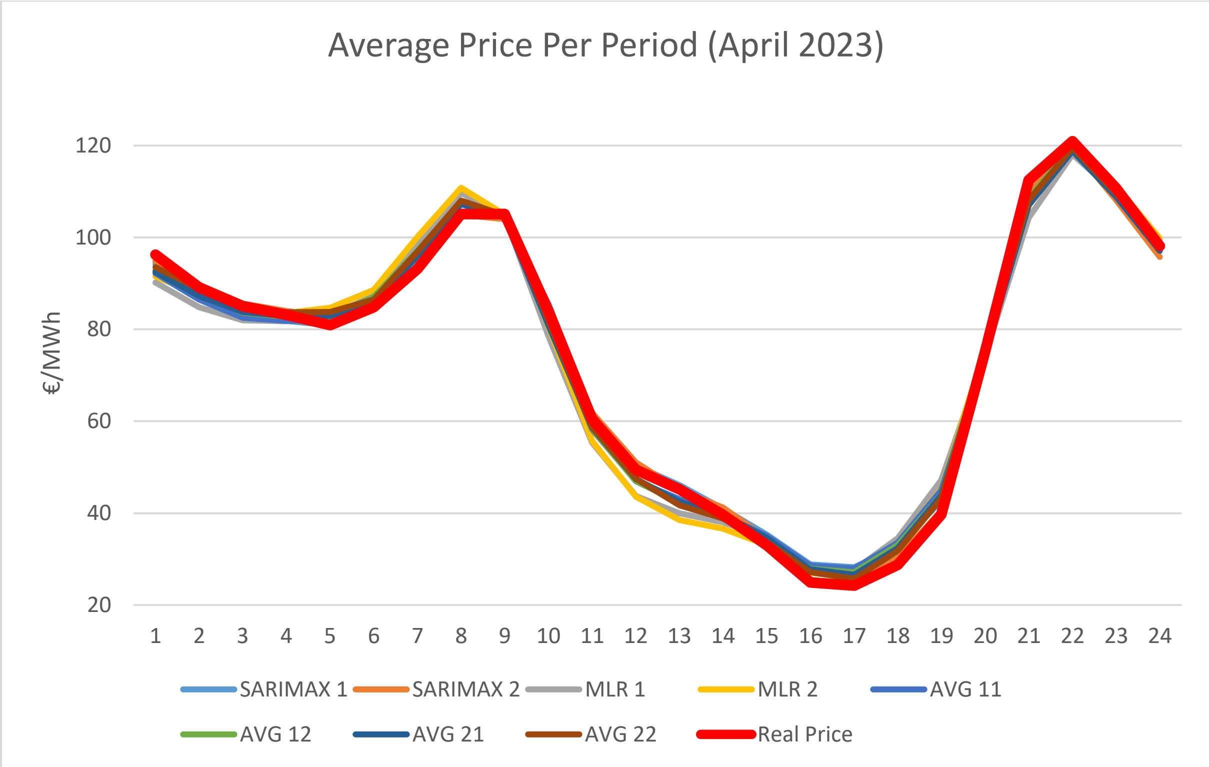

Figure 22. Comparison Average Hourly Price Forecasting Models (April)

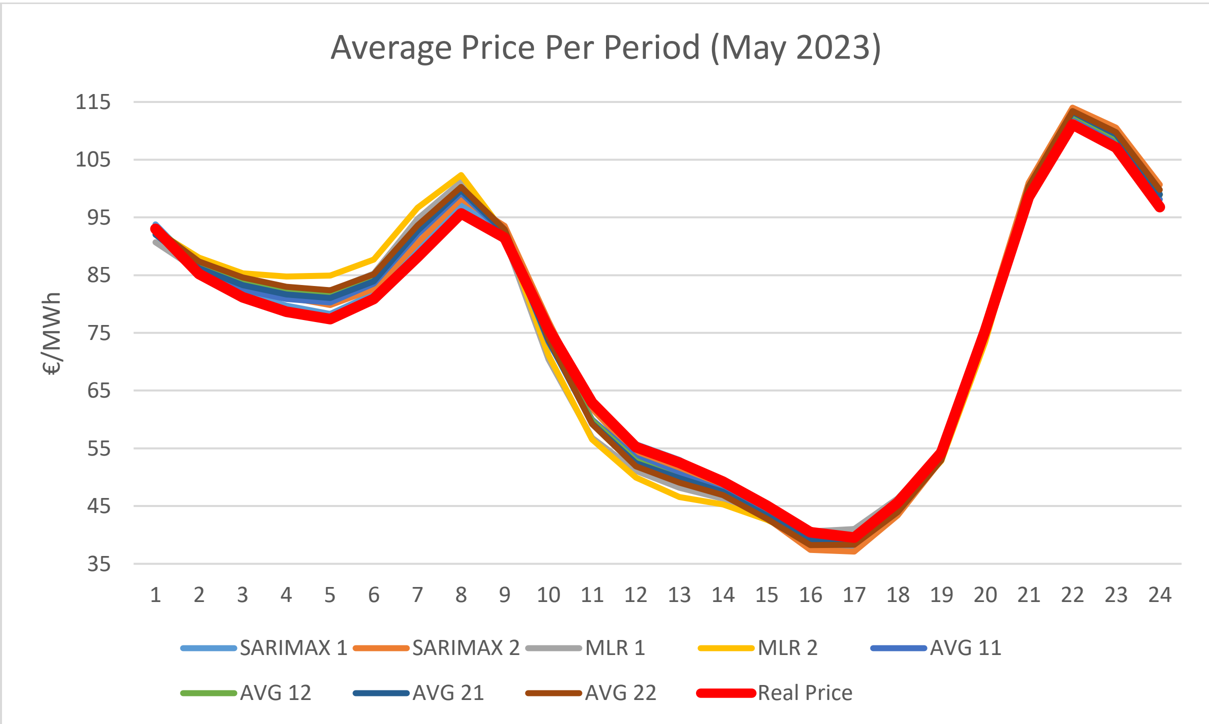

Figure 23. Comparison Average Hourly Price Forecasting Models (May)

The following graph illustrates well why the AVG models are better than the SARIMAX and MLR models: while it is true that the latter are the best possible models during longer periods of time, for each of the four models it is even more likely that they will be the worst possible model. On the other hand, even though there is a lower probability that the chosen model will be the best if one goes with an AVG model, the chances of it being the worst are almost nil. The best non-AVG, SARIMAX 2, has the highest percentage of best predictions and the smallest difference between the percentage of best and worst predictions.

