AAA: Report

**(b) Data description (the data that you end up using: size and dimension, what is a record, list of output and input variables, sample of 5 rows).**

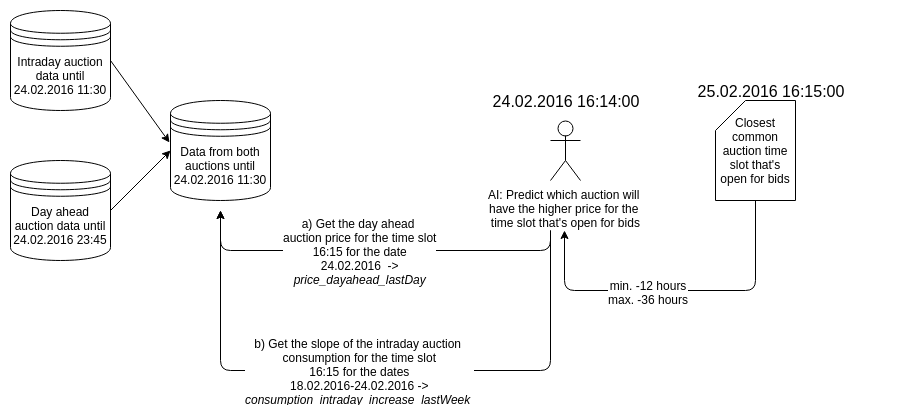
Same data set for descriptive and predictive, data set = basis (final.csv) set without price premium, proposal + scraping of markt participants, weather data, holidays

**(c) Brief data preparation details (how your data were created from the raw data) and key charts.**

General: Scaling, weather data drop, price premium

After analyzing the scraped weather data, we discarded the data containing information about the monthly number of days with snow and the monthly snow volume, due to extreme data scarcity causing the aggregated data to be too biased to be of any use.

Additionally, we discarded the data containing information about the average minimum and maximum daily temperature per month. The reason for that was that this data did not contain useful information in addition to the average of average daily temperature per month and leaving it would cause unnecessary multicollinearity.



Predictive:

In order to prepare the dataset for the predictive approach, the data had to be applied furthermore. Initially, the specification of the datetime was dropped and instead recorded by one hot encoding of the months, weekdays and hours. Furthermore, the original premium price was removed and replaced by a categorical value. This value indicates whether the price premium is positive or negative. To add a history to the records, the most relevant energy consumption and price for both auctions were calculated for each time slot (see figure xy). The most relevant data is made up of the auction data at the last date with that time slot, which is available at the prediction time, and the slope for the week before the target time slot. Because of this historic approach, the first eight days could not be included in the predictive dataset.

**(d) Data mining solution: Methods applied (with sufficient detail and screenshots; use Appendix if needed) and appropriate performance evaluation (proper choice of measures, benchmarking).**

For the predictive classification the following models were used: SGD Classification, Binary Support Vector Classification, Gaussian Naive Bayes, Soft Voting Ensemble, Hard Voting Ensemble, Logistic Regression, Recurrent Neural Network, Decision Tree, ARIMA and Random Forest.

For each model a time series compatible grid and random search was implemented. After the search finished, the hyperparameter combination of the best performing model was picked and evaluated on the full dataset using daily split increments. Exemplary screenshots for Logistic Regression can be found in the Appendix. The implementation of a time series compatible search can be seen in figure xy. The analysis of the hyperparameter search is shown in figure xy. The implementation of the final run is visualized in figure xy.

For a performance comparison of the applied models see figure xy. Plot in Text?

The best performing model was XY with an accuracy of XY and ROC-score of XY. The confusion matrix can be found in figure xy. Analysis of matrix. The plot of the ROC curve is visualized in figure xy.

**(e) Conclusions (advantages and limitations) and operational recommendations**