AAA: Report

**(a) Problem description (business goal and data mining goal)**

Proposal + better relevance: high costs if promised energy cant be delivered at intraday market

**(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).**

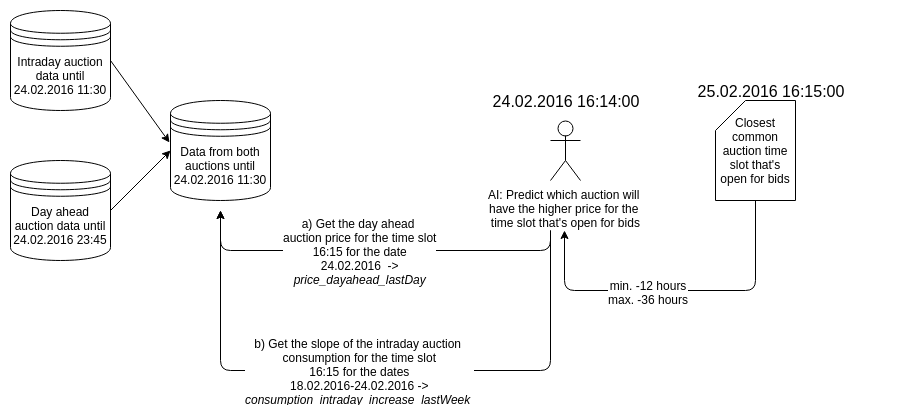
Same dataset for descriptive and predictive, dataset = basis (final.csv) set without price premium, proposal + scraping of market participants, weather data, holidays

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

General, additional to proposal: Scaling, calculation of price premium

An analysis of the scraped weather data revealed that the data containing information about the monthly number of days with snow and the monthly snow volume had to be discarded, due to extreme data scarcity causing the aggregated data to be too biased to be useful.

Additionally, the data containing information about the average minimum and maximum daily temperature per month had to be removed from the dataset. The reason for this step was that the named data did not add valuable information in addition to the average of the daily mean 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 represented by one hot encoding of the respective month, weekday and hour. Furthermore, the original price premium was removed and replaced by a categorical value. This binary value indicates whether the premium is positive or negative. To add a history of past energy consumption and price to the records, the most relevant historic energy data for both auction types were calculated for each time slot (see figure “Aggregated Columns Example.png”). The most relevant data for a time slot includes the auction data at the last date, which is available at the prediction time and which also contains the slot. Furthermore, it comprises the slope value of price and consumption for the week before the target time slot. This additional historic approach causes the exclusion of the first eight days for the predictive dataset since the former energy data for these entries is not included in the 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 trained and tested: Stochastic Gradient Descent (SGD) Classification, Binary Support Vector Classification (SVC), Gaussian Naive Bayes, Soft Voting Ensemble, Hard Voting Ensemble, Logistic Regression, Recurrent Neural Network (RNN), Decision Tree, Autoregressive-Moving Average (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, measured by accuracy rate, was picked and evaluated on the full dataset using daily split increments. Exemplary screenshots for the Logistic Regression model can be found in the appendix. The implementation of a time series compatible search can be seen in figure xy. An analysis of the hyperparameter search is shown in figure xy and figure xy visualizes the implementation of the final model run.

For a performance comparison of the applied models see figure “modelComparision.png”. Plot in text (high relevance)?

The best performing model was XY with an accuracy of XY and ROC-score of XY. The confusion matrix is displayed through figure xy. Analysis of matrix. The plot of the ROC curve is depicted by figure xy.

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