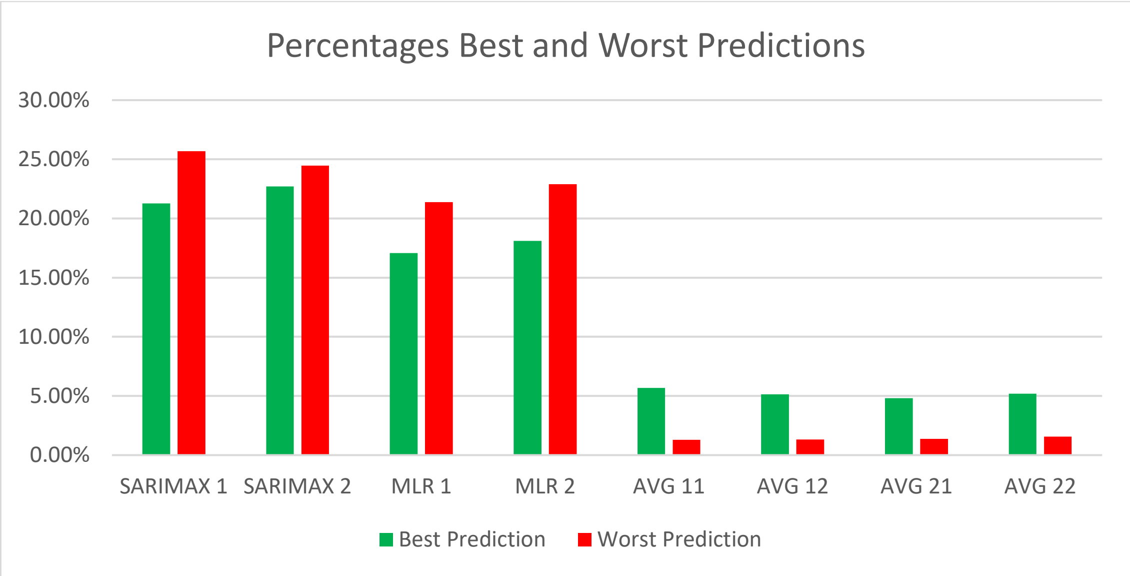

Figure 24. Comparison of the Best and Worst Prediction (%)

It is foreseeable that the standard deviation and the variance of the errors of the different models will be somewhat proportional to the results in red in Figure 24. In Table 24 we can confirm that. All AVG models have the lowest values while the SARIMAX models have the highest.

Model name	***Standard Deviation***	***Variance***
SARIMAX 1	*23,103*	*533,741*
SARIMAX 2	*22,853*	*522,244*
MLR 1	*22,942*	*526,347*
MLR 2	*21,375*	*456,902*
AVG 11	*21,332*	*455,036*
AVG 12	*23,074*	*532,432*
AVG 21	*21,064*	*443,681*
AVG 22	*21,137*	*446,756*

Table 24. Standard Deviation and Variance of Errors for each Model

Before moving on with the models, we need to check for one last thing: forecast bias. Forecast bias occurs if there is a tendency to predict with positive errors/differences (over-forecasting) or with negative errors/differences (under-forecasting). The presence of a forecast bias implies there is an issue with the method. If it is present in most of the predicted data, the average of the forecasted values would not be equal to the average of the expected values. Instead of using a specific metric we will verify this graphically, looking at graphs of the errors in the different models. For the sake of clarity and brevity two models are represented in each of the graphs (Figure 25, Figure 26, Figure 27 and Figure 28).

