BOĞAZİÇİ ÜNİVERSİTESİ

1863

Fall ‘21

IE 48B

# *Imbalance Prediction for Electricity Markets*

## Project Report

Mehmet Emin Öztürk 2017402000

Miraç Göktuğ Altuntaş 2016402198

# Introduction

In the modern world, one of the fundamental components of production is electricity. It is finite and needs efficient usage. At this very moment in Turkey, while this report is being written, most of the production sites have no electricity to operate. Which strengthens the argument of electricity being the most important feature.

Considering the size of the market, it is impossible for such an important resource to not get scarce. To understand it better, EPİAŞ is a good example. In this specific electricity market, there are some buyers (factories) and sellers (electricity producers). For every single hour, these two different sides act as a free market. One tries to buy with the minimum price and vice versa. If there is balance between supply and demand, there is a little room for profit. But most of the time there is either surplus or shortage. Which is the main problem for this project.

In this project, the main goal is to forecast this possible shortages, surpluses, or balances. To do that, the data of the last 3 years have been provided. This data contains the “Net” value, buy ins, and sell outs for EPİAŞ, and the weather information for every single hour. The aim is to forecast the “Net” status after 12:00.

To do that several approaches have been used in this study. Two different linear regression models and one tree model. After getting the predictions from them, the final prediction has been made as a hybrid model. When all models are giving the same result there is nothing to explain. But when there is conflict, the mean of the predictions is taken.

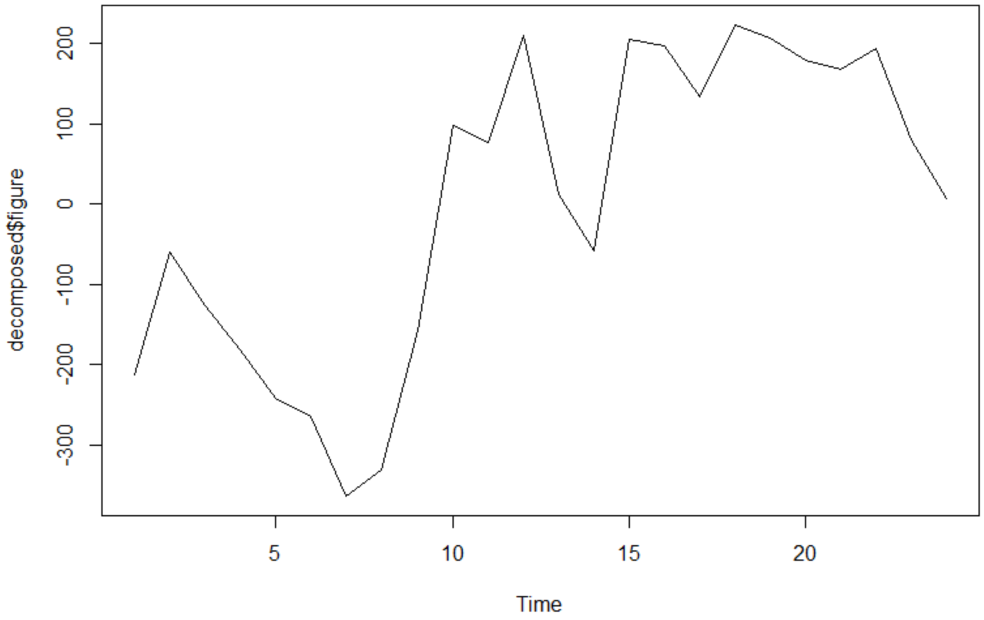
At this point it should be known that “Negative” class value means surplus and “Positive” class value means shortage.

Negative means “-1”, Neutral means “0” and lastly Positive means “+1”. If the mean X of the 3 predictions are $X< 0$, final prediction is Negative; if it is $X>0$, it is Positive and lastly if it is $X=0$ the prediction is neutral.

To make these predictions, it was crucial to understand the data.

