# IE 492-Electricity Market Clearing Price Forecasting

ADVISOR

Mustafa Gokce Baydogan, Bogazici University, Turkey

GROUP MEMBERS

Suheyla Seker, Alican Yilmaz, Fatma Nur Dumlupinar

**Abstract**

Electricity markets garner attention with their unique property of balancing requirements in both short and long term. In those competitive markets, market clearing price (MCP) forecasting is becoming quite important for the market participants as it can help them to adjust their bidding strategies and their own consumption and production schedule, which in turn increases their profits. A good price forecasting model requires a good understanding of the dynamics of the market of interest. This paper first analyzes the dynamics of the Turkish Electricity Market and then comes up with predictive approaches to forecast next-day MCP. Several statistical and machine learning models are used and compared in terms of their performance with performance measures across multiple time zones. (e.g., hour, season, weekday) A Bayesian Network model is also implemented in the Turkish Electricity Market and satisfactory results are obtained.

Index Terms: Electricity markets, forecasting, market clearing price, Bayesian network, SARIMA, random forest, decision tree, stochastic gradient boosting, linear regression.

Elektrik piyasaları hem kısa hem de uzun vadede dengeleme gerektirmesi özelliği ile dikkat çekmektedir. Bu rekabetçi piyasalarda piyasa takas fiyatı (MCP) tahmini, piyasa katılımcıları için teklif stratejilerini ve kendi tüketim/üretim programlarını ayarlamalarına, dolayısıyla karlarını artırmalarına

Date submitted: 2021-06-30

yardımcı olacağı için oldukça önemli hale gelmektedir. İyi bir fiyat tahmin modeli marketin dinamiklerinin iyi anlaşılmasını gerektirmektedir. Bu çalışma, ilk olarak Türkiye Elektrik Piyasasının dinamiklerini analiz etmekte, sonrasında ise bir sonraki gün MCP'yi tahmin etmek için tahmine dayalı yaklaşımlar öne sürmektedir. Bu çalışmada çeşitli istatistiksel ve makine öğrenmesi modelleri kullanılıp modellerin performansları birden çok zaman kırınımlarında(örn. saat, mevsim, haftanın günleri) performans ölçüm metodları ile karşılaştırılmıştır. Bayes ağı modeli (Bayesian network model) de Türkiye Elektrik Piyasasında uygulanmış ve tatmin edici sonuçlar elde edilmiştir.

Anahtar Kelimeler: Elektrik piyasaları, tahminleme, piyasa takas fiyatı, Bayes ağı, SARIMA, rassal orman, karar ağacı, stokastik gradyan artırma, lineer regresyon.

# Contents

## Introduction

### *Summary of the IE problem handled*

Turkish electricity market is divided into financial and physical markets and the physical part of the market is divided into three categories which are bilateral agreements, spot markets and real time markets. EXIST manages and regulates the day-ahead market which operates under the spot market, and prices are determined based on supply and demand. Day-ahead market regulator EXIST operates very actively since electricity is characteristically non-storable and must be delivered to consumers constantly.

With the new electricity market law put into force in 2001, the Turkish electricity market liberalized, and the market began to grow rapidly. Along with liberalization and rapid growth, the number of agents in the market began to increase rapidly, and there was also an increase in offers for the day-ahead market. After the price and quantity bids received from all market members, the market clearing price is determined by EXIST, and agents start transactions based on this price. If agents in the market could predict the market clearing price correctly, they could create their bids accordingly and maximize their profits.

The inability of market agents to accurately predict the market clearance price with high volatility and seasonality leads to a loss of profit. In this project, forecasting models were developed to estimate the market clearing price determined by EXIST, on an hourly basis by using general data, so that market agents can determine their bidding strategies. Models created to enable market agents to determine their bidding strategies allow individuals in the market to make more profit.

### *A short overview of identification, analysis, and solution methodologies*

The approach in the project is to estimate the market clearing price using data of the factors affecting supply and demand since these factors are known, while the individual data belonging to bids of market agents are unknown.

With this approach, first the correlation between market clearing price and factors which affect supply and demand of electricity such as fuel prices, special days and exchange rates was examined. The missing data was interpolated, and the market clearing price was converted into dollars to better examine the impact of exchange rates and inflation. Then, a regression analysis was performed to better examine the relationship between parameters and the market clearing price, and coefficients were examined. After the analysis, it was decided to examine the special days by dividing them into two groups, which increase and decrease the price, and to use rates such as the ratio of bilateral agreements to consumption and the ratio of the amount of electricity produced by renewable sources to total production in models.

Data of hourly electricity generation and consumption, market clearing price and bilateral agreements were downloaded from the Energy Exchange Istanbul (EXIST) Transparency Platform. In addition, data of the fuel price and exchange rates were downloaded from the data platform of the Central Bank of the Republic of Turkey.

Due to the seasonality of prices and frequent use in the literature, a SARIMA model was established. In addition to this model, two linear models which are Bayesian Network and linear regression were created in order to compare the performance of a model which mainly focuses on causal dependence between parameters. Besides, machine learning models such as decision tree, stochastic gradient boosting, and random forest were established to compare the performances of the models in forecasting market clearing price by using the aggregate level data. RMSE and MAE metrics were used to compare the performances of the models.

### ***Improvements (realized or expected) due to your work***

There are studies which forecast the market clearing price for different electricity markets as it is found in the literature survey. Many of those studies mainly focus on ARIMA and neural network models, but only a few of them focus on Bayesian network models.

Unlike most models in the literature, our study focuses on causal relationships affecting the mechanism of electricity generation and consumption, rather than focusing on the time series property of data. In addition, in our model, we used aggregate level data, as we do not have individual bidding data for market agents.

### ***Brief summary of conclusions***

Since it would be wrong to conclude whether a model works better in general, model performances were compared on a seasonal, monthly, weekly, and hourly basis.
In weekly analysis, Bayesian network, random forest and gradient boosting perform well but their performances are not significantly different, and almost all models perform worse on forecasting Market Clearing Price (MCP) on Sundays. According to hourly analysis, all models deteriorate at lunch hours and perform better at working hours. In seasonal analysis, all models perform better during fall and summer.
Among linear models, Bayesian network performs better for Turkish Electricity Market. The error variance of the Bayesian network model is small in hourly analysis. Among machine learning models, random forest performs better for Turkish Electricity Market.

### ***Contents of the report***

The report begins with an explanation of the problem and continues with the purpose of the project. After the aim of the project is identified clearly, requirements and limitations of the problem are clarified. After the source of the data used in the project and the metrics used to compare the performances of the models are mentioned, the stage of creating the models is explained. Just before models are clarified, similar studies in the literature and the necessary IE tools which should be used are explained.
The following sections contain technical information about data analysis and models. The parameters used in the models and the reason why they are selected are mentioned. After that, the models used and the performance of these models in different conditions are compared and the models which perform better under different conditions are mentioned. Finally, the results obtained from the project and additional work that can be applied in future periods are mentioned.

## Problem Definition, Requirements and Limitations

### *What seems to be the problem?*

The main focus of the project is analyzing the dynamics of Turkish Electricity Market and building a forecasting model to predict day ahead MCP at hourly level. Electricity trade in Turkey is operated through two main market structures, namely physical markets and financial markets. Financial market consists of a derivative market which is operated by Borsa İstanbul. Due to unpredictability of the electricity prices, the total volume of the derivative market is not the same level as other commodity markets.Physical markets are operated through three different markets (namely Day-Ahead Market, Intraday Markets and Balancing Power Markets) that are realized in different time zones. Electricity is a non-storable commodity and the balance between its consumption and generation must be satisfied constantly to maintain the stability in the overall system. Because of these two features, volatile behavior is observed in electricity prices [8].

The transactions in the Day-Ahead Market and Intraday Market are operated by EXIST. In this project, the main focus is on the Day-Ahead market where the MCP is determined via an optimization algorithm developed by EXIST. The bids consisting of price and quantity for each hour for the next day are collected by EXIST and the resulting MCP is obtained with the objective of maximizing total social welfare, which is total consumer and producer surplus. However, since this problem needs to be solved fast for each hour, tabu search and genetic heuristics are developed, and are being used today by EXIST [2].
Although the individual bids by all the market participants cannot be known beforehand, the factors that might affect their decisions, and supply/demand dynamics of the market can be analyzed, thus, a reliable forecasting model can be developed based on these input factors and previous MCP information.

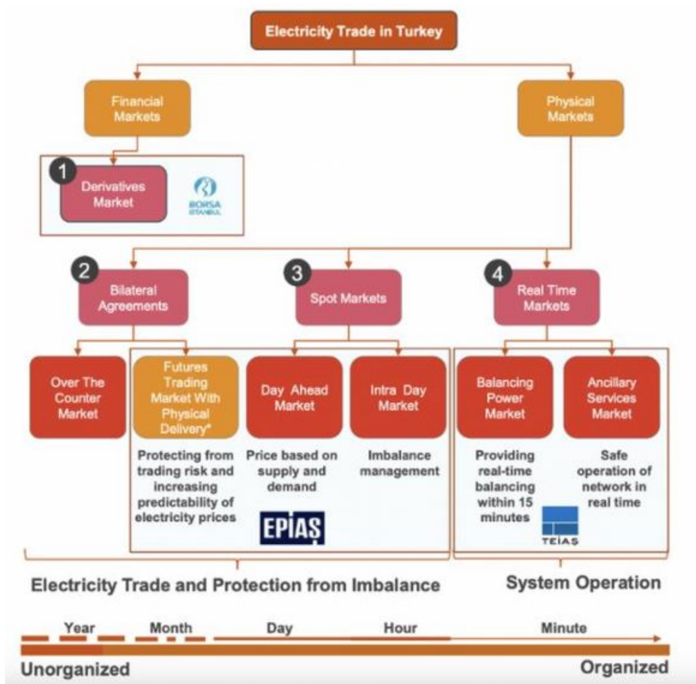



**Figure 1.** General Market Structure of the Turkish Electricity Market, Adapted from EXIST

***What is done to understand the causes of the problem?***

To understand the dynamics of the MCP, a detailed exploratory analysis is performed. The linear correlations between the variables are examined with correlation and regression analyses. However, additional visual and statistical analyses are needed to understand the non-linear behavior of the MCP. Causal relations between the variables are determined via statistical hypothesis tests (e.g., Granger Causality test). Even if the factors that have a potential impact on the MCP could be determined and predicted perfectly, the result might not turn out to be as expected, since the market participants might behave unrealistically. Similarly, a model which takes into account some factors, might fail in predicting specific(outlier) cases such as political turbulence times or major disruptions such as novel COVID-19.
The specific time intervals where MCP behaves unusual are analyzed separately to infer some conclusions about the influencing factors. After the analysis, it has been