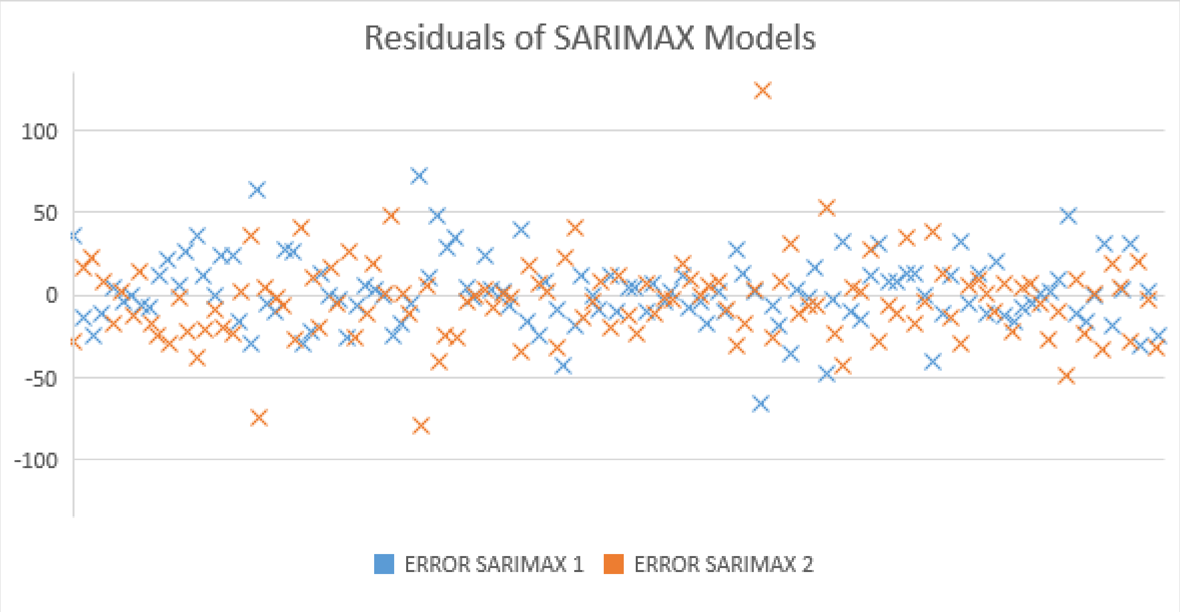

Figure 25. Residuals of Predictions with SARIMAX Models

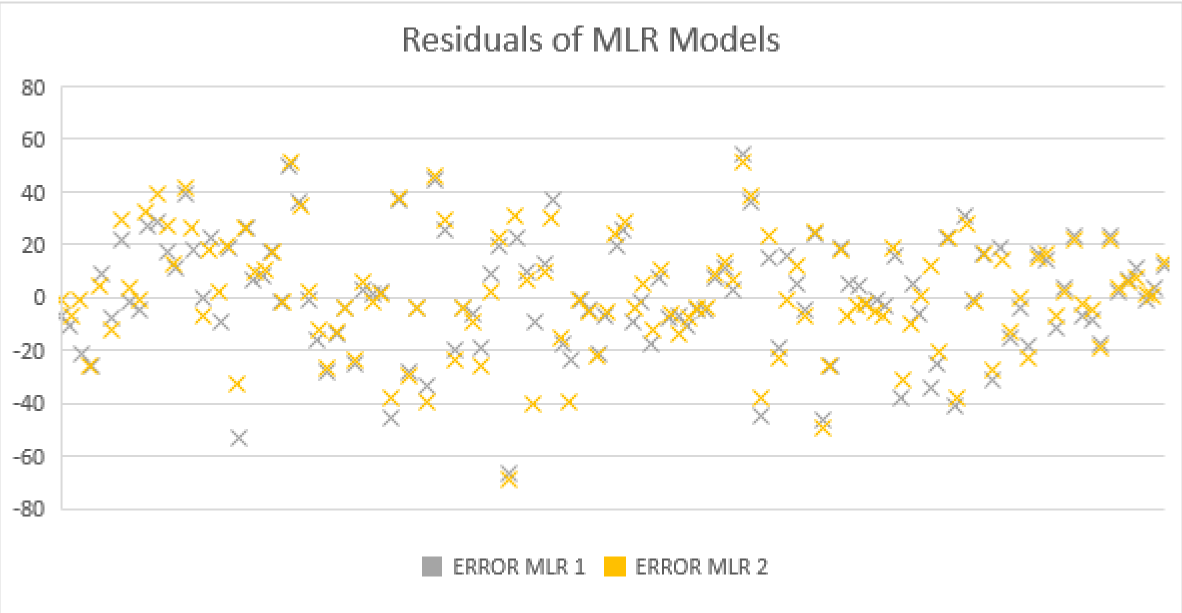

Figure 26. Residuals of Predictions with MLR Models

Figure 27. Residuals of Predictions with AVG 11 and AVG 12 Models

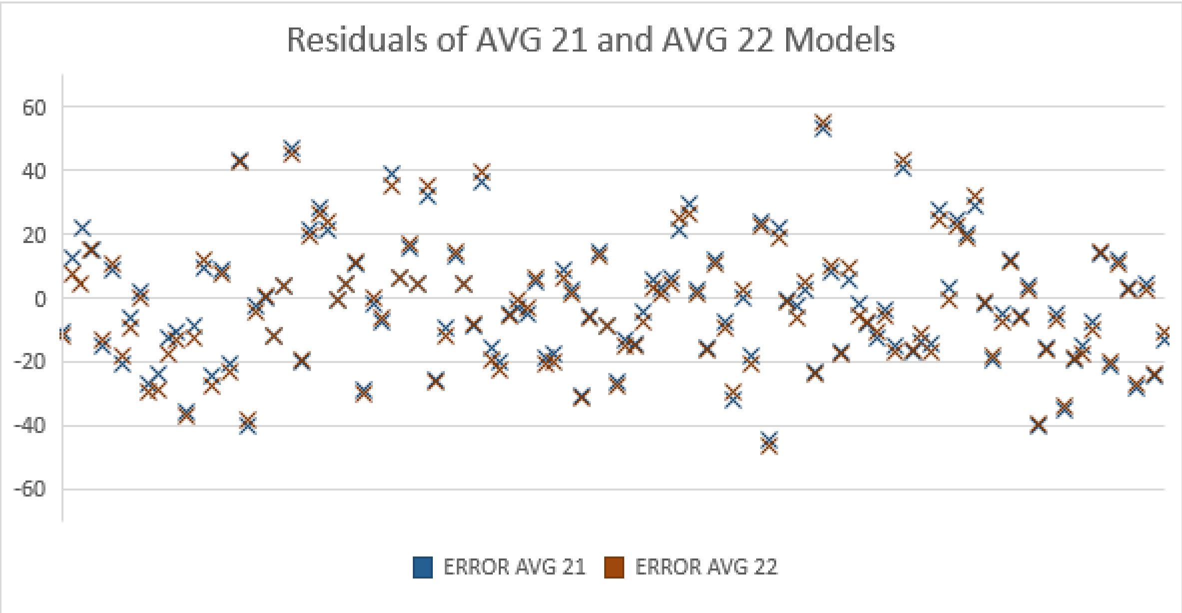

Figure 28. Residuals of Predictions with AVG 21 and AVG 22 Models

There is no higher density of errors in the positive nor negative side in any of the graphs of the residuals so there is no forecast bias of any type. Since there is also no pattern in the errors of any of the models, we can bring to a close the analysis of the results.

3.4. – MODEL CODING AND COMPUTING

The creation of the model and its format is a very important part of the project. The project has been coded using Python and SQL – where the datasets were stored, as mentioned previously. Since the project is focused on the study of the accuracy and performance of different models, it does not focus on the work related to the creation of the code itself nor its execution. For this reason, this section is centered around explaining the structure and methodology followed when creating said code, along with some comments on the computation of the models.

The idea behind the project is not just building models with the sole purpose of comparing and contrasting them, but also use them in practical applications in the near future. In order to make the model easy to understand and follow, the structure adopted was the following:

	Description	***Average Computational Time***
Data Loading	*Inputting of data from SQL using SQL commands*	*50s*
Data Standardization	*Creation of common indexes, transformation of data into correct units*	*<1s*
Data Auditing	*Examination of the data, removal of necessary data points*	*<1s*
Data Processing	*Organization and structuring of data into tables, creation of additional variables ('Closing or Marginal Technology' variables)*	*1s*
Feature Engineering	*Preparing the features so they are suitable to the model (removal of duplicates, fill in NaN values...)*	*2s*
Catch Up Function	*Path that will be triggered if the forecast for any previous date is not available*	-
Train Model	*Training of Multi Linear Regression, SARIMAX and Logistic Regression Models*	*15 mins*
Forecast Model	*Forecast of all models for 24 hours*	*<1s*
Insert Forecast	*Saving of the forecast into SQL database for future use*	*<1s*
Recommender System	*Selection of best model*	*5s – 1 min*

The model was tested for all previously mentioned models, as well as the recommender systems that will be described in the next chapters. Due to its nature as an optimizable model when finding orders, training ARIMA models takes a significant amount of time. The same can be said for all other seasonality-based models, that although not as slow still took some time. The information studied over the course of this project pertains to predictions from November 2022 till May 2023. This amounts to 5087 individual hours (not 5088 because of the time change to Daylight Saving Time on 26/03, which was considered by the model) for each model. This process added up to a total of over 400 computational hours.