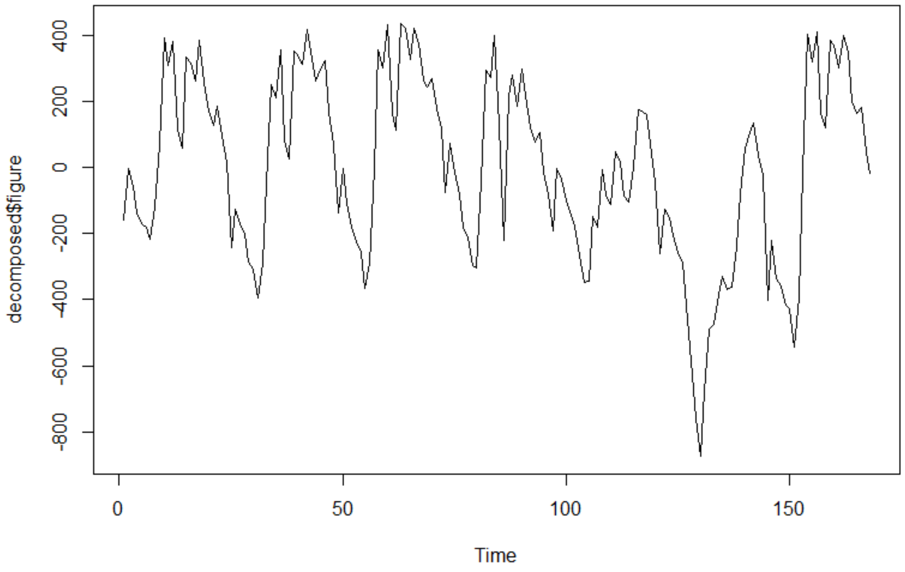
# Data

The data consists of 2 independent parts: Electricity and Weather. As the predictions were aimed to be made at hourly basis, it is beneficial to understand the effect of different hours during the day. To do this we've decomposed the data from trend and residuals to see the cyclic structure.



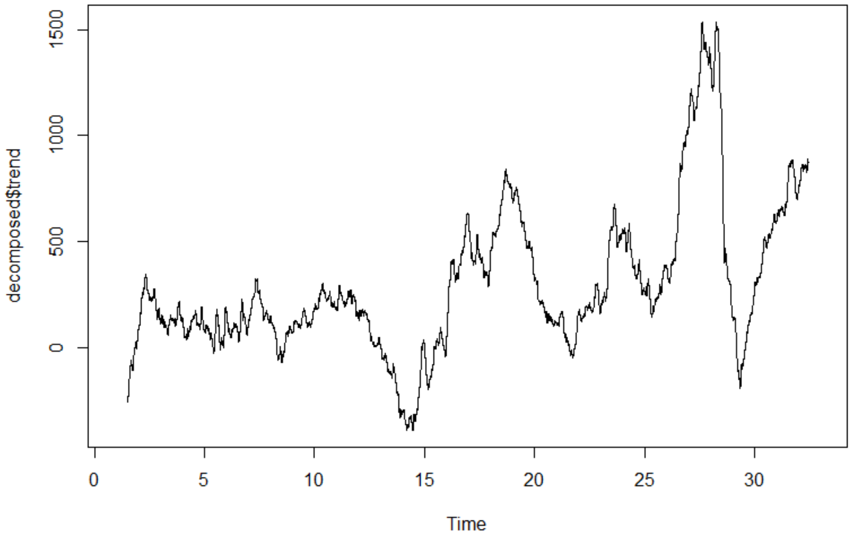
As it can be seen from the table there is a strong correlation between market status and working hours. In the working hours it is mostly positive with the little exception between 1 pm and 3 pm. Which may be resulted from lunch break.

The effect of working is also visible in weekly graph.