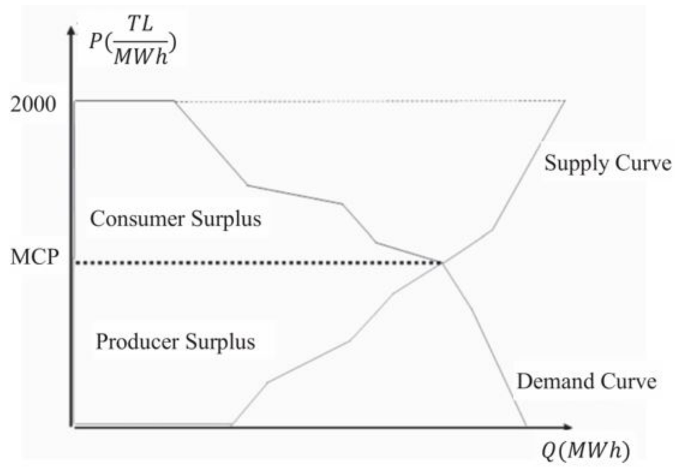



**Figure 2.** Consumer and Producer Surpluses from Hourly Bids [2, Figure 3]

observed that converting some inputs to ratios from their absolute values improves the model performance. (For instance, "proportion of Renewable Energy Generation to Total generation" is used as an input for the models instead of their absolute values.) The factors that are commonly used in the literature are analyzed and incorporated with our market specific domain knowledge to finally obtain input variables that will be used in the models. These factors are explained in section 4.

### ***Needs and requirements of the system/customer***

In Turkish electricity market, participants need to submit their bids until 12.30 for the following day. EXIST determines the prices after the bids are offered. Then, inconveniences in the bids are reported to participants, if any. Participants have a right to object to the results and finalized MCPs. Trade volumes are announced at 14.00 by EXIST. Due to the market structure, participants need to make their forecasting at least a day before to adjust their bidding strategies for the next day. Participants can resort to the intraday markets if they want to supply their needs. With the help of a balanced market one day ahead, suppliers can adjust their generation schedule and demand side can fine tune their consumption based on the price of the electricity. To develop a forecasting model at hourly level for the next day, related data of the variables that have a direct or indirect effect on the price needs to be collected and missing data should be imputed as correctly as possible.

***Limitations and constraints***

Obtaining, preprocessing and analyzing the data is a challenging task as some data might be missing and needs to be interpolated correctly, or some might require preprocessing (e.g. dollar/TL transformation, inflation adjustment). Also, many macroeconomic factors that might have an effect on the MCP have different time levels (i.e., year, month) which results in a complication for incorporating them into the models. The share of renewable energy sources (e.g., wind, solar, geothermal, hydropower) is in an uptrend, a factor that makes the accuracy of the forecasting model lower due to unpredictability.

***Data gathered and used in identification phase***

The main data source of the relevant data for our project is EXIST's own platform named "EXIST Transparency platform" where the data of the electricity and other relevant market information is accessible to all participants at the same time so that equality and fairness are maintained in the market. Adding to the transparency platform of EXIST, TCMB (Central Bank of the Republic of Turkey) and TPPD (Turkish Petroleum) sources are used to collect necessary data. Special days data (e.g., religious holidays, school holidays, new year eve, Ramadan etc.) is constructed by-hand for the train and test period.
Market specific hourly data is collected from EXIST Transparency Platform, namely total realized consumption (MWh), real-time generation (MWh), real-time renewable generation (MWh), Bilateral Contracts (MWh). However, after both data and market analysis, it has been concluded that "proportion of renewable generation over total generation", and "proportion of Bilateral Contracts over Consumption" has a significant effect on the MCP. Thus, proportions are used instead of their absolute values. Fuel price data collected from TPPD contains missing values at daily level. To overcome that, linear interpolation is implemented as it fills the missing values

with as little deviation as possible. The MCP values and fuel prices are transformed into dollars to deal with the inflation effect on price.

### *A context diagram (a systematic view) of the handled design problem*

The context diagram of the Turkish electricity market can be seen in the figure given below. For market clearing price forecasting, the day ahead market of the electricity market is focused. The players of this market are generators, electricity retailers, traders, utility companies and large industrial consumers. They can make bilateral agreements between each other, or they can submit their bids for the upcoming day. Then, EXIST determines the market clearing price using its optimization model, but they are actually using a heuristic algorithm. Thus, EXIST actually finds a good market clearing price, but it is not the optimum one. Finalized market clearing prices are announced after objections of the market agents are evaluated.

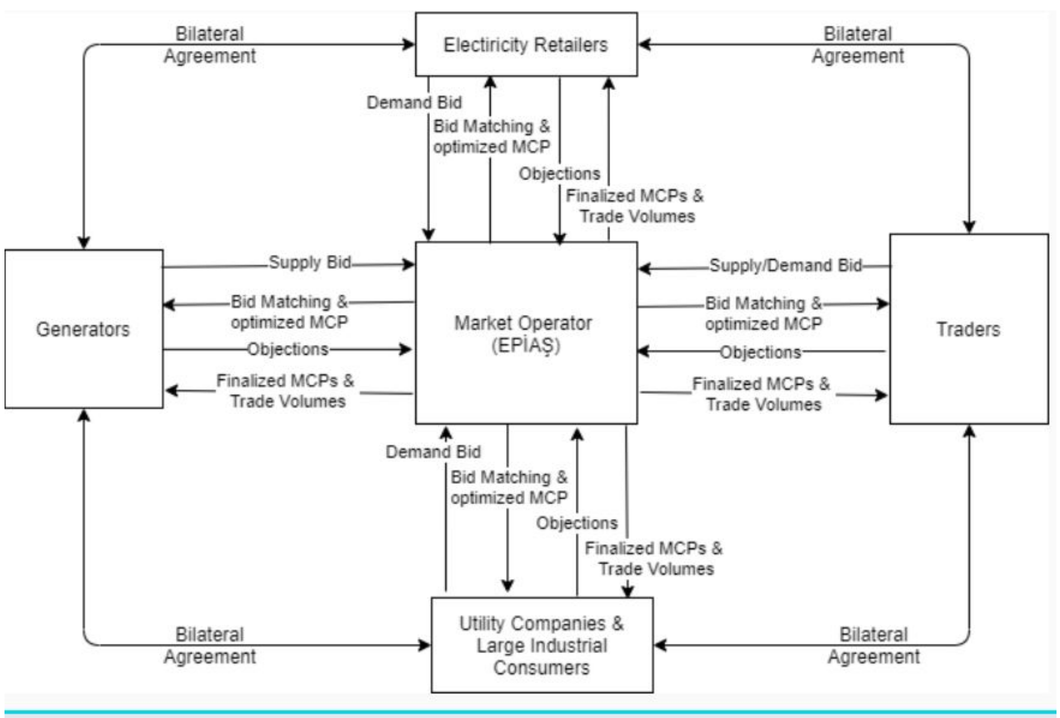



**Figure 3.** Context Diagram of the Turkish Electricity Market

### *Performance criteria and potential improvements*

Machine learning algorithms and other models used in the project allowed many performance metrics to be used but two metrics common in the models were selected for comparisons. RMSE and MAE metrics were used to compare model performances, as the project aimed to accurately estimate the market clearing price hourly. The MAPE metric used in some studies in the literature was not used in this study, because the market cleaning price in the Turkish electricity market is sometimes equal to zero and this prevented the use of this metric in comparisons. In order to make more detailed and accurate comparisons when comparing model performances, these two metrics were compared for specific conditions such as hourly, weekly, seasonal, and daily.

## Analysis for Solution/Design Methodology

### *Literature overview*

Bidding price is crucial to optimize the prosperity of both producers and consumers in the electricity market. According to Yan and Chowdhury [10], the most significant aspect in the customized electricity markets is to provide the needed electricity quantity at the right time, with the correct bidding price, and they also state that market clearing price (MCP) which comes up as a result of the market bidding price, is an estimate of the future price of electricity based on predictions of demand, temperature, sunlight, weather conditions etc.

Since many factors are effective in shaping the market clearing price (MCP), a single forecasting method will not be sufficient to estimate the price at the desired accuracy rate. Due to difficulty and complexity in estimation of MCP, many methods such as time series analysis, ARIMA and regression models, Bayesian and simulation techniques have been studied to most accurately predict MCP [3]. For these reasons, modeling consisting of combinations of several forecasting and statistical methods can also be used to predict MCP.

There are many factors that are effective in determining the market clearing price.

In the literature review for MCP forecasting, it can be seen that different input parameters are used in models created by different methods. Although some of these input parameters are used in common in different models, it has also been observed that some input parameters are market-specific parameters.
In [3], adaptively trained neural network model was created by using historical MCP, historical load, and forecasted load for data with price spikes for California Power Market for year 1999.
Kölmek and Navruz [4] built a model for forecasting day-ahead pricing of Turkish Electricity Market by using the weighted average value of temperatures in cities whose population is more than 3.000.00, estimated consumption for the day, the amount of bilateral contracts realized for the day and the past day-ahead price (SDAP) in their artificial neural networks (ANN) and ARIMA models.
In Yan and Chowdhury's article, PJM Interconnected Electric Market's mid-term market clearing price was modeled by using electricity hourly demand at hour t, electricity daily peak demand, electricity monthly average demand, daily price of natural gas, previous year's monthly average electricity MCP, month and hour of the day [11]. In study [7] New England Market was modeled by using input parameters predicted and historical market demand, historical MCP, power import to and export from other markets through inter-market contracts, fuel price, weather information and hydro energy prediction.
Lu et al. [5] used demand, reserve (supply-demand), supply-demand balance index and relative demand index to forecast market price spikes of for Queensland Electricity Market of Australian NEM.