CHAPTER 4. – MODEL SELECTION

4.1. – SELECTION METHODOLOGY

In this section we will investigate very different methods of approaching the concept of a selector system. The goal is to study two types of systems: systems that choose one of the available models (recommender systems) and systems that create a new prediction from the given models (ensemble averaging).

4.1.1 – RECOMMENDER SYSTEMS

The first and simplest model will be referred to as the 'D -1' Recommender Method (day minus 1) and it will choose as 'ideal' model the one that had the best prediction 24 hours before each point. The concept is graphically showed below:

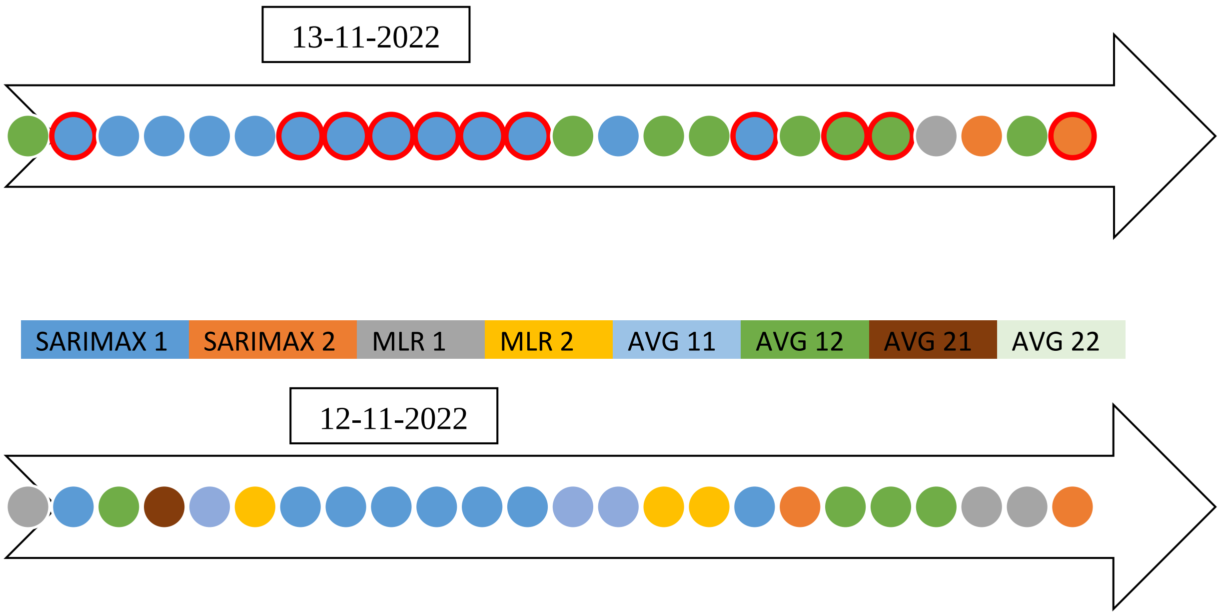

Figure 29. Diagram with example to explain D-1 Recommender Method (13 Nov. 2022)

In Figure 29 there are circles representing each hour of the day. The bottom arrow represents the day before the one that is being forecasted now (Nov. 13th) and the top arrow represents the day that is being predicted. Each circle has the color corresponding to the model that would have been the best option each of the hours (the legend in between the two arrows indicates which model that is). The hours where the best model would have been correctly chosen with the D-1 Method are the ones where equivalent circles are the same color. They have been circled in red above.

This method can be used with all eight methods or with as many or as little as it is considered necessary. It is possible that in this case there will be better results when only the AVG models are being compared. The Table 25 bellow has the results of both methods implemented for 5064 data points. It clearly shows a big difference in accuracy between the models. For the final comparison of all recommender models, we will only follow up on the D-1 AVG Method.

	D-1 ALL Method	*D-1 AVG Method*
RMSE Total	*16,689*	*15,872*
RMSE November	*18,886*	*18,390*
RMSE December	*19,493*	*18,112*
RMSE January	*18,871*	*17,582*
RMSE February	*13,367*	*12,632*
RMSE March	*19,944*	*19,010*
RMSE April	*16,180*	*16,532*
RMSE May	*12,231*	*12,035*

Table 25. Comparison of RMSE Values for D-1 Recommender Methods

If we are working with only two models that do not work well at the same time – in general, they are accurate in different circumstances – they could complement each other if the model is chosen correctly. One such pair in the case of our models consists of SARIMAX 1 and MLR 1.

Why only two methods? Because we will not be predicting directly which is a better model, but a ratio. In other words, RMSE values of MLR 1 will be divided by each corresponding RMSE value of SARIMAX 1. If the resulting value is higher than 1, the RMSE MLR 1 value is higher and therefore SARIMAX 1 is the better model. The opposite happens if the ratio is lower than 1. With this information we will attempt three new methods:

- Multi Linear Regression Model whose dependent variable is the ratio.
- SARIMAX Model whose dependent variable is the ratio.
- Logistic Regression Model that predicts the probability that SARIMAX is the better model (this last method is not as related to the other two, but Logistic Regression models work best with two clear categories so it should be compared with other models that choose between two options).

Method Name	***RMSE Value***
RMSE SARIMAX Method	*57,650*
RMSE MLR Method	*36,919*
LR Method	*37,233*

Table 26. Comparison of RMSE Values for RMSE Focused Methods (Dec. 2022 - Jan. 2023)

It is not surprising that Table 26 shows such high values for the RMSE of all the methods: neither MLR 1 nor SARIMAX 1 where great methods in the first place and any kind of mistake when choosing one of them results in a very high error. Overall, these methods do not work in any way that makes it necessary to compare them to any other method and they will be dropped now.

The next method we will try is a Decision Tree. A decision tree is a classification method based around the concept of nodes and branches. Based on information from a series of independent variables, every time the model gets to a node, it chooses the appropriate branch. The goal is to get to an end node that represents a specific subset or subgroup of the data (Song, Yan-Yan, and Ying Lu, 2015).

To reduce the possible residuals, the decision tree will only have four possible results (AVG 11, AVG 12, AVG 21 and AVG 22). The independent variables we will introduce are the same ones from MLR 1 and the predictions of the four models. We will feed the algorithm information about which was the best model out of the four in the past with a new variable 'Best'.

In Table 27 we can see how accurate the prediction (the percentage of the time that it correctly chose the best possible model out of the four options). However, this information is not the measure of accuracy we are interested in. The second column is much more relevant in assessing the overall effectiveness of the method. Both columns look very promising: the percentage of best model is around the same percentage of the SARIMAX models, knowing that in this case we do not have the issue with the worst model as none of the AVG are a great risk. The RMSE values are lower than the ones obtained in the best model during some months, but further analysis is necessary to reach any definite conclusions. We will analyze the model comparing it to the other recommenders later since they have not been tested for the same periods of time and they must be compared in equal circumstances.