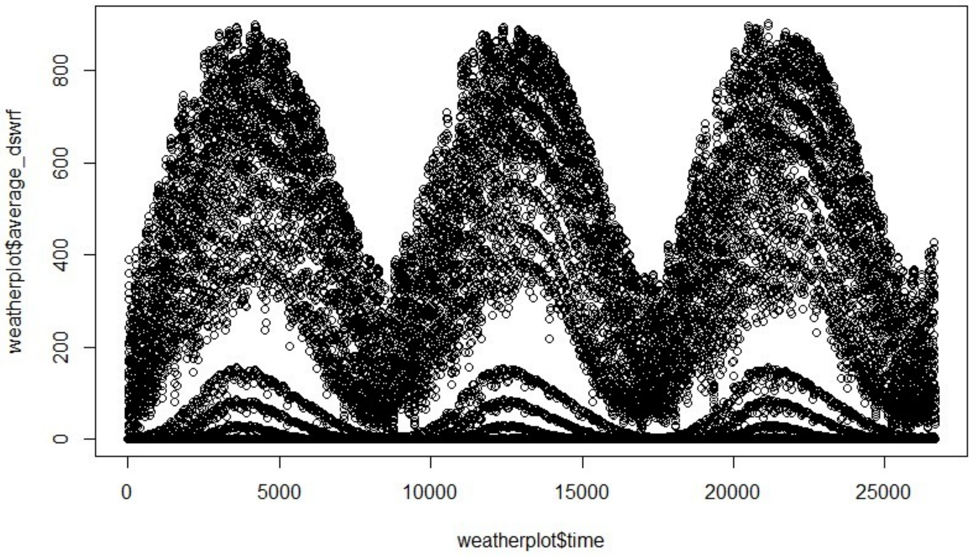
Since the data begins with 01.01.2019 Tuesday, the graph also begins with it. The presence of week days and weekends can be observed. When the factories are active there is a visible shortage represented by “Positive”. And when they are not they are mostly “Negative”.

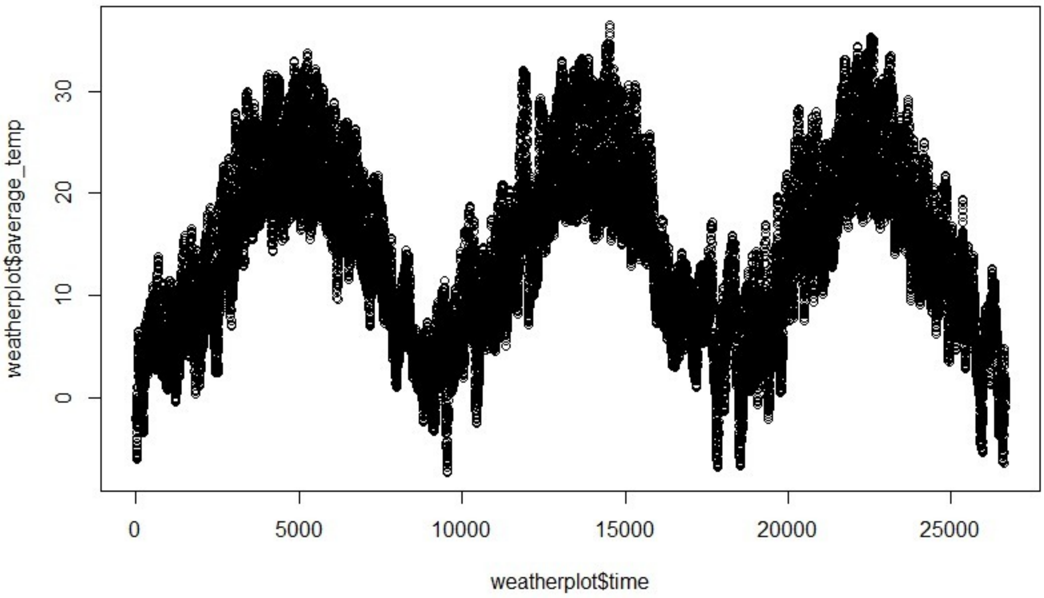
Both graphs were representing cycles without trends. Hence it would be beneficial to give a look at the trend of the electric data.