### ***Alternative solution/design approaches***

Weron [9] created a comprehensive article summarizing the models used to forecast MCP and tried to classify electricity price models in the literature, summarized them as five main approaches: Multi-agent models, Fundamental methods, Reduced-form models, Statistical approaches, Computational intelligence techniques.
Yan and Chowdhury [10] used a hybrid LSSVM and ARMAX approach to forecast

MCP. Then next year, Yan and Chowdhury [11] tried to make MCP predictions by using hybrid support vector machine and auto-regressive moving average with external input, and and they noted that from the models they applied, the hybrid SVM and ARMAX prediction model outperformed the hybrid LSSVM and ARMAX model, and the models they applied alone.

In a study [2] that looks at the problem from a mathematical programming perspective, day-ahead combinatorial auction is modeled in which the total surplus of the market is maximized. Then authors approached the problem with heuristics algorithms which are Tabu Search and Genetic Algorithm and found that Genetic Algorithm performs better [2]. Georgilakis used adaptively rained neural networks (ANN) method to forecast MCP of California Power Market for year 1999 by using data with price spikes and without price spikes [3].

Kölmek and Navruz [4] created two different model for Turkish electricity Market by using Artificial Neural Networks and Time Series approaches and compared their performances.

Auto-regression, Iterative clustering algorithm and Gaussian kernel methods were used to forecast 1999's New England Market's MCP and got better results on average on-peak MCP prediction than average off-peak MCP prediction [7].

Lu et al. [5] benefited from Bayesian categorization approach for Queensland Electricity Market of Australian NEM market clearing price forecasting for 2003.

## *Assumptions*

Before creating forecasting models, some assumptions were made due to missing data and implementation difficulties. The first assumption used in the study was to use data from the previous hour for missing dollar exchange rate data because many of the hourly macroeconomic data were missing, and the most appropriate data that could be used was the data of the previous hour.

In addition, it was assumed that electricity producers produced electricity under all appropriate conditions in the perfect market assumption. This assumption was made because there was no official data, although it was estimated that there were producers who did not produce to protect their plants when market clearing prices were low.

### ***Brief overview of the selected approach(es)***

I. Machine Learning Approach

- Decision Tree
- Stochastic Gradient Boosting
- Random Forest

II. Statistical Approach

- Seasonal ARIMA
- Linear Regression
- Bayesian Network

### ***IE skills/tools/techniques/methods to be integrated to implement the proposed methodology***

In this project, the industrial engineering curriculum contributed greatly to understanding market dynamics and model development. At the stage of understanding the problem, being able to understand the optimization model that determines the market clearing price allowed us to have an insight about the problem at the initial stage. In this way, we have seen an application of the topics in the operations research courses.

The basic topics covered in the IE department have been widely used. Since the data used is time series, time series analysis and forecasting are made with the seasonal ARIMA model. In linear regression and Bayesian network applications, the topics in statistics and probability courses are used. In addition, the data mining course enables us to build models and interpret with the machine learning approach.

## Development of Alternative Solutions

### *Detailed technical description of the design process for generating meaningful alternatives*

### I.Feature Selection

While deciding on the features, the parameters previously used for electricity price estimation in the literature are examined at the beginning. Correlation and causality between all parameters are analyzed with various visualizations, correlation matrices and stepwise regression approaches. As a result of the analysis, it is determined which parameters are suitable for Turkish electricity market data and may be important in price estimation.

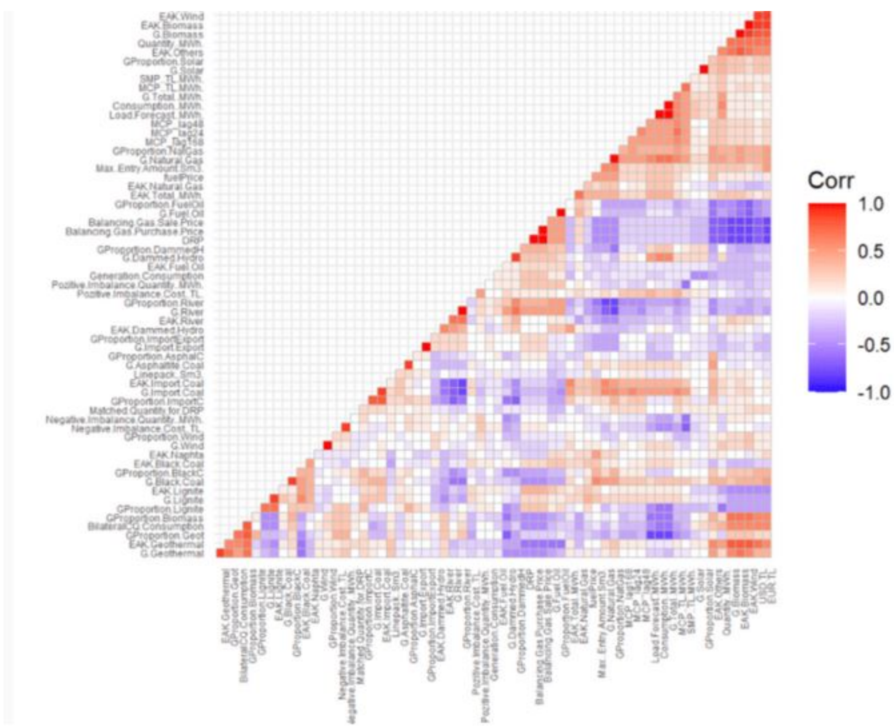



**Figure 4.** Correlation Matrix of the All Parameters

According to the correlation matrix above, MCP is highly correlated with its lagged values, consumption, and total generation amount.

- *Daily Mean Fuel Price:*

The scatter plot below shows the daily mean of MCP values. Green points means that the MCP value decreased more than twenty percent compared to previous day, and red points means that the MCP value increased more than twenty percent compared to previous day.Blue points say the change is less than twenty percent. Also, the line on the plot indicates daily mean of fuel price. Before 2019, the relation between two

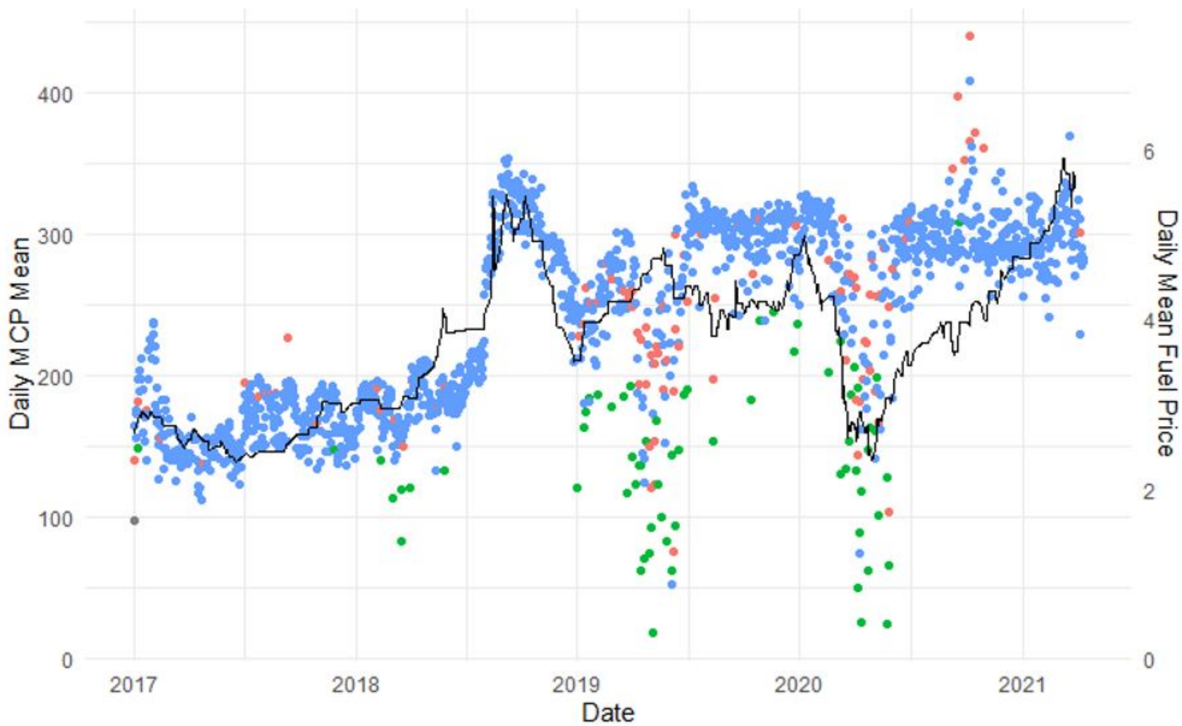



**Figure 5.** Analysis of Daily Mean of Fuel Price and Daily Mean of MCP

is clear, the patterns overlap. However, after that year, more explanation about the deviation on the periods when green and red points increase is needed.

- *Proportion of Daily Renewable Energy Generation:*

Since the amount of renewable energy generation can be easily affected by external factors such as weather conditions during the year, it also causes a change in the total amount of generation. These changes happen in spring seasons.

In the scatter plot below, the dots show the MCP values, and the black line shows the percentage of renewable energy in total energy generation. The coloring of the MCP values is the same as the previous plot. It can be said that high changes occur in MCP during periods when the share of renewable energy production exceeds fifty percent.