	RMSE	*% Best Model*
Total	***15,931***	***28,32***
December	*17,870*	*26,75*
January	*17,901*	*27,69*
February	*12,614*	*29,91*
March	*18,825*	*26,61*
April	*16,045*	*27,78*
May	*12,011*	*31,32*

Table 27. RMSE Values and Performance Accuracy for Decision Tree Recommender Method

In Figure 30 we can see a visual example of data from a specific date (02/05/2023), the order of the AVG models (from best to worst) and the choices the Decision Tree made for each hour (circled in red). As previously clarified, an example is not enough to know how well the model works. Still, with the information from the previous table it is not far-fetched to derive that the model is successful in, at the very least, not choosing the worst available option.

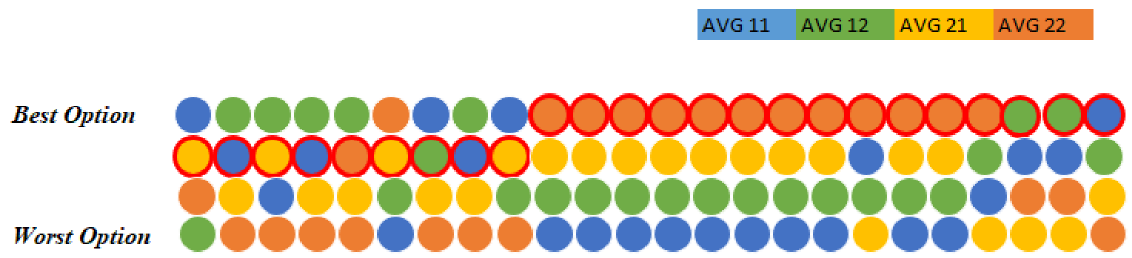

Figure 30. Diagram with example to Decision Tree Recommender Method (2 May. 2023)

The last method we will try for choosing the optimal model is called Random Forest. It is not a coincidence that the name of this method is so closely related to that of the Decision Tree. For starters, they are both machine learning models designed to make decisions. While a Decision Tree takes a single series of sequential decisions until it reaches its final node and therefore, its final decision, a Random Forest works slightly differently. Random Forest is still based on the concept of branches and nodes, but as its name hints at, it is formed by many trees. Simply put, a Random Forest algorithm is made of decision trees that are randomly generated. Together they generate a final output (Schonlau, Matthias, and Rosie Yuyan Zou, 2020). It is to be expected that Random Forest will outperform Decision Tree since the latter will always have a bigger overfitting problem. Decision Trees are easier to understand so they would be better in cases of processes that we wanted to understand step by step. In every other case and when the data pool is big enough to allow it, we should use Random Forest.

Table 28 below mirrors the one we just analyzed for Decision Tree. As expected, the results are better in this case in terms of performance in every month; Random Forest always manages to be right about the best model a higher percentage of the time than Decision Tree does.

	RMSE	***% Best Model***
Total	***15,878***	***32,19***
December	*18,313*	*26,21*
January	*17,230*	*32,66*
February	*12,350*	*34,52*
March	*19,114*	*30,11*
April	*16,067*	*31,11*
May	*11,888*	*38,71*

Table 28. RMSE Values and Performance Accuracy for Random Forest Recommender Method

Studying RMSE is a bit harder, so the information is visually shown in Figure 31. Even though Decision Tree performs better some months (namely December and March), the overall results are better with Random Forest (works best for more months and has a lower total RMSE value). For these reasons, it will be the model chosen between the two, which was the expected results as was explained above.

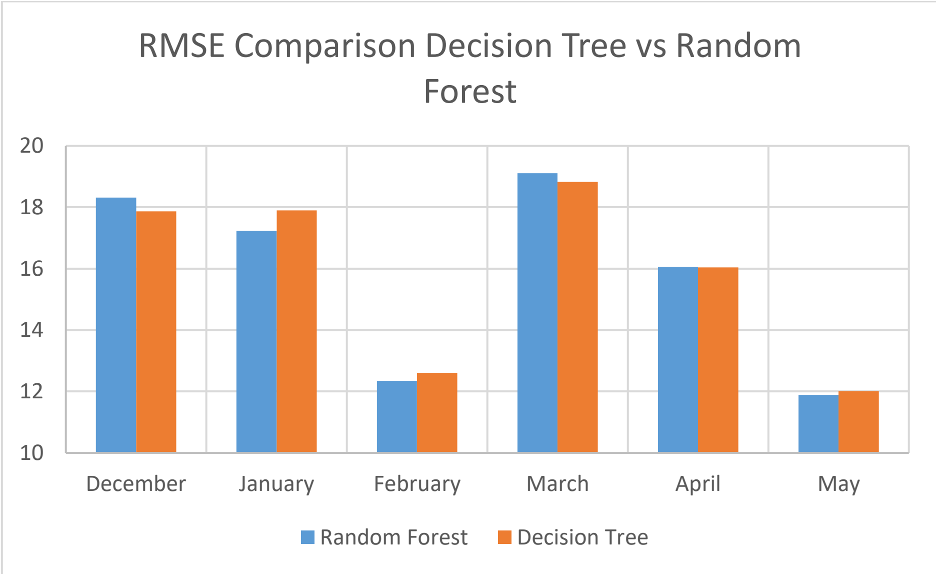

Figure 31. Comparison of RMSE Values per Month Decision Tree Method vs Random Forest Method

We have analyzed different recommender systems and the two that outperformed the others are 'D-1 AVG' Method and 'Random Forest' Method. We will compare them to the best method from the following section to reach an overall conclusion on the best method.

4.1.2 – ENSEMBLE AVERAGING

Ensemble averaging allows us to create a new model as a combination of the other existing models. The AVG models are examples of Ensemble Averaging although as we will explain in this section, ensemble averaging should be created with informed decisions. That is to say, explicitly choosing the coefficients that go with each of the models.

Any prediction method comes with the issue of finding a balance between minimizing the bias implicit to a model and reducing the variance of the model. Complex models have low bias and high variance, but simpler models have the opposite effect. Ensemble averaging helps reduce the variance and bias implicitly present in all models. It allows the model to sufficiently represent the true relationship between inputs and outputs but does not fall into overfitting. It has been empirically proven that "the performance of an ensemble is better than individual models" (Lim, Yenwee, 2022).