The trend shows us that even it may show some ups and downs there is a significant increase in “Net” if the beginning and the end of the data gets compared. Which means there is a shortage for the electricity nowadays (which can be also represented from submission data).

The weather data shows stronger signs of seasonality without much trend. For example, the temperature and the downward short-wave radiation flux graphs can be seen below.





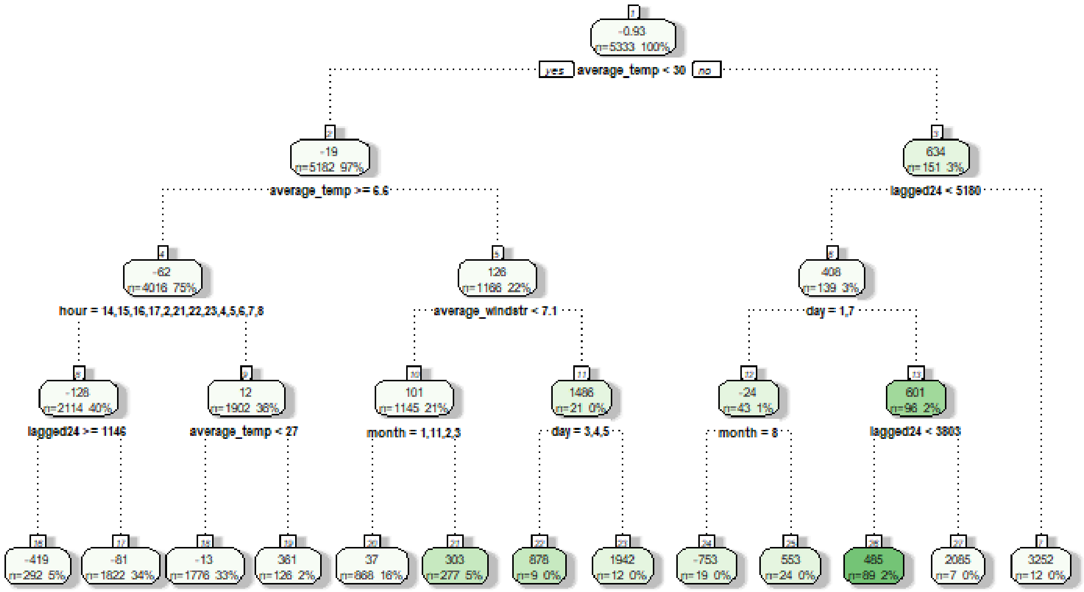
# Approach

The first model of the study was “developed” linear model. It was built with the present data and then some extra variables were added after examining the residual graph.

```
Call:
lm(formula = net ~ average_dswrf + average_rh2m + average_windstr +
    average_winddir + average_temp + average_tdcdlow + lagged24 +
    day + hour + month + lagged168, data = train_1)


Residuals:
    Min      1Q  Median      3Q     Max
-5545.3  -531.3   -18.2   494.6  6687.0


Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.043e+03  8.990e+01  11.598  < 2e-16 ***
average_dswrf   -1.892e-01  7.379e-02  -2.564 0.010346 *
average_rh2m    -1.581e+01  9.391e-01 -16.831  < 2e-16 ***
average_windstr -6.091e+01  6.002e+00 -10.147  < 2e-16 ***
average_winddir -2.655e-01  1.502e-01  -1.767 0.077256 .
average_temp    -2.994e+00  2.182e+00  -1.372 0.169982
average_tdcdlow  8.254e+00  4.574e-01  18.045  < 2e-16 ***
lagged24         4.163e-01  6.168e-03  67.488  < 2e-16 ***
day2             3.900e+02  2.337e+01  16.689  < 2e-16 ***
day3             3.208e+02  2.320e+01  13.826  < 2e-16 ***
day4             2.703e+02  2.326e+01  11.618  < 2e-16 ***
day5             3.309e+02  2.333e+01  14.179  < 2e-16 ***
day6             2.747e+02  2.328e+01  11.803  < 2e-16 ***
day7             1.482e+02  2.321e+01   6.386 1.74e-10 ***
hour1            9.072e+01  4.302e+01   2.109 0.034958 *
hour10           4.853e+01  4.969e+01   0.977 0.328759
hour11           5.484e+01  5.242e+01   1.046 0.295520
