Johannes Weyers: Scraping day-ahead and intraday data, descriptive analysis

Julian Cornea: Scraping Weather Data, predictive analysis

Kim Ferres: Scraping market participants and holidays, predictive analysis

Rufin KorbmacherEin Bild, das Elektronik enthält.

Automatisch generierte Beschreibung: Interpolation, descriptive analysis

Team Assignment AAA Group 2

Advanced Analysis and Applications

Johannes Weyers, Julian Cornea, Kim Ferres, Rufin Korbmacher | 30.01.2019

List of content

[1. Executive Summary 1](#_Toc536549084)

[2. Problem Description 1](#_Toc536549085)

[3. Data Description: 2](#_Toc536549086)

[4. Data Preparation 2](#_Toc536549087)

[5. Data Analysis 3](#_Toc536549088)

[6. Conclusions and Operational Recommendations 5](#_Toc536549089)

[**Appendix** 6](#_Toc536549090)

# Executive Summary

The overall business objective of this project is to exploit differences in the two major electricity markets: The day-ahead and the intraday market. Both are available online and provide platforms to exchange electricity for specific timeslots. For the day-ahead market, offers and bids can be placed up to 12 pm and are than cleared in the following fifteen minutes, after which the respective prices are also published. The exchanged electricity can be purchased as Mwh for a specific hour and is traded one day in advance. The intraday market in contradiction to the day-ahead market provides the ability of quarter-hourly trading timeslots, which can be placed up to fifteen minutes before delivery. From 15 pm the next 24 hours are open for trading. So both intraday and day-ahead market make prices available for the same timeslots, but most likely are not equal. Sometimes the day-ahead and sometimes the intraday market has economic advantages. Hence a measurement for this price differences is introduced: The price premium. Patterns that can be found in the derived price premium can generate economic value, if it is utilized by only buying or selling at advantageous markets or advantageous timeslots.

In order to identify meaningful patterns, the analysis is divided into the examination of the data in terms of time related and non-time related patterns. In general, over 70% of all price premiums are positive, which induces that the intraday market seems to be economically preferable. Therefore, the occurrences of positive and negative price premiums are counted for each hour, weekday and even week. Striking timeslots are between 10 am and 13 pm, since the price premium is high and therefore it is advantageous to buy on intraday market on these slots. Also noticeable is the increase of the positive price premium in the end of the year. Here it also recommends buying electricity on intraday market, since the risk that the day-ahead market price is more expensive rises during October, November and December.

The non-time pattern analysis resulted in recommendations that partly support the findings of the time related pattern analysis. For example, the more freezing days occurred during the last 30 days the more likely the intraday market price is cheaper than the day-ahead market. In contradiction to that the snow volume implies the only rule that recommends a purchase on day-ahead market, due to a negative price premium. If the snow volume is 300 centimeters the probability of a negative price premium is above 50%.

Since a set of manually derived rules is difficult to optimize, deploy and maintain in a next step the weather and price data is used in order to derive predictive models that are able to predict whether the price premium is positive or negative. Therefore, seven predictive models are built and the best model, a Recurrent Neural Network, is able to predict with an accuracy of over 80%. It is expected that the performance of the model can be improved by using recent news reports and analyze them in order to detect potential catastrophes which can have impact on the prices of the markets.

# Problem Description

The essential objective of this investigation is to adopt the perspective of an electricity trading company and to develop data-based recommendations to support its profit generation by electricity trading. The business goal in this case is to maximize the turnover of the total amount of electricity, which is offered on sequential markets and is deliverable by the electricity provider for a period of time. This shall be achieved by following a smart approach in order to split the sales between the sequential short-term markets, here the day-ahead and intraday auctions of the German market, more effectively. Therefore, arbitrage opportunities could be exploited. Such an approach needs to seek for systematic price differences, fluctuations and deviations in sequential markets to be capable of evaluating which trading possibility is advantageous for the provider concerning a certain time slot. In concrete terms, an electricity provider can generate additional profits by offering more electricity on that market, where conditions are expected to be more lucrative. Hence an electricity trader can significantly benefit from analysis and predictions of price differences, fluctuations and volatilities. This study especially addresses the challenges created by parallel participation in several markets, which has become essential mostly due to the incalculability of renewable energy generation. As a growth for the ratio of renewable energy sources over the next couple years is planned, it can also be assumed that short-time trading of electricity will be of increasing importance in the soon future and therefore will gain relevance for providers. Identifying a systematism in the price differences of sequential markets and deriving a recommendation for pricing auction offers would be considered a success of this study. This would increase the effectiveness of dealing with the electricity market’s price volatility and decrease current opportunity costs.

To achieve the overall business goal, the exploitation of systematic price differences, it is necessary to identify parameters which influence the price differences between the day-ahead and intraday auctions. These features can be used to build a model that is able to explain the relationship between the price for both auctions at a given time. Another part of the data mining goal is the development of a predictive model, which should be capable of forecasting the direction of price differences. Both models solve a supervised task, due to the availability of price data. The first model is retrospective, whereas the second is predictive. The main outcome variables of interest are the parameters which influence the prices of both auctions, including the quantification of each variable’s influence and the resulting price differences.