There are various methods for ensemble averaging, some extremely complex, some that use the previously mentioned Random Forest and Decision Tree methods and some that take very little time to compute. We will only look into the latter, at a direct combination of the four standard models (it would make no sense to include the AVG models since they are combinations of the SARIMAX and MLR models). It could be interesting to investigate different ensemble averaging methods in the future as they could help improve the performance of the models even more.

This process will be approached as a simple optimization process where we will try to minimize the RMSE value for the data Nov. 2022 – May 2023. The nature of this decision – choosing what exactly is the target that wants to be minimized – is not trivial. One could choose to minimize the RMSE of the latest month (not May but make the model one that recalculates itself with data from the latest month) or remove months from 2022 since they might not be relevant anymore. In this case we have chosen to make one decision and not a 'moving average'-type model primarily for two reasons:

1. To be able to understand how well that model would work would require a lot of time and computational capacity, making all predictions for each period of the day as well as optimizing the best coefficients each time.
2. The model could easily overfit and make bigger mistakes when it did make them, because it would be working with less data.

Therefore, only one model will be created, using information from months where predictions were good and had lower RMSE values (like April and May) and months that had higher RMSE values (like March and January). That way we can ensure that it is not biased either way and does not work extremely well in certain circumstances and not so great in other circumstances.

Since this is an elaborate weighted average, but an average all the same, the conditions of the optimization were set so that the coefficients all added up to a total of 1 with a step of 0,005. The weight each of the models has on the optimized product is represented in Figure 32. The SARIMAX models represent over half of the model. This is not unexpected because they had better results than the MLR models overall. SARIMAX 2 has a higher coefficient than SARIMAX 1 and MLR 2 has a higher coefficient than MLR 1, which also fits with their previously studied performance.

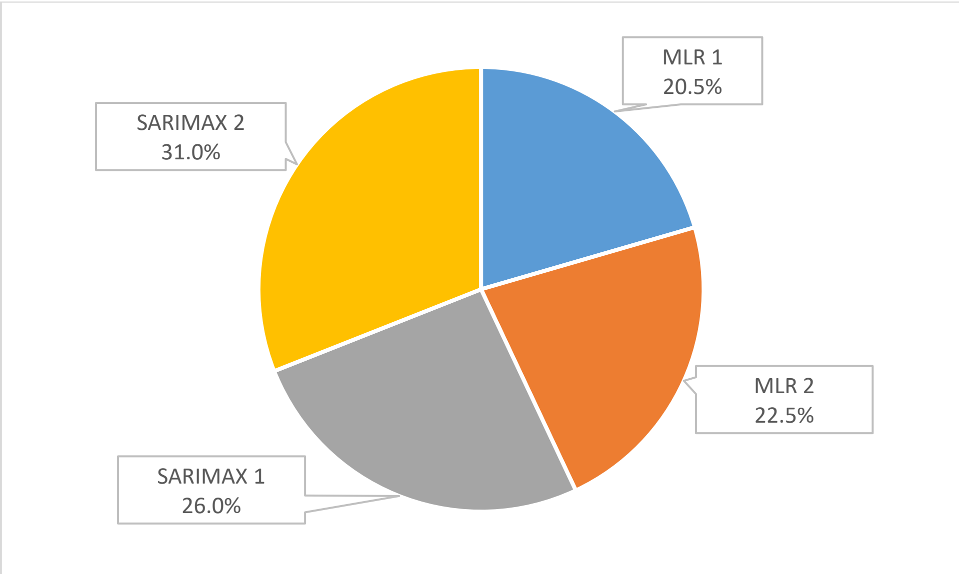

Figure 32. Optimized Coefficients of Weighted Average of all Models

4.2. – SELECTOR COMPARISON AND ANALYSIS OF THE RESULTS

The final methods that have potential as the best method to make a final decision are 'D-1 AVG', 'Random Forest' and 'Ensemble Averaging'. There is no way to predict Random Forest accurately for dates prior to December 2022 (it is necessary to have a month of information, or the results could be easily overfitted) so the comparison of models will be done in the period December 2022 – May 2023. They will be compared with all AVG models because the purpose of this section was to investigate whether there was a model that outperformed them.

RMSE/Model	***D -1 AVG***	***RF***	***EA***	***AVG 11***	***AVG 12***	***AVG 21***	***AVG22***
December	*18,112*	*18,313*	*17,637*	*18,062*	*18,143*	*17,880*	*17,985*
January	*17,582*	*17,230*	*17,570*	*17,914*	*17,947*	*17,155*	*17,429*
February	*12,632*	*12,350*	*12,448*	*12,626*	*12,405*	*13,042*	*12,979*
March	*19,010*	*19,114*	*18,532*	*18,919*	*18,785*	*18,857*	*19,071*
April	*16,180*	*16,067*	*16,077*	*16,692*	*16,251*	*15,969*	*15,739*
May	*12,035*	*11,888*	*12,233*	*12,177*	*12,209*	*12,057*	*12,145*
Total RMSE	***15,976***	***15,931***	***15,802***	***16,118***	***16,013***	***15,868***	***15,938***

Table 29. Final Comparison of RMSE Values of the Best Models

As discussed previously, the information we will give more importance to are total RMSE values to make sure that the model will stand the test of time. In Table 29 we are presented with the final comparison of RMSE values in this project. There is no clear best model if one looks at the performance per month, but one model stands out as having the best results overall: Ensemble Averaging. It is the only one of the new models that has a better performance than all AVG models (Random Forest is better than two of them and D-1 AVG only surpasses AVG 11).

We can conclude that the creation of all four models was not wasteful since all of them are present in the final chosen model: Ensemble Averaging Model. However, in the following pages we will elaborate on the possibility of improving the chosen model in future projects given the potential that these models offer as can be seen in the table below. In the table we can see the RMSE values that would result from choosing the best model every single time.

Month	***RMSE Value***
December	*10,065*
January	*9,746*
February	*6,618*
March	*11,458*
April	*9,140*
May	*7,674*

Table 30. Best possible RMSE Values with the Models per Month

CHAPTER 5. – CONCLUSIONS

5.1. – CONCLUSIONS ABOUT THE METHODOLOGY

The approach to the project has been highly effective. The work has been planned around a spiral method. A spiral method allows for a near infinite potential in terms of improvements while making sure that at any point the result is one that can be studied to reach conclusions. In this case, every time a model is created, all the information from that model is saved and compared with previously obtained information before choosing the next course of action. Instead of planning at the beginning of the project all the methods and models that were going to be analyzed, each step of the way one would question what the next logical proceeding would be. When the time came to finish the project, even if not everything that could have possibly been done was, it still allowed us to reach a lot of valuable conclusions.