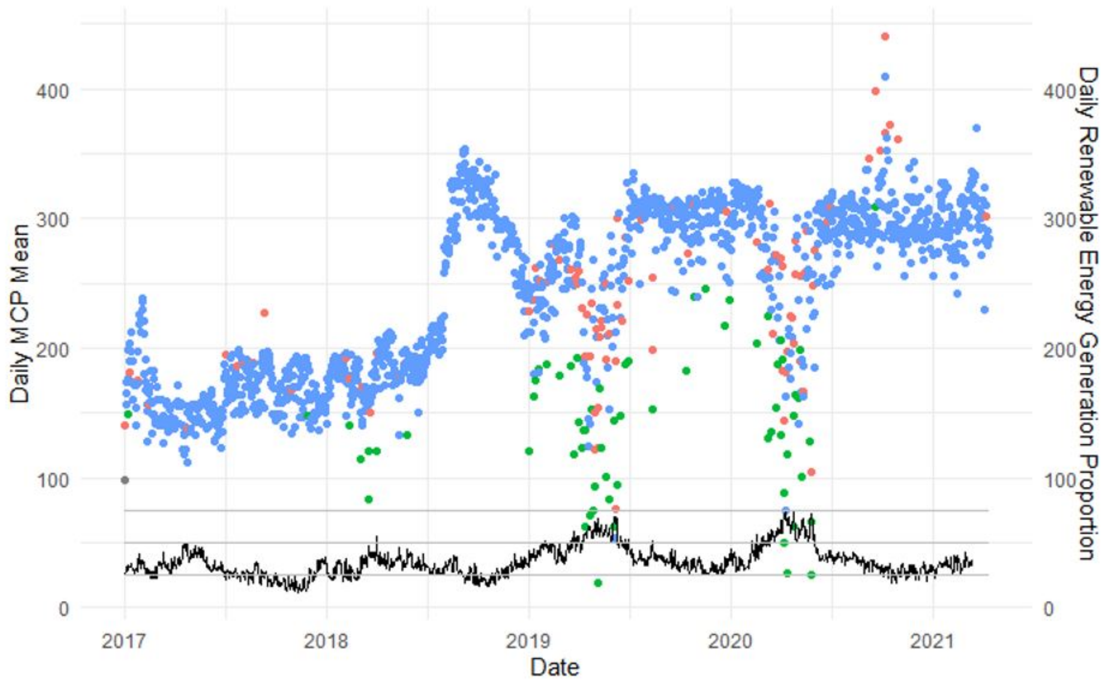



**Figure 6.** Analysis of Daily Generation Proportion of Renewable Energy and Daily Mean of MCP

Deviations that cannot be explained by the fuel price plot can be explained by the renewable energy effect. In 2019, an increase in renewable proportion comes into the play. The proportion goes beyond fifty percent in spring seasons, and we thought this causes MCP to decrease in spring seasons. Red points are also seen during those periods probably because of fluctuations on varying renewable energy capacity.

- *The Ratio of Bilateral Agreement Quantity to Consumption:*

Since bilateral agreements made before the market opens affect the remaining un-

matched production and consumption quantities, it also indirectly affects the MCP resulting from the matches.
The same logic applies for this plot in terms of coloring. This time, the line indicates the ratio of bilateral agreement quantity to consumption. Like the renewable energy proportion, the increase in the ratio becomes effective in spring.

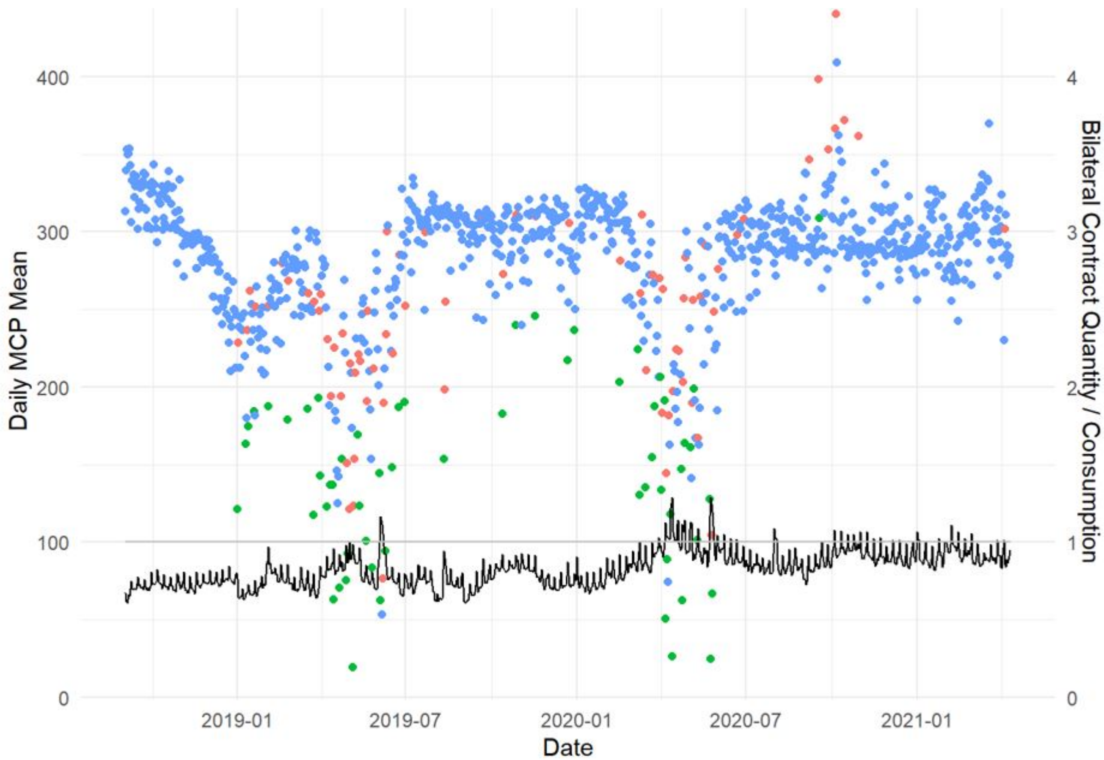



**Figure 7.** Analysis of Bilateral Contract Quantity/Consumption and Daily Mean of MCP

At the end, some features are determined to use in models. These are consumption, generation, the ratio of bilateral quantity to consumption, fuel price, proportion of renewable energy generation, MCP of the previous day, days of week, hour, and special days. Special days are divided into two categories. These are the special days that affect MCP positively and negatively.

## II.Train and Test Period

Two things are considered when choosing the time interval of the data: pandemic and data availability. Since the fullness in the data before 2017 decreased, the part after this year is taken. In addition, pandemic data is not included as different dynamics developed that could affect the electricity market with the spread of the covid pandemic in the world. For this reason, the data after 2020 is not included in the test period. As a result, the data from 2017 to the end of 2018 is chosen as the train and 2019 as the test period.
Due to the high computational effort, models that make hourly price estimations are run every month, not every day. While training the models, new realized values are included in the model by extending the train period every month using the sliding window approach.

## III.Modelling

After determining periods, the models that are frequently used in the literature for the electricity market price forecasting are evaluated. The models which are easy to apply and interpret like linear regression and seasonal ARIMA are also chosen. On the other hand, nonlinear models are needed because there are nonlinear relationships in the data.
Therefore, the models that predict the electricity price with statistical and machine learning approaches in the literature are examined. As a result, 5 models are identified to compare their prediction performances. Seasonal ARIMA, linear regression from statistical models, and decision tree, stochastic gradient boosting, random forest models from machine learning models are applied. In addition to these, due to the high causality between the parameters in the market structure, Bayesian network model may also be suitable for the problem although it is not used so much before for the electricity market. So, the Bayesian network is added to the statistical models to be established.

In the machine learning approach, parameters of the models are required to be tuned for more accurate results. Although cross validation method is commonly used for this purpose, time slice method is used since the data used is time series and the order of the time series should be saved when tuning parameters.

```
fitControl <- trainControl(method = "timeslice",
                           initialWindow=24*30*4, fixedWindow=TRUE,
                           horizon=24*30,
                           skip=24*30-1)
```

**Figure 8.** Parameters of Time Slice Method

## 1. Machine Learning Models

In this approach, nonlinear models are built since it can be appropriate for the data containing nonlinear relationships.
Parameters of the models are tuned to give the least RMSE value. Optimal values of the related parameters are given in the figures.

### Decision Tree

```
CART

17496 samples
   10 predictor

No pre-processing
Resampling: Rolling Forecasting Origin Resampling (720 held-out with a fixed window)
Summary of sample sizes: 2880, 2880, 2880, 2880, 2880, 2880, ...
Resampling results across tuning parameters:

  cp          RMSE      Rsquared   MAE
  0.04452381  7.566532  0.3285360  5.587201
  0.07233983  7.857856  0.2931450  5.799630
  0.34714775  8.923329  0.2768151  6.831317

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.04452381.
>
```

**Figure 9.** Best Tunes of the Decision Tree

Complexity parameter of the decision tree is tuned as 0.0445 for the smallest RMSE value which is 7.5665.
The MCP value of the 24 hours before, generation, consumption, the ratio of bilateral contracts to consumption and the proportion of renewable energy generation are the most important variables of this model.

```
> varImp(DecTree)
rpart variable importance

  only 20 most important variables shown (out of 37)

                             Overall
lag_24                       100.000
Generation                    84.923
Consumption                   84.849
BilateralContacts.Consumption 46.848
Hour4                          9.766
RenewableProportion            8.873
Hour12                         0.000
Hour6                          0.000
Hour7                          0.000
daySaturday                    0.000
Hour10                         0.000
dayThursday                    0.000
dayTuesday                     0.000
Hour20                         0.000
PositiveSpecialDays1           0.000
Hour18                         0.000
Hour3                          0.000
Hour15                         0.000
Hour21                         0.000
Hour11                         0.000
```

**Figure 10.** Importance of the Variables for Decision Tree

### Stochastic Gradient Boosting

Number of trees(n.trees) is tuned as 100, maximum nodes per tree (interaction.depth) as 3, shrinkage as 0.1 and the minimum number of observations allowed in the trees terminal nodes(n.minobsinnode) as 10 for the smallest RMSE value which is 6.4614 .

