Data Science Methods in Finance

Take Home Exam

Jens Kvaerner, Ole Wilms

Submission deadline: April 2, 2021 at 23.59

Important Instructions

- This is an **individual assignment** and hence, you must work on it by yourself.
- Copying codes from others will be counted as fraud and lead to a failure of the exam as well as a report to the exam committee.
- You are allowed to use ment.io to discuss questions with other students. However, the questions must be about the general task of a problem. For example asking other students to look for errors in your code is not allowed.
- **To submit your final report and code, please use the corresponding Assignment on Canvas. Do NOT send files via mail as originally stated in the syllabus.**
- **Late submissions will not be considered and result in a grade of 0.**
- **Please do NOT submit datafiles. Only use the original datafile from Canvas and upload your codes and report.**
- The code must run without errors and generate the same solutions as you have reported.
- The data you will use is "2021_Electricity_data.RDS" available on Canvas by March 31.
- The exam has a total of 5 questions with multiple sub-questions which account for a total of 100 points:

Question	1	2	3	4	5	Total
Points	18	16	18	25	23	100

- The bonus points you earned for the participation on ment.io will be added to the points you achieve in this exam (maximum 10).
- We care about the layout, and we want a “consultant" professional report. That means you should prepare a set of slides, and not a long document (more on this below). We give up to 5 bonus points for good layouts.

General Information

For this assignment, please create a report in power point or with Beamer (we recommend overleaf) where you summarize the results. Each item should be answered on a different slide. Only if it is mentioned otherwise in the subquestion or if you think it is really necessary, you can use more slides. Try to be precise. Ambiguous and overly long answers will lead to a reduction of points. Put additional explanations or calculations in the appendix. Start all headers with the question you are answering. For example, when presenting summary statistics for the first question, write “Q1.1: Summary statistics", etc. Make sure that the font size is well readable (not too small) and that you do not put too much information (text) on one slide. You might want to use wide format.[1]

[1] The template “example_template.pdf" on Canvas is a good starting point. We strongly encourage you to follow this structure as close as possible.

Dataset and Setup

In this exam you are asked to predict hourly energy prices in the Netherlands and build a trading strategy based on your predictions. The data you will use for this exam is "2021_Electricity_data.RDS". It contains, among other variables, hourly day-ahead electricity prices in Euro/MWh for the Netherlands from 2018-2020. The power prices are made publicly available the day before delivery. For example, in the dataset you see the price 2018-01-02 14:00 of 46.10 Euro. This means that this is the price of 1 Mega Watt an Hour (MWh) for delivery between 2018-01-02 14:00 and 2018-01-02 14:59. This price is made public on 1-Jan-2020 (hence the name day-ahead). The main goal of this assignment is to predict day-ahead prices and build a trading strategy based on the predictions.

Forecasting Problem

After having taken the Data Science Course at Tilburg University, you want to apply your new knowledge and make money by trading in the electricity market. For that you bought Tesla batteries with a total capacity of 1 MWh from which you can buy and sell electricity to the grid. Your goal is to apply machine learning techniques to forecast electricity prices in the Netherlands and build a profitable trading strategy based on the predictions.

Market Setup

The market you are looking at is the so called day-ahead market. The market works as follows:

- Each day at 14.00 you must submit an order at which hours of the subsequent day you want to charge or discharge your battery. In this order you must specify how much electricity you want to buy or sell, however the price at which you trade is yet unknown.
- Each day at 16.00, the hourly day-ahead electricity prices at which your order is executed are released.
- Assume for example that you decide at 14.00 to buy 1 MWh at 8.00 the next day and you sell it at 20.00. At 16.00 today the day-ahead prices are released and it turns out that the 8.00 price is 23.01 Euro/MWh and the 20.00 price is 32.21 Euro/MWh. Your profit is hence (32.21 Euro/MWh - 23.01 Euro/MWh) x 1 MWh = 9.20 Euro.
- So each day at 14.00 you try to forecast day-ahead prices for tomorrow. For this you can use all the hourly information (day-ahead prices and other predictor variables) from the previous day.