The methodology also was not conventional when it came to comparing models. Not all models were contrasted at the end and that was purposeful. Instead, methods were eliminated earlier in the process, when they were shown to be worse than models that had similar principles in the way they worked. Rather than waste time and resources in models that were never going to be better than their counterparts, that time and those resources were used to create more models and methods that brought something new to the table.

5.2. – CONCLUSIONS ABOUT THE RESULTS

The results of the different models and methods have been extensively discussed and shown throughout these pages. Therefore, in this section we will draw attention to the most important conclusions reached from the results that perhaps have not been as emphasized as they should have.

The results from the comparison of all seasonal and autoregressive models are very conclusive: they show that the trend and seasonality as well as any other autoregressive term is not good enough to explain the behavior of the day-ahead electricity price. In the past, the prices behaved more consistently so past patterns were enough information to create a model that gave a good understanding of the present. However, that is no longer the case as could be seen when comparing the performance of SARIMAX with all other seasonal models that did not include exogenous variables.

The difference in performance between SARIMAX models and Multi Linear Regression models demonstrate that the seasonal and other autoregressive components are in fact present and that their addition improves the quality of the model overall (as SARIMAX models outperformed Multi Linear Regression models).

The results from the AVG models, in the same way as the Ensemble Averaging model, confirm that there is much value in creating different models that can be used to create a model with less overfitting and lower variance values. The best models on the whole were the ones created this way.

Unsurprisingly the outcomes that resulted from the use of a simple Decision Tree could be improved with Random Forest, although the difference was not strikingly big. Random Forest primarily helps with overfitting issues, but the data with which the Decision Tree was trained was enough that that was already not the biggest of issues.

The project can be closed successfully: it has shed light on many methods and models that can be used to forecast electricity prices and their true efficiency and accuracy, finding a resulting method that is able to follow the trend of the actual prices with great closeness.

5.3. – RECOMMENDATIONS FOR FUTURE STUDIES

The models created for this project (SARIMAX and Multi Linear Regression models with and without the variables generated via Logistic Regression) have a potential that was not fully explored due to time constraints. As previously shown in Table 30, the best possible RMSE values that are possible just by improving the recommender are close to half the ones obtained with the best model.

There is a very real possibility of improving these results. Directly related to the work done in the project there are two advisable paths that could be explored in the future:

- Ensemble Averaging: making the Ensemble Averaging Model train every day with data from the previous month has the potential of working better than using one model for all forecasts. On the other hand, there exist more complex Ensemble Averaging methods that are relatively easy to implement on Python that could be viable as improvements on the model.
- Recommender System: the methods to create such a system tested in the project did not produce the best results. However, they are not the only methods that exist that help with classification. One of the methods that was considered for the project and

that there was no time for is called Preference Ranking Organization METHod for Enrichment of Evaluations and its descriptive complement Geometrical Analysis for Interactive aid, commonly called Promethee and Gaia. It would be interesting to study its efficiency with these models.

If one had even more time on their hands, the original models could also be modified and possibly improved. One way that could be done is with the addition of new variables. For example, variables related to the temperature in key places of the peninsula or holiday effects. New models could be created that included the new variables. They could be compared to the old models and used to improve the Ensemble Averaging.

CHAPTER 6. – ANNEXES

ANNEX I: ALIGNMENT OF THE PROJECT WITH THE SUSTAINABLE DEVELOPMENT GOALS (SDG) OF THE UNITED NATIONS

The main purpose of this project is to investigate methods to forecast day-ahead electricity prices. It has been previously mentioned that this can "facilitate the integration of renewable energy sources by accounting for their volatility and variability". This directly relates to the UN's Sustainable Development Goals. One of the goals under goal 13 (Climate Action) is to achieve net-zero emissions by 2050. This goal aims to take urgent action to combat climate change and its impacts by strengthening and more relevant to our case, implementing technologies that support low-carbon development. This process requires a transition to renewable and clean energy sources.

Forecasting day-ahead electricity prices has the potential of helping with this transition: it will allow market agents to optimize their use of renewable energy sources and help reduce their dependence on fossil fuels. This means that by predicting how the electricity prices will behave depending on variable such as weather conditions and the electricity demand, these agents can better plan their production and consumption strategies and optimize their use of renewable energy sources. This can also reduce the need for conventional power plants and increase the share of clean energy in the electricity mix.

This project's purpose once put into use after the investigation of the models is over and a conclusion has been reached is precisely to help one such market agent with a hydraulic power plant, directly aligning it with the 13th Sustainable Development Goal of the United Nations.

ANNEX II: BIBLIOGRAPHY

Anderson, Daivid, et al. "Numerical Measures." *Encyclopædia Britannica*, 8 May 2023, www.britannica.com/science/statistics/Numerical-measures.

Andrés Palacios, Marcos. "TFM. Estudio de Los Factores Que Afectan al Precio de La Electricidad En El Mercado Eléctrico Español y Análisis de Posibles Mejoras." *Escuela Técnica Superior de Ingeniería Industrial de Barcelona*, Jan. 2019, https://upcommons.upc.edu/bitstream/handle/2117/128950/marcos-andres-palacios-tfm.pdf?sequence=1&isAllowed=y .

Artley, Brendan. "Time Series Forecasting with Arima , Sarima and SARIMAX." *Medium*, 26 Apr. 2022, towardsdatascience.com/time-series-forecasting-with-arima-sarima-and-sarimax-ee61099e78f6.

Azaizah, Nadeem. "Time Series Forecasting Using TBATS Model." *Medium*, 21 Nov. 2021, medium.com/analytics-vidhya/time-series-forecasting-using-tbats-model-ce8c429442a9.

Bhandari, Aniruddha. "Multicollinearity: Causes, Effects and Detection Using VIF (Updated 2023)." *Analytics Vidhya*, 26 Apr. 2023, www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/#What_Is_Multicollinearity?

Bobbitt, Zach. "How to Test for Multicollinearity in Python." *Statology*, 12 Oct. 2022, www.statology.org/multicollinearity-in-python/.

Butlion, Justin. "Data Audit Guide: How To Run A Data Audit at Your Startup." *RSS*, 29 May 2021, www.projectbi.net/blog/data-audit-guide-startup.

Caro, Eduardo, and Jesús Juan. "Short-Term Load Forecasting for Spanish Insular Electric Systems." *MDPI*, 15 July 2020, www.mdpi.com/1996-1073/13/14/3645.