# Data Description:

The dataset to be examined includes 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 intervals. As a whole the dataset encompasses 731 days, two years, which results in 17544 data instances for the day-ahead data and 70176 instances for the intraday auction data during the years 2016 and 2017. The very first step was to scrape the needed market data from the website <https://www.epexspot.com/de/>. After the successful collection of day-ahead and intraday auction price and volume information, the hourly data was converted into 15-minute increments and attached to the data file containing the intraday auction data. Since there is no missing data, no respective preparations were required. A timestamp is used as the unique identifier of each row in the merged data file. Figure 3-1 shows an exemplary subset of our core data set. You can find several plots of the scraped auction data in the appendix. Figure 3-2 shows the consumption on the intraday market throughout the years 2016 and 2017. Figure 3-3 shows the price fluctuations in the same auction. Figure 3-4 shows the consumption on the day-ahead auction. Figure 3-5 shows the price fluctuations on that auction. Figure 3-6 shows the price fluctuations of both auctions, while figure 3-7 compares the consumption.

# Data Preparation

In this section it is described which steps the data preparation included. For details concerning the scraping see section 4-0 the appendix. The German holidays were scraped from the page <https://feiertage-api.de>. Two other features specify the days distance of the observed day from the last and the next holiday. Another added feature is the number of participants in the market, the electricity providers. The data for the participants was scraped from <https://www.acer-remit.eu/portal/register-download?fileType=XML&euregId=> on a monthly level. Various weather data was scraped from the website <https://www.weatheronline.de/Deutschland/>. Next, the data hat to be interpolated. This addresses the circumstance that all data had to be adjusted to the same 15-minute granularity. For this a series with a 15-minute granularity was created, which was later connected as a dataframe with the actual data by a left merge via the respective datetime features to a dataframe. The result of this merge was a dataframe with a 15-minute granularity, but the features were only filled for every full hour. This was solved by using a forward fill. After this, all the still unrelated files of scraped data could be merged into a large file via the datetime column. Also, the price premium, which is calculated by the difference of the day-ahead price and the intraday price at time t, was added. The following formula represents this calculation:

To fix the notation, let fit denote the electricity day-ahead price observed on day t for delivery during hour i of day t + 1, and let Si,t+1 denote the intraday price for hour i of day t + 1. Figure 4-8 shows the price premium over time. 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 (the country wide average was usually influenced by the location “Zugspitze” by at least 80%). 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 was that the named data did not add valuable information in addition to the average of the daily mean temperature per month and keeping it would have caused unnecessary multicollinearity. This can be seen when looking at Figure 4-7. The StandardScaler was used to scale our data set because, for this data set, it was estimated to offer the best balance between avoiding high sensitivity to outliers and keeping the range of the scaled values relatively dense. This balance is important to optimize model performance. The previously described data set is used for the descriptive part of the task. In order to prepare the dataset for the predictive approach, the data had to be adjusted furthermore. The datetime column was replaced by new columns containing one hot encoded months, weekdays and hours of each data instance. Additionally, the original price premium was replaced by a categorical value. This binary value indicates whether the premium is positive or negative (1 means that it’s positive, v.v.). To add a history of past electricity consumption and price to the records, the most relevant historic market data for both auction types were calculated for each time slot (see figure 4-9 for an example). The most relevant data for a time slot includes the auction data at the last date, which is available at the prediction time, that matches the time of the slot to be predicted. Furthermore, it comprises the slope value of price and consumption for the week before the target time slot. This additional historic approach requires the exclusion of the first eight days for the predictive dataset, since the former auction data for these entries is not included in the dataset.

# Data Analysis

* 1. Descriptive

In the descriptive part of our analysis we will now try to discover hidden systematologies and use them to suggest how a good trader should behave. The analysis is divided into two parts, the examination of time patterns and secondly non time patterns.

Starting by uncover monthly systematologies in Figure 5-1-1, it can be stated out that the price premium drops is especially increased from September to March. Thereafter, it remains at an almost constant level throughout the summer. In the Figure 5-1-2 for the standard deviation, however, it can be seen that in the months from September to march, the standard deviation is also particularly high, until it remains at a constantly low level over the summer. Therefore, it can only be suggested that from September to March it is particularly worthwhile to buy on the intraday market, but there may also be a greater risk of losses due to a high standard deviation. Finally, in Figure 5-1-4 for the hourly average prices it can be seen that price premium is particularly high from nine to thirteen o'clock and from eighteen to one o'clock at night. Nevertheless, the standard deviation in Figure 5-1-5 from seventeen to twenty o'clock is exceptionally high, which leads to the conclusion that it is particularly worthwhile to buy from nine to thirteen o'clock and from twenty to one o'clock at night on the intraday market. The Figure 5-1-6 states out that there is a high autocorrelation between a timeslot t and the timeslot t-96, which is the timeslot 24 hours before t. Another peak can be observed at 672, which describes a high autocorrelation between t and the timeslot one week earlier.

