EXAMINATION OF OPPORTUNITIES ON PARTICIPATING IN SEQUENTIAL MARKETS AND IDENTIFICATION OF POSSIBLE SYSTEMATIC PRICE DIFFERENCES

Written by: Johannes Weyers, Julian Cornea, Kim Ferres, Mert Erdogan and Rufin Korbmacher

University of Cologne
Advanced Analytics and Algorithms

1. Business Goal:

In the following, the intended business goal of examining sequential markets is briefly explained in its essential core points:

1.1 An energy trading company represents the client in the case of this investigation.

1.2 This stakeholder wants a study of price fluctuations and deviations in sequential markets in order to exploit arbitrage opportunities. In this way, the profit from energy sales for the company can be optimized or even maximized. Therefore, the client could benefit significantly from the analysis and prediction of price differences, fluctuations and volatilities.

1.3 A permanent implementation of the data analysis required to fulfil the pursued business goal would allow the company to split its available energy in the different markets more effectively. The planned undertaking could also have the side effect that mostly incalculable energy yields, such as the yields through different renewable energies sources, could experience a better integration into the system of the existing energy market. More specifically, these incalculable energy quantities could benefit from strengthening the short-term electricity markets due to a more strategic placement of energy across different markets. The business goal is related to the problem of the insufficient orientation of energy prices to the respectively current demand. As well it addresses the challenges created by parallel participation in several markets, which has become essential mostly due to renewable energies.

1.4 Identifying a systematic in the price differences of sequential markets and through this deriving a recommendation for pricing auction offers would be considered a success. From that would increase the effectiveness of dealing with the energy market's price volatility and this would ensue a decrease of current opportunity costs.

2. Analytics and Data Mining Goal:

2.1 Accordingly, to the issue description of the analytics objective it is necessary to identify parameters which are responsible for the existing price difference between the day-ahead and intraday auctions. Based on these parameters it would be needful to develop a regression model to emphasize this issue in a graphical way.

2.2 Furthermore, this task is supervised. The reason in order to that is the available price data. The first part of the regression is obviously descriptive, but the second part cannot exactly be estimated. It is predictive and things with a predictive character are never exact. Both parts are also retrospective. But the regression model can additionally be used for prospective predictions, required the utilized parameters data is available at the point of prediction.

2.3 In term of the main outcome variables it is important to consider several parameters which have a causal relationship with the prices at both auctions and the kind of their relationship.

3. Data:

3.1 The dataset to be examined shall include volume (in MWh) and price (in €/MWh) of the energy sold at each auction sale. For the day-ahead auction the data is provided in hourly-, for the intraday auction in 15-minute resolution. As a whole the dataset encompasses 731 days, so two years, which results in 17544 data points for the day-ahead data and 70176 data points for the intraday auction data of the year 2016 and 2017.

3.2 The first step in order to create the dataset to be used for advanced analysis will be the scraping of relevant information from the respective websites. After the successful collection and persistence of day-ahead and intra-day auction price and volume information the hourly data needs to be converted into 15-minute increments and attached to the data file of the intra-day auction data. Since there is no missing data, no respective preparations are required. As a unique identifier of each row in the merged data file, a timestamp will be used. Additionally, the difference between the prices of the two power markets are mapped to a new column, which is called price premium throughout our analysis.

3.3 In order to get an overview of the features and the nature of the data, a copy of ten rows of the dataset is provided.

Timestamp	15-minute increment	MWh (day-ahead)	€/MWh (day-ahead)	MWh (intraday)	€/MWh (intraday)
2017-01-01 00:00:00	00:00	69,20	2736,55	46,56	616,5
2017-01-01 00:15:00	00:15	69,20	2736,55	39,96	282,6
2017-01-01 00:30:00	00:30	69,20	2736,55	34,44	257,1
2017-01-01 00:45:00	00:45	69,20	2736,55	27,18	426,2
2017-01-01 01:00:00	01:00	61,07	2642,975	38,68	456,9
2017-01-01 01:15:00	01:15	61,07	2642,975	34,1	326,7
2017-01-01 01:30:00	01:30	61,07	2642,975	40,52	396,2
2017-01-01 01:45:00	01:45	61,07	2642,975	36,41	404,5
2017-01-01 02:00:00	02:00	55,47	2651,8	39,83	348,2
2017-01-01 02:15:00	02:15	55,47	2651,8	37,76	231,4
2017-01-01 02:30:00	02:30	55,47	2651,8	35,24	277,4
2017-01-01 02:45:00	02:45	55,47	2651,8	32,79	288

Figure 1: A sample overview over the first few lines of the data

4. Methods:

4.1 To find underlying patterns in the given dataset based on the price premium, a supervised data mining task needs to be solved. This results out of the fact that price premium is our target variable which should be explained through a good fitting descriptive model. The values of price premiums are already known prior to the model step in our analysis. Due to the fact that the target variable is continuous regression models will be considered. For the following step, a forecast on whether the price premiums will rise or fall in the near future, a predictive model is necessary which can be the previous build regression model, but not necessarily has to. In order to get discrete values from the pre-trained regression model the output needs to be categorized in rise or fall through a threshold. Apart from that model any classification model can be used for this predictive task as well. For example, Neural Networks, Decision Trees or Naïve Bayes. The final model selection process will be determined through performance measurements. Hence the choice is objectified.

4.2 The crucial point in any data mining task is evaluation of models and therefore selecting the right performance measures, not only to select the best model, but also to infer the real performance without the interference of random data constellations. Each task implies specific Evaluation approaches and therefore specific measures. The descriptive task needs a measure to evaluate the goodness of fit and the fitting of the regression. The predictive task requires a good reflection of the performance on new data. A common method to simulate the performance on new data is to split the original dataset into training and holdout data. From that, advanced measures are derived from the comparison of actual prices of a holdout dataset and the predictions of an applied predictive model. Furthermore, cross-validation as an industry standard is used to validate the model and test the generalization performance.

4.3 The goodness of fit shows if the derived parameters are responsible for the price difference between the auctions. The prediction accuracy shows how well the causal relationship between the parameters and the price has been estimated.

5. Implementation and Production:

5.1 The predictive model will not be continuously run in real-time, however it certainly can be used for that. This would however require automated data collection, an automated data preparation, a thorough monitoring system, and a backup system in case the prediction performance deteriorates too much.