Cruz García, Alberto Miguel. "Contribuciones a La Predicción a Corto Plazo Del Precio de La Electricidad Mediante Modelos de Series Temporales." *Repositorio Comillas*, 2013, repositorio.comillas.edu/xmlui/handle/11531/51505.

Duca, Angelica Lo. "Understanding the Seasonal Order of the Sarima Model." *Medium*, 13 July 2021, towardsdatascience.com/understanding-the-seasonal-order-of-the-sarima-model-ebef613e40fa.

Electricity in Spain. "Electricity Prices in Spain." *Electricity In Spain*, 6 May 2023, electricityinspain.com/electricity-prices-in-spain/.

Elseidi, Mohammed. "Forecasting Temperature Data with Complex Seasonality Using Time Series Methods - Modeling Earth Systems and Environment." *Springer Link*, 27 Dec. 2022, link.springer.com/article/10.1007/s40808-022-01632-y.

Endesa. “¿Por Qué El Precio Del Co2 Influye En La Factura de La Luz?” *Endesa*, 19 Nov. 2021, www.endesa.com/es/la-cara-e/sector-energetico/efectos-precio-carbono-factura-luz#:~:text=El%20precio%20lo%20fija%20el,para%20alcanzar%20el%20objetivo%20propuesto.

Facebook Open Source. “Package ‘Prophet.’” *CRAN*, 14 Oct. 2022, cran.r-project.org/web/packages/prophet/prophet.pdf.

Frost, Jim. “How to Interpret P-Values and Coefficients in Regression Analysis.” *Statistics By Jim*, 21 Mar. 2023, statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/.

Graduate Programs Staff. “What Is Statistical Modeling for Data Analysis?” *Graduate Blog*, 26 Apr. 2023, www.northeastern.edu/graduate/blog/statistical-modeling-for-data-analysis/#:~:text=Statistical%20modeling%20is%20the%20process,mathematical%20model)%20of%20observed%20data.

Hayes, Adam. “Autoregressive Integrated Moving Average (ARIMA) Prediction Model.” *Investopedia*, 22 May 2023, www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp.

Hayes, Spencer. “Finding Seasonal Trends in Time-Series Data with Python.” *Medium*, 31 Aug. 2022, towardsdatascience.com/finding-seasonal-trends-in-time-series-data-with-python-ce10c37aa861.

Jafari, Mohieddin, and Naser Ansari-Pour. “Why, When and How to Adjust Your P Values?” *Cell Journal*, 7 Aug. 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC6099145/.

Jose, Victor Richmond R., and Robert L. Winkler. “Simple Robust Averages of Forecasts: Some Empirical Results.” *International Journal of Forecasting*, 30 Aug. 2007, www.sciencedirect.com/science/article/pii/S0169207007000878.

Lim, Yenwee. “Ensemble Averaging-Improve Machine Learning Performance by Voting.” *Medium*, 24 Mar. 2022, towardsdatascience.com/ensemble-averaging-improve-machine-learning-performance-by-voting-246106c753ee#:~:text=Ensemble%20averaging%20is%20the%20technique,opinion%20towards%20the%20final%20prediction.

Manani, Kishan. “Multi-Seasonal Time Series Decomposition Using MSTL in Python.” *Medium*, 28 Apr. 2022, towardsdatascience.com/multi-seasonal-time-series-decomposition-using-mstl-in-python-136630e67530.

Marcos Peirotén, Rodrigo Alejandro. “Short-Term Forecasting of Electricity Prices: A Hybrid Methodology Based on Fundamental and Statistical Analysis.” *Repositorio Comillas*, Feb. 2020, repositorio.comillas.edu/xmlui/handle/11531/53177.

Moody, James. “What Does RMSE Really Mean?” *Medium*, 5 Sept. 2019, towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e.

OMIE. “Tecnologías Que Marcan Precio Marginal Del Mercado Diario.” *OMIE*, 2023, www.omie.es/es/market-results/monthly/daily-market/marginal-price-technologies?scope=monthly&year=2023&month=1&country=1.

Pacce, Matías José, et al. *Repositorio Institucional - Bde*, 10 Aug. 2021, repositorio.bde.es/handle/123456789/17351.

Pannell, Reagan. “The Box-Cox Transformation: What It Is and How to Use It.” *LeanScape*, 4 Oct. 2022, leanscape.io/the-box-cox-transformation-what-it-is-and-how-to-use-it/.

Prabhakaran, Selva. “Augmented Dickey-Fuller (ADF) Test - Must Read Guide - Ml+.” *Machine Learning Plus*, 4 Apr. 2022, www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/.

Product Team. “Forecasting Time Series with Multiple Seasonal.” *Algoritma Data Science School*, 9 Sept. 2021, algoritmaonline.com/multiple-seasonal/.

Querol Herrá, Javier. “Desarrollo de Un Modelo de Predicción Del Precio de La Energía Eléctrica Para El Mercado a Plazo Mediante Redes Neuronales.” *Universitat Politécnica de Valencia*, 2019, riunet.upv.es/bitstream/handle/10251/121891/Querol%20Herr%C3%A1%20-%20Desarrollo%20de%20un%20modelo%20de%20predicci%C3%B3n%20del%20precio%20de%20la%20energ%C3%ADa%20el%C3%A9ctrica%20para%20el%20mercado%20a%20plazo%20mediante%20redes%20neuronales.pdf?sequence=3.

Saltz, Jeff. “What Is a Data Science Workflow?” *Data Science Process Alliance*, 5 Sept. 2022, www.datascience-pm.com/data-science-workflow/.

Schonlau, Matthias, and Rosie Yuyan Zou. “The Random Forest Algorithm for Statistical Learning.” *The Stata Journal*, 2020, journals.sagepub.com/doi/pdf/10.1177/1536867X20909688.

Song, Yan-Yan, and Ying Lu. “Decision Tree Methods: Applications for Classification and Prediction.” *Shanghai Archives of Psychiatry*, 25 Apr. 2015, www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/.

Verma, Yugesh. “Quick Way to Find P, D and Q Values for Arima.” *Analytics India Magazine*, 24 May 2022, analyticsindiamag.com/quick-way-to-find-p-d-and-q-values-for-arima/.

Xing, Chen. “Why TBATS?” *RPubs*, 7 May 2022, rpubs.com/chenx/tbats_notes.

Zajic, Alexandre. “What Is Akaike Information Criterion (AIC)?” *Built In*, 29 Nov. 2022, builtin.com/data-science/what-is-aic.