The MCP value of the 24 hours before, generation, consumption, the ratio of bilateral contracts to consumption, the proportion of renewable energy generation and fuel price are the most important variables of this model.

```
Stochastic Gradient Boosting

17496 samples
   10 predictor

No pre-processing
Resampling: Rolling Forecasting Origin Resampling (720 held-out with a fixed window)
Summary of sample sizes: 2880, 2880, 2880, 2880, 2880, 2880, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  RMSE      Rsquared   MAE
  1                   50      6.807849  0.4807816  4.972531
  1                  100      6.593183  0.5033370  4.774100
  1                  150      6.493676  0.5125786  4.695723
  2                   50      6.689359  0.5001594  4.849609
  2                  100      6.507077  0.5204319  4.704186
  2                  150      6.541276  0.5241714  4.743485
  3                   50      6.537775  0.5142953  4.724080
  3                  100      6.461413  0.5294012  4.688267
  3                  150      6.538704  0.5301520  4.739808

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning parameter 'n.minobsinnode' was held constant at a value of 10
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were n.trees = 100, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10.
```

**Figure 11.** Best Tunes of the Stochastic Gradient Boosting

### Random Forest

Number of predictors sampled for splitting at each node (mtry) is tuned as 5 for the smallest RMSE value which is 6.4336.

```
rf variable importance

  only 20 most important variables shown (out of 37)

                              Overall
lag_24                       100.0000
Generation                    56.6941
Consumption                   45.6211
BilateralContacts.Consumption 41.1543
FuelPrice                     21.5472
RenewableProportion           20.0916
Hour4                          4.0504
Hour5                          3.7985
dayMonday                      3.2619
Hour3                          3.0128
NegativeSpecialDays1           2.8810
daySunday                      2.7258
PositiveSpecialDays1           2.6055
Hour23                         1.7466
Hour6                          1.2897
Hour9                          1.0319
Hour2                          1.0030
Hour8                          0.9372
Hour14                         0.9190
Hour11                         0.7813
```

**Figure 14.** Importance of the Variables for Random Forest

```
> summary(gbmFit)
                                                     var     rel.inf
lag_24                                            lag_24 61.79225425
Consumption                                  Consumption 16.27214574
BilateralContacts.Consumption BilateralContacts.Consumption  8.52665618
Generation                                    Generation  5.00585503
RenewableProportion                  RenewableProportion  2.95201849
dayMonday                                      dayMonday  1.98040583
FuelPrice                                      FuelPrice  1.68488085
PositiveSpecialDays1                PositiveSpecialDays1  0.58849196
Hour23                                            Hour23  0.29395676
Hour14                                            Hour14  0.20196345
Hour4                                              Hour4  0.16966380
NegativeSpecialDays1                NegativeSpecialDays1  0.14233681
Hour8                                              Hour8  0.14230415
Hour9                                              Hour9  0.05693125
Hour1                                              Hour1  0.04420511
Hour11                                            Hour11  0.03678991
daySunday                                      daySunday  0.02763939
Hour6                                              Hour6  0.02326580
Hour5                                              Hour5  0.02139164
Hour3                                              Hour3  0.01845727
Hour22                                            Hour22  0.01838632
Hour2                                              Hour2  0.00000000
Hour7                                              Hour7  0.00000000
Hour10                                            Hour10  0.00000000
Hour12                                            Hour12  0.00000000
Hour13                                            Hour13  0.00000000
Hour15                                            Hour15  0.00000000
Hour16                                            Hour16  0.00000000
Hour17                                            Hour17  0.00000000
Hour18                                            Hour18  0.00000000
Hour19                                            Hour19  0.00000000
Hour20                                            Hour20  0.00000000
Hour21                                            Hour21  0.00000000
daySaturday                                  daySaturday  0.00000000
dayThursday                                  dayThursday  0.00000000
dayTuesday                                    dayTuesday  0.00000000
dayWednesday                                dayWednesday  0.00000000
```

**Figure 12.** Importance of the Variables for Stochastic Gradient Boosting

```
Random Forest

17496 samples
   10 predictor

No pre-processing
Resampling: Rolling Forecasting Origin Resampling (720 held-out with a fixed window)
Summary of sample sizes: 2880, 2880, 2880, 2880, 2880, 2880, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
   5    6.433675  0.5466319  4.724137
  15    6.663390  0.5148534  4.834293

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 5.
>
```

**Figure 13.** Best Tunes of Random Forest

The MCP value of the 24 hours before, generation, consumption, the ratio of bilateral contracts to consumption, fuel price and the proportion of renewable energy generation are the most important variables of this model. Also, the importance of fuel price and the proportion of renewable energy generation is higher than the ones in the decision tree and stochastic gradient boosting models.

## 2. Statistical Models

### Linear Regression

Linear regression is one of the models that is easy to apply and requires low computational effort. So, it is a good alternative to use as a benchmark model.
In the figure below, it is demonstrated that all the features used have high significance.
Also, the features can be analyzed in terms of whether they have a positive effect or not from the first column.

```
Call:
lm(formula = MCP ~ ., data = (BNDataset %>% filter(Date < date))[,
    -1])

Residuals:
    Min      1Q  Median      3Q     Max
-80.407  -3.247   0.219   3.587 228.204

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.108e+01  1.451e+00  28.310  < 2e-16 ***
Hour1                   -1.294e+00  3.388e-01  -3.820 0.000134 ***
Hour2                   -3.251e+00  3.430e-01  -9.479  < 2e-16 ***
Hour3                   -4.141e+00  3.466e-01 -11.949  < 2e-16 ***
Hour4                   -4.878e+00  3.489e-01 -13.978  < 2e-16 ***
Hour5                   -4.279e+00  3.486e-01 -12.278  < 2e-16 ***
Hour6                   -4.261e+00  3.495e-01 -12.189  < 2e-16 ***
Hour7                   -3.142e+00  3.449e-01  -9.110  < 2e-16 ***
Hour8                   -1.312e+00  3.397e-01  -3.861 0.000113 ***
Hour9                    3.334e-01  3.431e-01   0.972 0.331166
Hour10                   2.541e-01  3.454e-01   0.736 0.461858
Hour11                   6.945e-01  3.478e-01   1.997 0.045860 *
Hour12                  -1.497e+00  3.426e-01  -4.369 1.26e-05 ***
Hour13                  -7.704e-01  3.438e-01  -2.241 0.025032 *
Hour14                   1.048e-01  3.468e-01   0.302 0.762435
Hour15                  -6.326e-01  3.460e-01  -1.828 0.067496 .
Hour16                  -6.984e-01  3.465e-01  -2.016 0.043855 *
Hour17                  -1.382e+00  3.469e-01  -3.983 6.82e-05 ***
Hour18                  -1.292e+00  3.470e-01  -3.722 0.000198 ***
Hour19                  -7.072e-01  3.482e-01  -2.031 0.042287 *
Hour20                  -4.828e-01  3.483e-01  -1.386 0.165660
Hour21                  -6.908e-01  3.455e-01  -2.000 0.045560 *
```

**Figure 15.** Summary of the Linear Regression Model-Part A

```
Hour22                       -1.949e+00  3.442e-01  -5.663 1.51e-08 ***
Hour23                       -3.238e+00  3.412e-01  -9.490  < 2e-16 ***
RenewableProportion          -1.975e+01  6.609e-01 -29.887  < 2e-16 ***
Consumption                   1.449e-03  2.712e-04   5.344 9.21e-08 ***
FuelPrice                    -1.149e+01  1.082e+00 -10.617  < 2e-16 ***
Generation                   -9.323e-04  2.717e-04  -3.431 0.000603 ***
BilateralContacts.Consumption -1.482e+01  4.316e-01 -34.340  < 2e-16 ***
dayMonday                     2.137e+00  1.854e-01  11.525  < 2e-16 ***
daySaturday                  -4.065e-02  1.841e-01  -0.221 0.825309
daySunday                    -5.293e-01  2.022e-01  -2.618 0.008849 **
dayThursday                  -4.707e-01  1.824e-01  -2.581 0.009858 **
dayTuesday                    3.075e-01  1.823e-01   1.687 0.091646 .
dayWednesday                 -1.349e-01  1.823e-01  -0.740 0.459259
lag_24                        3.921e-01  6.337e-03  61.878  < 2e-16 ***
NegativeSpecialDays1         -1.060e+00  1.090e-01  -9.723  < 2e-16 ***
PositiveSpecialDays1          4.920e+00  2.415e-01  20.370  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.434 on 17458 degrees of freedom
Multiple R-squared:  0.6036,    Adjusted R-squared:  0.6028
F-statistic: 718.5 on 37 and 17458 DF,  p-value: < 2.2e-16
```

**Figure 16.** Summary of the Linear Regression Model-Part B

## Bayesian Network

A Bayesian network is a probabilistic graph model and expresses a set of random variables that have conditional dependencies with each other.

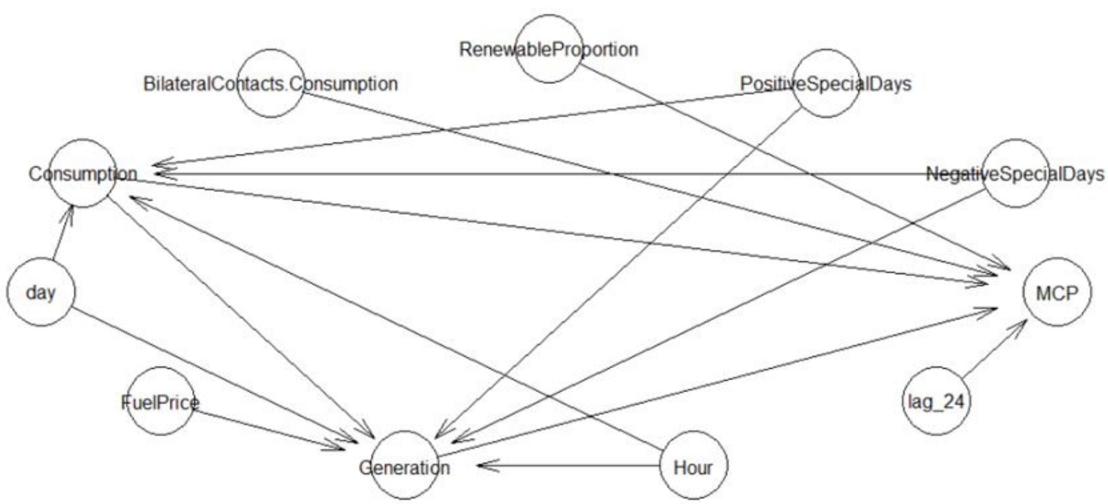



**Figure 17.** Graph of the Bayesian Network

The causal relations in the network are created investigating the data automatically. Then some arcs are updated by using our data analysis and market knowledge to obtain the final network.
Linear formulation of the MCP is given below. It is directly affected by the ratio of the bilateral contracts to consumption, consumption, generation, the proportion of the renewable energy generation and the MCP value of 24 hours before as demonstrated in the network in figure 16 above.