Variables in the Dataset

Note that all variables are preprocessed so that they are aligned with regard to their time of publication. For example all variables with date 2018-01-02 are available the day before at 16.00.

- *dutch_*, *german_*, *norway_* and *belgium_power*: day-ahead electricity prices in the corresponding country.
- *dutch_load*: Day-ahead load forecast in MWh on the electricity grid. On the power grid supply equals demand. Hence the load forecast can be seen as a demand forecast.
- *dutch_generation_forecast*: MWh estimate of day-ahead generation (supply) of dutch power.
- *solar*: solar power generation forecast in MWh.
- *wind*: wind power generation forecast in MWh (on-shore: windmills on land; offshore: windmills in the sea).

Question 1 (18 points): Report summary graphs for the data

1.1 Plot average hourly prices over a day (x-axis hours from 1-24, y-axis average price in the first hour of the day, second hour of the day,...). In the same graph add average hourly prices over a day but only take into account weekends

1.2 Plot the volatility of hourly prices over a day (x-axis hours from 1-24, y-axis volatility of price in the first hour of the day, second hour of the day,...)

1.3 Plot average daily prices over a year. For this, first take daily averages and then compute average daily prices over the year. (x-axis days from 1 to 365, y-axis average over hourly prices per day and average over the three years)

1.4 Plot average hourly solar generation over a day. Once for the months from April-September and once from October-March (both in one plot)

1.5 Plot average daily solar and off-shore wind generation over a year (x-axis days from 1 to 365, y-axis average over hourly generation per day and average over the three years)

Briefly interpret your results. You might want to look at other summary statistics as well to obtain more information about the data but you do not have to report these in the write-up.

Question 2 (16 points): Prepare the dataset for making daily price predictions

We are interested in making daily forecasts of hourly prices. For example, we might be interested in predicting the price in the first hour tomorrow, using all the data that we have available today—that is, the previous 24 hourly prices and the hourly values of the other predictors.

2.1 **Prepare the data:** The first 24 columns, should contain the hourly dutch electricity prices, the following 24 columns, the 24 hourly German electricity prices and so on. So at the end the dataframe should be of size $n \times 24p$ where n is the number of days in the dataset and p is the number of features (9 in our example). Appendix 1 contains a draft of what the dataframe should look like. Report a table which shows a subset of your data.

2.2 Split the data into a dataframe with the outcomes ($((n-1) \times 24$ of hourly electricity prices) and the features ($(n-1) \times 24p$) features lagged by 1 day. (No output slide needed.)

Important: If you are not able to construct the data above, you can use the pre-processed data "2021_Electricity_data_x_preprocessed.RDS" (feature data, all variables are already lagged by one day) and "2021_Electricity_data_y_preprocessed.RDS" (outcome data) available on Canvas. However, if you do so, you must clearly indicate this in your report (it will be considered as fraud if you do otherwise). This will lead to 0 points for the first part of this question.

2.3 **Clean the data**

- Get rid of the features in this dataframe that have too many 0 entries. It's up to you to choose a reasonable cutoff. Report the number of features in your dataset.
- Split the sample into a training set of 2.5 years and a test set of 0.5 years. Explain how you split the data.
- Normalize the data so that all features have mean 0 and a standard deviation of 1. Explain how you normalize the training and test data.

Question 3 (18 points): Predicting electricity prices in the first hour of the day

In this exercise, you are asked to predict the dutch electricity price in the first hour of the day, that is the price from 0.00-0.59. You can use all the information available the day before. For example, you make the prediction at 14.00 on a given day where the day-ahead prices are not released yet but all the day-ahead prices from the previous day are available.

3.1 **Build a benchmark model**. Use a linear regression to predict the dutch electricity price in the first hour of the day using the price in the last hour of the previous day (and a constant) as a predictor. Report the in-sample R^2 for the training set as well as the out-of-sample root mean squared prediction error (RMSE) for the test data. Interpret your results.