hour12          -4.682e+01  5.444e+01  -0.860 0.389856
hour13          -1.031e+02  5.607e+01  -1.839 0.065921 .
hour14           3.856e+01  5.688e+01   0.678 0.497892
```

```
hour15           9.050e+01  5.711e+01   1.585 0.113034
hour16          -7.369e+00  5.366e+01  -0.137 0.890762
hour17           8.752e+01  5.147e+01   1.700 0.089067 .
hour18           9.557e+01  4.873e+01   1.961 0.049877 *
hour19           1.234e+02  4.668e+01   2.643 0.008234 **
hour2            8.485e+01  4.279e+01   1.983 0.047372 *
hour20           1.640e+02  4.551e+01   3.604 0.000314 ***
hour21           2.086e+02  4.469e+01   4.667 3.08e-06 ***
hour22           1.501e+02  4.279e+01   3.508 0.000452 ***
hour23           1.332e+02  4.264e+01   3.125 0.001783 **
hour3            4.047e+01  4.281e+01   0.945 0.344493
hour4            3.859e+01  4.289e+01   0.900 0.368193
hour5            3.745e+01  4.289e+01   0.873 0.382589
hour6           -1.022e+01  4.311e+01  -0.237 0.812667
hour7            1.990e+00  4.269e+01   0.047 0.962826
hour8            2.241e+01  4.261e+01   0.526 0.598927
hour9            5.904e+01  4.337e+01   1.361 0.173482
month10         -1.045e+02  3.969e+01  -2.634 0.008445 **
month11         -9.857e+01  3.386e+01  -2.912 0.003600 **
month12         -6.876e+01  3.000e+01  -2.291 0.021946 *
month2          -5.175e+01  2.994e+01  -1.728 0.083916 .
month3          -9.592e+01  3.038e+01  -3.158 0.001592 **
month4          -2.005e+02  3.377e+01  -5.939 2.92e-09 ***
month5          -1.386e+02  4.090e+01  -3.389 0.000702 ***
month6           1.369e+01  4.537e+01   0.302 0.762799
month7           1.031e+02  4.955e+01   2.080 0.037493 *
month8           8.467e+01  4.982e+01   1.700 0.089221 .
month9          -1.282e+02  4.488e+01  -2.857 0.004277 **
lagged168        8.950e-02  6.154e-03  14.543  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 899.3 on 21282 degrees of freedom
Multiple R-squared:  0.2749,	Adjusted R-squared:  0.2733
F-statistic: 168.1 on 48 and 21282 DF,  p-value: < 2.2e-16
```