**SARIMA**

ARIMA (Autoregressive Integrated Moving Average) is a class of statistical model that is used in forecasting and analyzing time series data. ARIMA models are also capable of forecasting data that have seasonal properties. S in SARIMA stands for that seasonal component of the model. The formulation of the SARIMA model is as follows: ARIMA (p,d,q) (P, D, Q)m .

In this formulation p,d and q refers to lag order, differencing and order of moving average, respectively. Upper case letter notation is used for seasonal parts of the model, which functions the same as non-seasonal components, but instead, involves backshifts of the seasonal component.

In the literature of MCP, ARIMA models have been used extensively in different electricity markets due to its mathematical soundness and accuracy. Because of the satisfactory results obtained from different studies and suitability of the inherent structure of the model to the problem of interest, it is preferred to be used as one of the forecasting models in this work [1]. For the modelling phase, standard Box-Jenkins ARIMA methodology is applied. This methodology refers to the iterative application of the following steps [6].

1. Identification: Using plots of the data, autocorrelations, partial autocorrelations, and other information, a class of simple ARIMA models is selected. This amounts to estimating appropriate values for p, d, and q.

2. Estimation: The phis and thetas of the selected model are estimated using maximum likelihood techniques, back casting, etc., as outlined in Box-Jenkins (1976).

3. Diagnostic Checking: The fitted model is checked for inadequacies by considering the autocorrelations of the residual series (the series of residual, or error, values). Seasonality at multiple periods is observed in MCP as can be seen in Figure 1.

A SARIMA model with frequency 7 is developed for each hour and forecasting is performed accordingly. Thus, for each day, a total of 24 forecasting model results are obtained, each of which corresponds to an hour to be forecasted. The model is updated on a monthly basis during the forecasting test period. (For instance, the SARIMA model used for the test period January at Hour=12:00 is ARIMA(1,1,1)(2,0,0)7). The performance of the SARIMA model developed will be discussed in section 5.

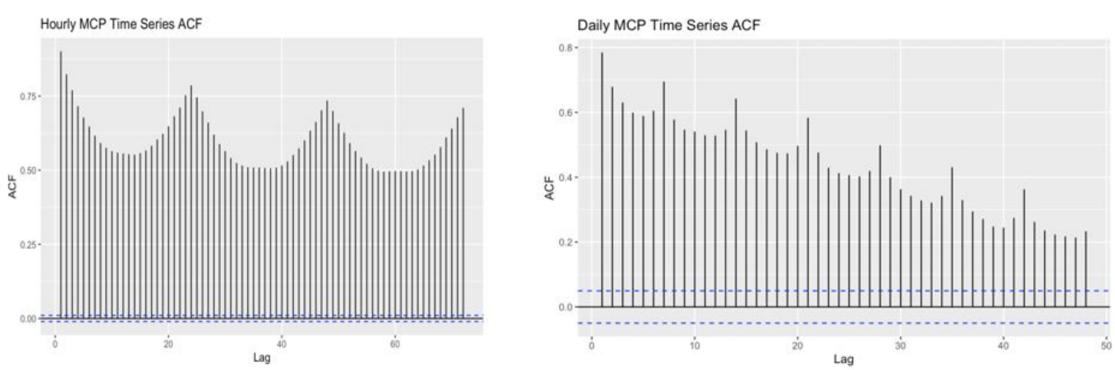



**Figure 18.** ACF Plots of MCP at Hourly and Daily Level

## Comparison of Alternatives and Recommendation

### *Numerical studies or evaluation procedure*

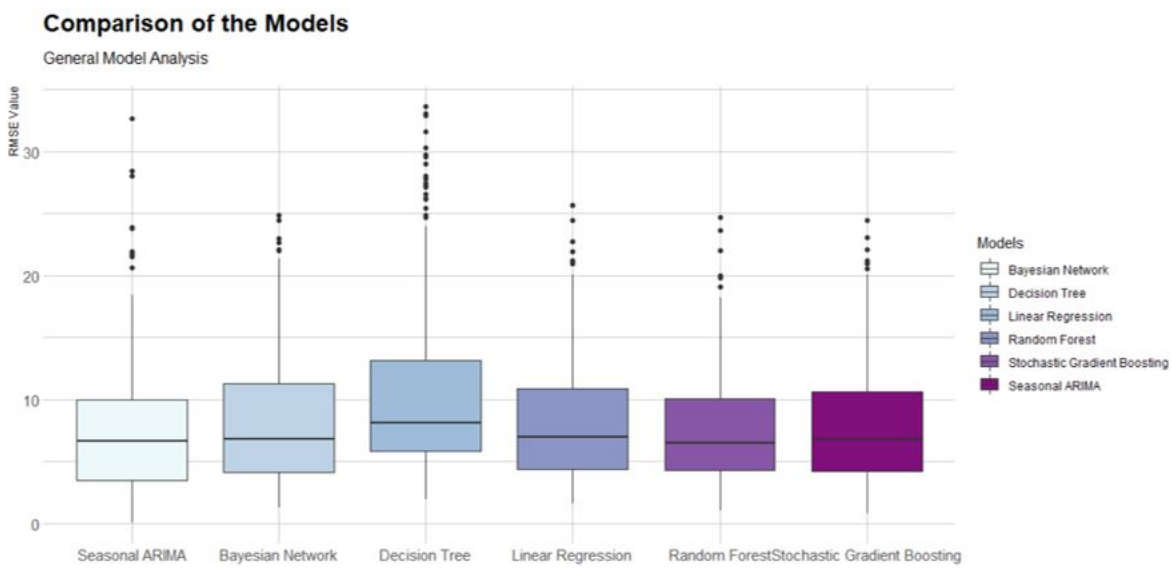



**Figure 19.** General Model Performance Comparison

RMSE, MAE and MAPE performance metrics, which are frequently used in the literature, are examined to evaluate the hourly, seasonal, and monthly performances of the established models.

However, it is decided that MAPE is not a suitable performance metric for the problem since it showed misleading results by taking a high value when the price is close to zero and a low value when the price is high. For this reason, RMSE and MAE values of the models are calculated to see hourly, seasonal, and monthly performances and their distributions are examined.

**Weekday Analysis:**

Almost all models perform worse to predict MCP on Sundays. In general, Bayesian network, random forest and gradient boosting perform well but their performances are not significantly different.

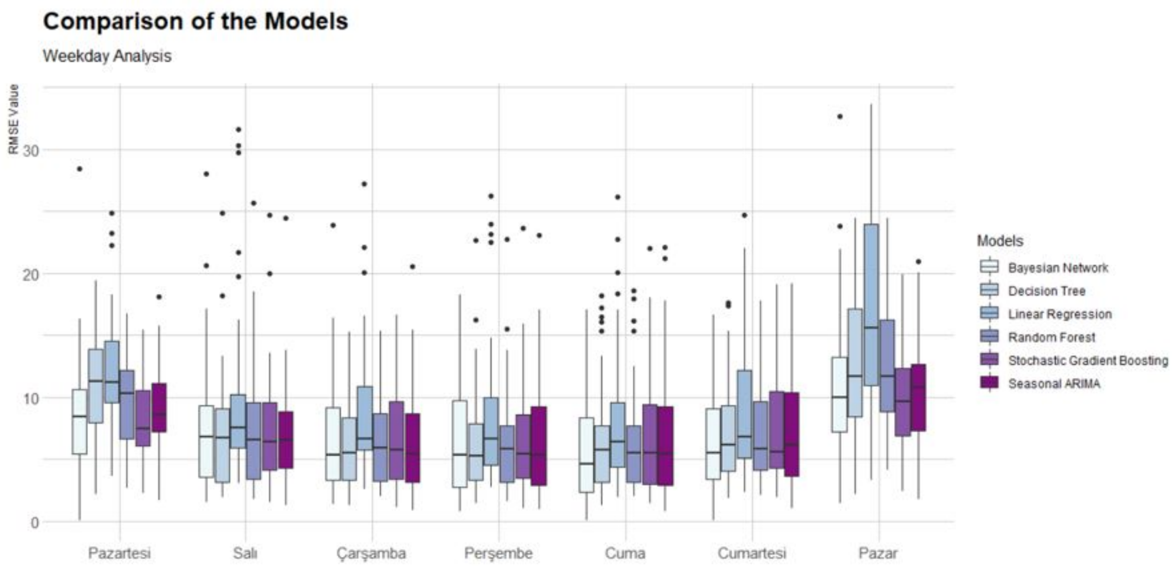



**Figure 20.** Model Performance Comparison for Weekdays

**Seasonal Analysis:**

All models perform better during fall and summer, and they perform relatively worse during spring. The fluctuations during spring times might be due to the increase in renewable energy proportion.Also, linear regression performs significantly worse. This can be because of the nonlinear volatile structure of the market clearing price.