3.2 **Machine learning models:** Use different machine learning models to predict the electricity price in the first hour of the day. You can use all the available features or a subset. Please motivate your choice. To tune the hyper-parameter(s) of your method, split the training sample into training (first 2 years) and validation data (last half year). Report the optimal tuning parameter(s) as well as the in-sample R^2 and the out-of-sample RMSE for the test data. Interpret your results for the following machine learning techniques:

- Lasso
- Partial Least Squares
- Random Forest

(Prepare one slide for each model)

Question 4 (25 points): Predicting hourly electricity prices

Now redo Question 3, but not only for predicting the price in the first hour of the day, but to predict all 24 hourly prices. That is, build 24 models where the first model predicts the price in the first hour (as in Question 3), the second model predicts the hourly price from 1.00-1.59, the third model predicts the hourly price from 2.00-2.59 and so on. Note, that all of these models are independent, but they all use the same input data.

4.1 **Build a benchmark model**. Use 24 linear regressions to predict the 24 hourly dutch electricity prices over the day using the price in the last hour of the previous day (and a constant) as a predictor. Plot the in-sample R^2 for the 24 models in one graph (x-axis goes from 1 to 24 for the 24 hourly prices we predict, y-axis shows the in-sample R^2 for each horizon). Plot a corresponding graph for the out-of-sample RMSE and interpret the figures. (Prepare two slides, one for each figure.)

4.2 **Machine learning models:** As in Question 3.2, now use different machine learning methods to predict the 24 hourly electricity prices. That is, using each method build 24 models. Tune your hyperparameter(s) for each of the models as in Question 3.2 using the same training and validation set. Use the following machine learning techniques:

- Lasso
- Partial Least Squares
- Random Forest

Add the in-sample R^2 for each model to the plot from Question 4.1. So the plot should contain 4 lines, one for each model. Plot a corresponding figure for the out-of-sample RMSE and interpret your results. (Prepare one slide for each model.)

Question 5 (23 points): Trading strategy

The final task is to build a (profitable) trading strategy based on your forecasts. You have a Tesla battery with a capacity of 1 MWh from which you can buy and sell electricity to the grid. Each day at 14.00 you must submit your order at which hours you want to buy or sell how much electricity for the subsequent day. So we need to predict day-ahead prices for tomorrow using the feature data of today as in Question 4 and build a trading strategy based on these forecast.

5.1 **Benchmark strategy**. First we build a simple benchmark trading strategy. Each day, we want to charge our battery when the price is lowest on average and discharge the battery when the hourly price is highest on average. For this, compute the hour of the day with the lowest and highest average price. Charge your battery at the lowest and discharge at the highest price. Report the cumulative profit as well as the daily volatility of this strategy for the test data.

5.2 **Machine learning models**. Use the 24 hourly forecasts from the Lasso model from Question 4.2 to predict the hour with the lowest and the hour with the highest price for the subsequent day. Charge your battery at the hour with the lowest predicted price and discharge at the hour with the highest price. If the highest price comes before the lowest price, do not charge your battery at all on this day. Report the cumulative profit as well as the daily volatility of this strategy for the test data. Redo the exercise using your forecasts from the Partial Least Squares model and the Random Forest from Question 4.2 and compare your results.

5.3 **Build your own trading strategy**. Instead of only charging and discharging once a day, now build a trading strategy that uses all the 24 hourly predicting. You are free to charge and discharge your battery at any hour of the day. But make sure that after charging, you must discharge first before you can charge again because of the maximum capacity of 1 MWh of your battery. Apply the trading strategy using your best performing machine learning model from Question 4 and report the cumulative profit as well as the daily volatility of this strategy for the test data.

1 Appendix

Table 1: Example of data structure

Date	dutch_power_1	dutch_power_2	...	dutch_power4	german_power_1	german_power_2	...	german_power_24	...
2018-01-01	27.20	27.30	...	23.79	-5.27	-29.99	...	18.96	...
2018-01-02	25.07	23.19	...	28.51	18.12	14.99	...	9.94	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