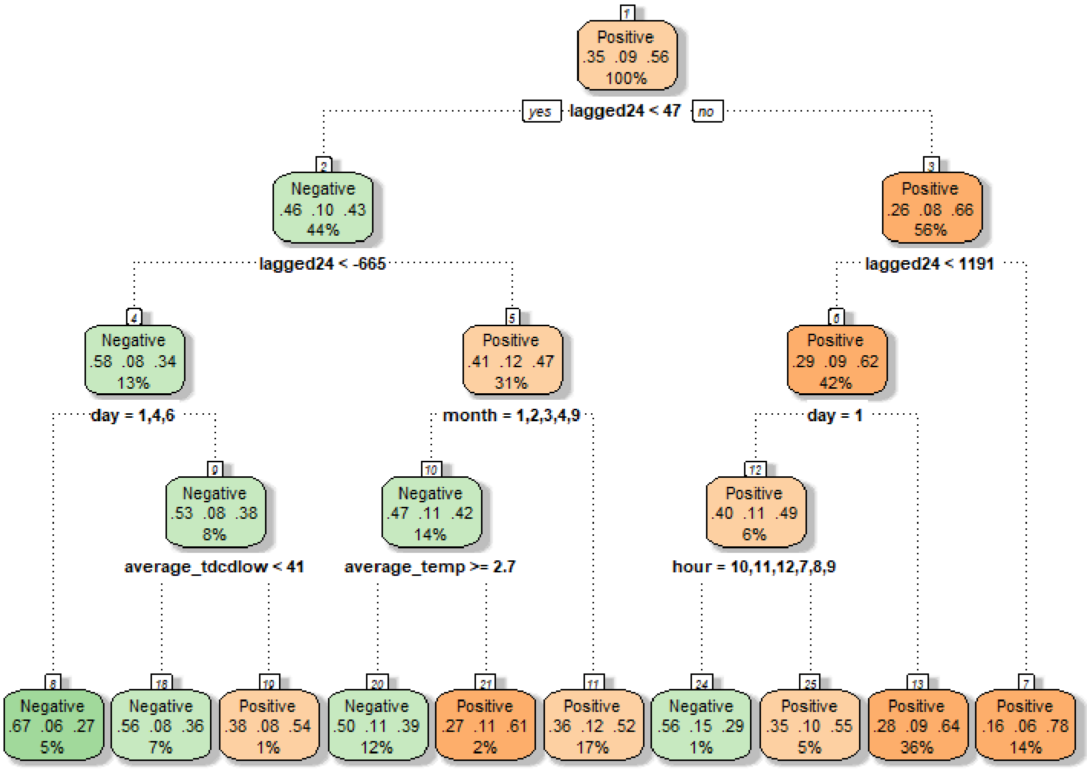
After making predictions with this model the residual tree was drawn to check for any improvements.



Rattle 2022-Oca-24 22:50:27 mehmet

We’ve decided to add 2 different variables for leaf 16 and 17 as most of the data is there and there is a slight bias. Then the predictions were made with this developed linear model.

The second model is based on tree approach.



Rattle 2022-Oca-25 00:18:45 mehmet

The distribution of the classes can be seen below. Even though there is no presence of “Neutral” class the result is better than both developed linear model and naïve in the all data.

Third and the last model is a new perspective for linear model. As the trend of the data is both increasing and unstable, it is decided to discard most of the data for the model. To make a prediction only the weather of the first 12 hours of the day were used. Then, by using this model the prediction is made for the PM values.

At this point it should be known that for the first 2 models the data is divided randomly to two as train and test. Which can also be seen in the code.

# Results

The naïve method was decided as a benchmark for different models. In the all data, it made correct predictions with about 54% probability.

The first model performed with 57% probability.

The second model performed with 60% probability.

All three percentages above show the success of the model on the full data with different train and test splits. However, the third approach uses a different model for each day, which makes it harder for it to calculate for whole data. Yet it is observed that it performed better than the first two in random dates.

As the third model doesn't give a certain result, the same condition also goes for the hybrid one.

Yet in the end, it can be said that the models could perform better in a more stable environment.

The further results can be observed in the code.

# Conclusion and Discussion

The results show that the present status of the electricity market is far from constant. The trend is unstable and increasing in a rapid manner. Even though it can be associated with the growing market, the reality is mostly due to the unstable economic atmosphere.

Besides that, the weather data is not sufficient to explain the movements of the market. Considering that the most of the electricity production comes from fossil fuels it is no surprise.

In addition to that, the submission phase came in an unfortunate time when factories were getting no electricity and the electric is scarce.

To get better results the live exchange rates and the price of natural gas may also be used for this study.