**Figure 21.** Model Performance Comparison at Seasonal Level

## Hourly Analysis:

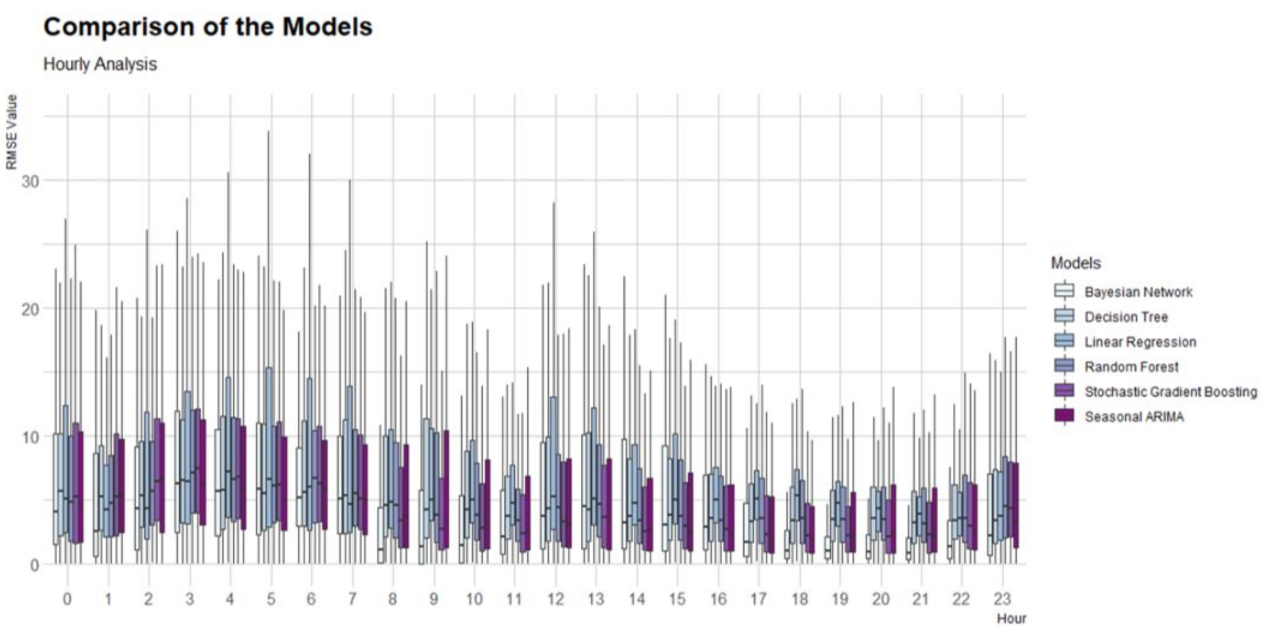



**Figure 22.** Model Performance Comparison at Hourly Level

In general, models perform better at working hours and perform worse between hours 0 and 7. All models deteriorate at lunch hours.

**Monthly Analysis:**

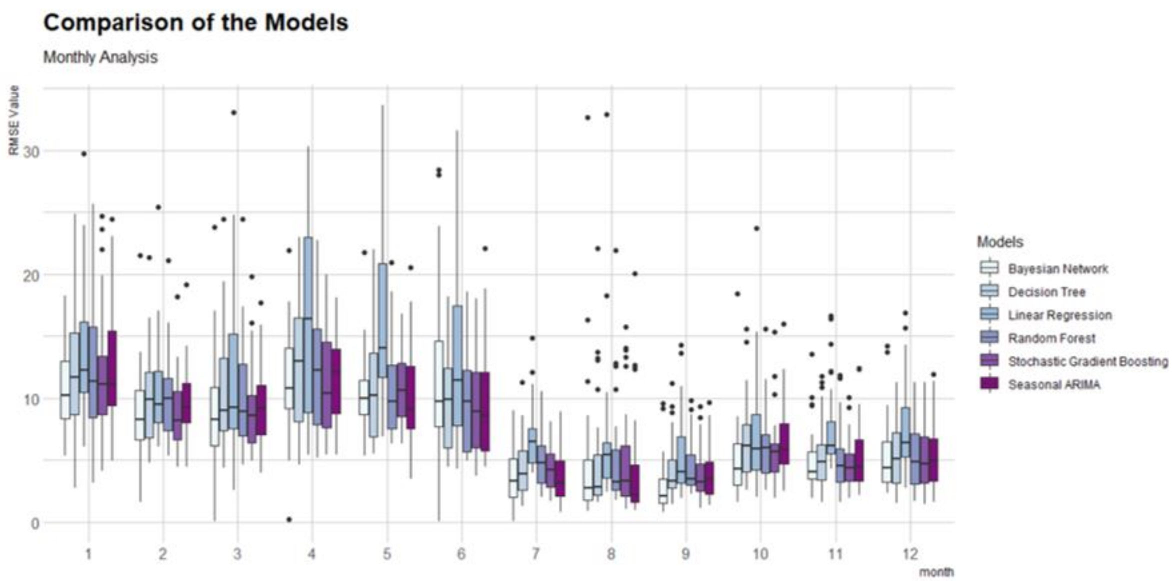



**Figure 23.** Model Performance Comparison at Monthly Level

Again, all models perform better during summer and models in general perform worse in April and May. This can be because of the renewable energy effect.

## *Proposing a promising solution along with justification in terms of predefined criteria*

Some data from more than a few years ago are missing and many of the macroeconomic data have different time horizon (like yearly GDP). Bayesian network model would be better under these limitations since it considers holistic causal structure unlike the other models.

### *Further assessment of the recommended solution*

Since the computational effort of running some of the models daily would be too much, the train data was updated monthly, and the models were run monthly. However, the Bayesian network works faster than the others and it is one of the top three models in general as in the figure. On the other hand, since the network is determined considering the structure of causal relations that is hard to change over time, it is possible to say the model is robust and sustainable.

## Suggestions for a Successful Implementation

### *How can the design be implemented?*

The implementation consists of two main steps: Collecting the necessary data and preprocessing it regularly and making predictions for the next day using the forecasting model developed. The first phase should be done for each hour and missing information should be imputed using several imputation methods. (Data of the last day can also be used if the information is available). Second phase, first, requires building and training the model which might take too much time. However, once the model is fitted, it is relatively easy to run the model and obtain the day-ahead predictions. Thus, this process can be scheduled and regularized. In fitting the model and making predictions, it is preferred to use the "monthly slide" approach, that is, the model is updated at each month so that the information from the new data is incorporated into the model itself. Similar procedures could be applied to improve the performance of the model developed.

Finally, after the performance analysis at different time levels, it has been observed that there is not just one model which outperforms the other in every time zone. Thus, using a combination of models, or using different models when it is at its best in terms of time (e.g., SARIMA model outperforms the others at specific hours, or during summer times Bayesian network model outperforms the others) could improve the prediction accuracy.

### *Can the design be integrated with the overall system?*

The design can be used by participants in the market. However, one must realize that the actions taken by participants after forecasting would differ. For instance, hydroelectric, wind, solar power plants, electricity retailers, distributors and utility companies mostly offer hourly bids whereas coal, natural gas-powered plants and large industrial consumers mostly offer block bids, in which offers can be made for more than one consecutive hour. Based on the forecast, a power plant might decide on whether to stop or continue the operations whereas consumers decide on to fine tune their consumptions. From a broader perspective, MCP forecasting might help the policy makers to adjust their strategies in the long term. In that sense, the forecasting model can be integrated with the overall system although the resulting actions of the participants could be varying.

Finally, it is important to mention that the forecasting model developed for the project is fundamentally different from the model developed by EXIST for price determination. The core aim of the forecasting model is helping agents (consumers, producers etc.) to adjust their bidding strategies to maximize their profits whereas, it is matching bidders maximizing social welfare in the EXIST model. Secondly, the forecasting model functions to forecast market clearing price (MCP) at hourly level for the next day before the market opens, whereas the model of EXIST must solve nonlinear integer programming (or find a reasonably good feasible solution) within minutes for each hour and each day.

### *How often should any solution obtained be revised?*

The performance of the predictions should be revised regularly to notice the extraordinary behavior of the model in time. However, there can be unexplained behaviors where the model fails to detect it. This can be due to major societal, political, or economic turbulence which has not been observed before, thus not considered by the modeler. (e.g., novel COVID-19 pandemic or dollar crisis in Turkey). To overcome those, the modeler should revise the model and try to find ways to incorporate those

factors into the model. Similarly, a model could perform poorly during a specific season due to its inherent structure. In that case, another model which can handle the seasonality better could be implemented instead. One of the significant results of the project was that some of the models can outperform the other during specific time intervals, due to its capability.
The MCP can easily be affected by sudden external policy changes or special occasions which might deteriorate the performance of the model. Those types of failures could be tolerated as it is not possible to perfectly predict them. However, if a failure in the predictions of the model is detected on a regular basis, this might be a sign of "overlooking a structural factor" that might have an effect on the dynamics of the MCP. Thus, in those cases, the developed model should be revised and, if possible, this factor should be included in the model.

## Conclusions and Discussion

### ***Summarize IE tools, techniques, methods that are integrated towards a successful completion of the project***

Industrial engineering techniques learned in statistics, probability, data mining, statistical forecasting and time series courses were often used in the project. Correlation analysis was performed between the parameters in the data and the output parameter market clearing price. In addition, in the regression analysis, the relationship of parameters with the market clearing price was analyzed by paying attention to the coefficients. For example, the special days parameter is divided into those that affect the price positively and negatively in subsequent analyses.
In the Bayesian network model, conditional probabilities which are often used in the probability course were used, and the causal relationship between parameters was studied. Random forest, decision tree, gradient boosting and linear regression models were created using the information learned in the data mining, statistics, and probability courses. In addition, the SARIMA model learned in the statistical forecasting and time series course was also used as a forecasting model in the project and

compared with other models.

### *Summarize the merits and significance of your design*

In the project, a Bayesian network model was established, a model that is less used than other models in the literature, and we brought another study into the literature in this area. In addition, in this project, which compared the price estimation performance of two linear models, linear regression and Bayesian network models, it was found that the Bayesian network works better in aggregate level data than linear regression.
In addition, since the performances of the models are compared on an hourly, monthly, seasonal, weekly basis, it is possible to observe the effects of different characteristics of the models on forecasting market clearing price performances.

### *Economic, environmental, ethical, and societal impacts of your design*

The project helps agents in the Turkish Electricity Market to determine their bidding strategies and contributes to profit increase, so its economic impact is significant. Manufacturers who can predict the market clearing price might adjust their capacity accordingly and can maximize the profit they can earn, but manufacturers who cannot predict the market clearing price accordingly might offer a low bid and even if they have capacity, they earn less profit. From another point of view, if the demand is low, the excess electricity produced remains to be sold at a lower price since electricity cannot be stored, and the profit falls.
On the other hand, the contribution and impact of renewable energy sources which are often mentioned in the project in electricity generation, makes the project important from an environmental point of view. According to the results obtained from the data analysis in the project, although most of the electricity production in Turkey is produced by non-renewable energy sources, there has been an increase in renewable electricity production, especially in the years after 2017. Instead of using fossil fuels that pollute the air and the environment by increasing the emission of greenhouse gases into the atmosphere, it is important to conclude that there has been a trend in

renewable electricity generation that has increased in recent years.
Finally, as for the social and ethical contributions and analysis of the project, it is estimated that some manufacturers who want to protect production facilities when prices are low do not produce and the authorities are insufficient to control the market. This situation theoretically prevents the establishment of a perfect market equilibrium. In addition, a more efficient market can be created by correctly estimating market clearing prices and preventing power outages that people will experience if manufacturers use their capacity in the most appropriate way.

## References

[1] Contreras, J., Espinola, R., Nogales, F. J., and Conejo, A. J. (2003). Arima models to predict next-day electricity prices. *IEEE transactions on power systems*, 18(3):1014–1020.

[2] Derinkuyu, K., Tanrisever, F., Kurt, N., and Ceyhan, G. (2020). Optimizing day-ahead electricity market prices: Increasing the total surplus for energy exchange istanbul. *Manufacturing & Service Operations Management*, 22(4):700–716.

[3] Georgilakis, P. (2006). Market clearing price forecasting in deregulated electricity markets using adaptively trained neural networks. pages 56–66.

[4] KÖLMEK, M. A. and Navruz, İ. (2015). Forecasting the day-ahead price in electricity balancing and settlement market of turkey by using artificial neural networks. *Turkish Journal of Electrical Engineering & Computer Sciences*, 23(3):841–852.

[5] Lu, X., Dong, Z. Y., and Li, X. (2005). Electricity market price spike forecast with data mining techniques. *Electric power systems research*, 73(1):19–29.

[6] NCSS (2021). The box-jenkins method.

[7] Ni, E. and Luh, P. B. (2001). Forecasting power market clearing price and its discrete pdf using a bayesian-based classification method. In *2001 IEEE Power Engineering Society Winter Meeting. Conference Proceedings (Cat. No. 01CH37194)*, volume 3, pages 1518–1523. IEEE.

[8] PricewaterhouseCoopers (2020). Overview of the turkish electricity market.

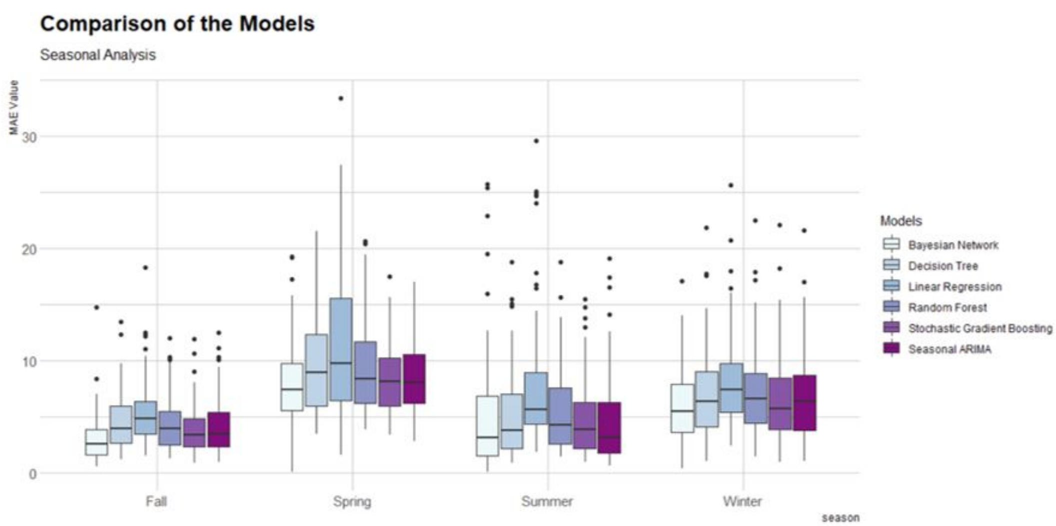
[9] Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International journal of forecasting*, 30(4):1030–1081.

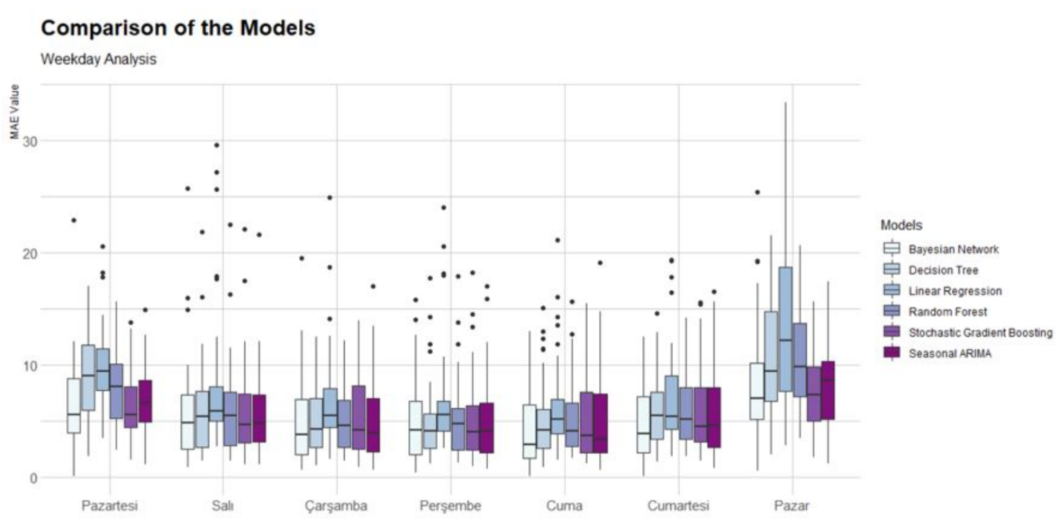
[10] Yan, X. and Chowdhury, N. A. (2013). Mid-term electricity market clearing price forecasting: A hybrid lssvm and armax approach. *International Journal of Electrical Power & Energy Systems*, 53:20–26.

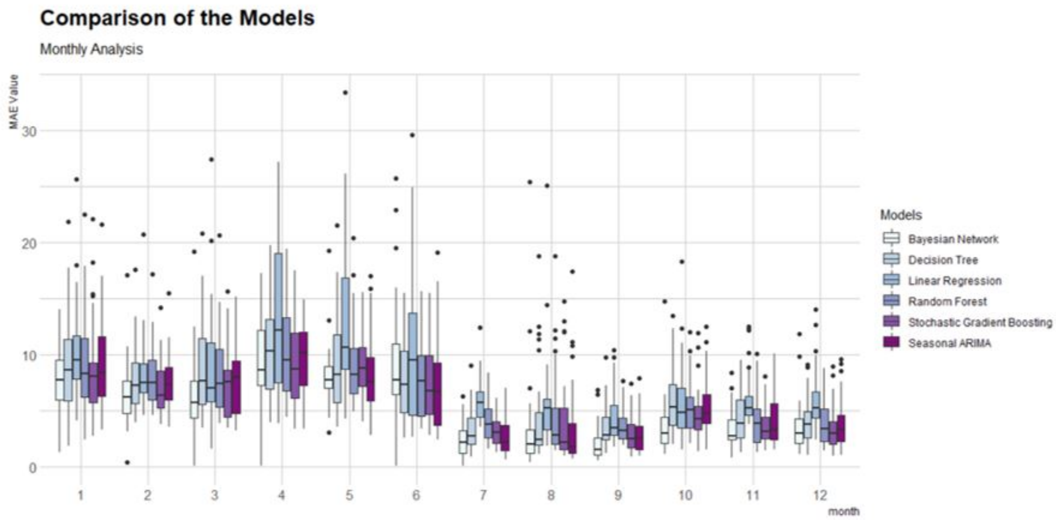
[11] Yan, X. and Chowdhury, N. A. (2014). Mid-term electricity market clearing price forecasting utilizing hybrid support vector machine and auto-regressive moving average with external input. *International Journal of Electrical Power & Energy Systems*, 63:64–70.

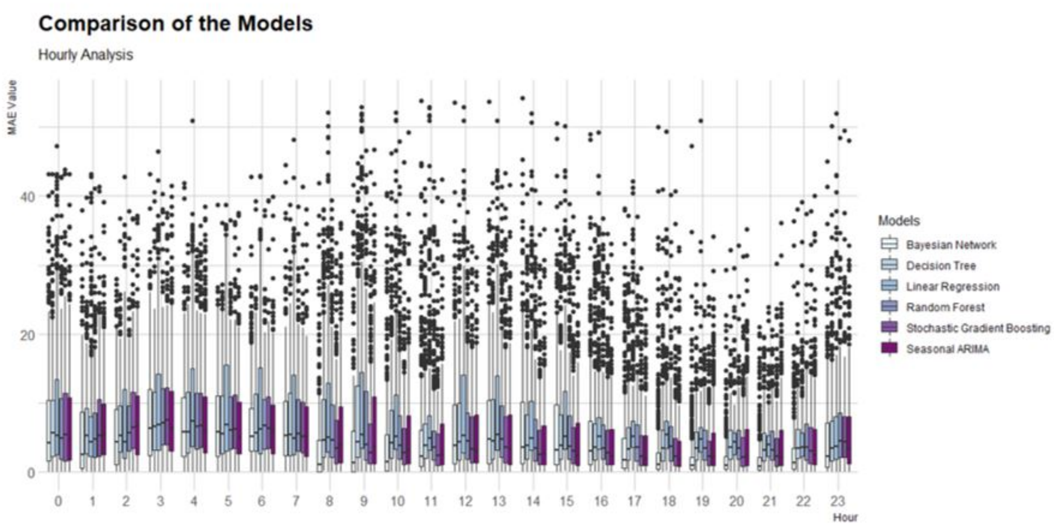
# Appendix

## *Appendix A- Comparison of Models with MAE*






## *Appendix B- Github Page of The Project*

https://github.com/alicanylmz/ie492-