In general, a tendency to positive price premium can be found in the data, which can also be seen on the distribution on the provided Figure 5-1-7. To get a good overview which exogenous variable have unequally distributions of positive and negative price premiums, several violin plots are created and qualitatively analysed from Figure 5-1-8 to Figure 5-1-13. Some of the variables show some obvious discrepancy between the distribution of positive and negative price premium, especially the weather-related attributes like “daily sunny hours” in Figure 5-1-11 or “daily Snow Volume” in Figure 5-1-13. Since a visual analysis alone is not that precise, another approach is used by dividing the occurrences of a positive premium by the number of all price premiums, in order to quantify the share of positive price premiums against negative ones in terms of the different variables. This approach can be seen from Figure 1-5-13 to Figure 1-5-18. It can be observed from the visualizations that only for one value, when the daily snow volume is around 300 cm in Figure 5-1-18, it can be recommended to buy on day ahead market instead of intraday market. It should be considered that this analysis does not take into account any balancing penalties or limited electricity market volumes. For such restrictions it can be possible that the expected price premium is more balanced. Apart from the previous analysis, a tree classification model is used to uncover the most informative features in terms of the feature price premium category. The major advantage of a tree model is its interpretability. In order to find the best model parameter with the least number of leaves that fits the data well, a grid search is implemented for the model. In addition, a time series compatible cross validation is implemented to verify the accuracy, which is used to measure how well the model fits the data. The Figure 5-1-19 shows that the model uses only three variables to identify whether the “price premium category” is positive or negative: “dailySunnyHoursAvg”, “numberRainyDays” and “participants”. In the figure 5-1-20 the confusion matrix and the accuracy of the tree-classifier can be viewed. It can be recognized that the accuracy of the model is 78.89%.

5.2 Predictive

The predictive analysis is, in contrast to the descriptive part, not focused on interpretability and goodness of fit. The only dimension of interest is the predictive power, which is represented through accuracy throughout this project. In order to simplify the interpretation of the performance through the accuracy a baseline is calculated first. Therefore, a simple “ZeroR” classifier is introduced, which predicts only the class which has the most observations in the dataset. This technique simulates random predictions. With regard to our dataset the baseline is: 64.25%=44594/69408. Above mentioned the tree classifier can also be used for out-of-sample predictions and with reference to its accuracy 78.89% it can be stated out that it is around 14% better than the defined baseline. 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. Both ensembles consist of the following classifiers: SGD, Decision Tree, Logistic Regression and Gaussian Naive Bayes.

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. The only exception to that is the Soft Voting Classifier which was evaluated on increments containing about 9 days each, due to a relatively high computing cost. 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 5-2-1. An analysis of the hyperparameter search is shown in Figure 5-2-2 and Figure 5-2-3 visualizes the implementation of the final model run. For a performance comparison of the applied models see Figure 5-2-4. The best performing model was the Recurrent Neural Network (RNN) with an accuracy of 81.4% and ROC-score of 75%. The confusion matrix is displayed through Figure 5-2-5. The plot of the ROC curve is depicted by Figure 5-2-6. Here the neural net receives an input similar to a normal neural net. In addition, as with a neural network, various complex calculations are carried out in the network and we finally get a prediction for a certain input. First the weights at the knots must be trained. What’s peculiar about a Recurrent Neural Network is that the previously calculated values still affect future values. This is especially useful in this case, because values for a certain point in time are searched, which are affected by the values of past time slots.

# Conclusions and Operational Recommendations

To tackle the task of predicting which auction will have a higher energy price for a given time slot, a Recurrent Neural Network turned out to be the best performing model. The main downside of using it however, is the lack of explanatory power. While the model itself offers the best predictive performance, it’s output can’t be used to gain a deeper understanding of the market. To use it in a productive environment several things have to be considered. First of all, the weather data fed to the model will consist of predicted weather data. However, once real weather data is available for a given time slot, it might be useful to add it as historical data, similar to the historical auction data approach used in our solution. Secondly, the historical auction data approach used by us can and should be evaluated on a regular basis to feed the model the most relevant historical data. To do that it would be useful to create regular explanatory analyses of the auctions so that relevant factors can be kept current. Finally, once put into production the model’s performance has to be carefully monitored and the model needs to be backed up regularly to prevent, or minimize, model degradation. However, if the predictive task is altered to predicting how much energy should be offered, or bought, at which auction, and at what price, it would make sense to test out a reinforcement learning model, as the task’s type would change from classification to a multi-layered regression. The reward type used for it would be the net gains and losses from its predictions.

# **Appendix**

Fig. 3-1



Fig. 3-2

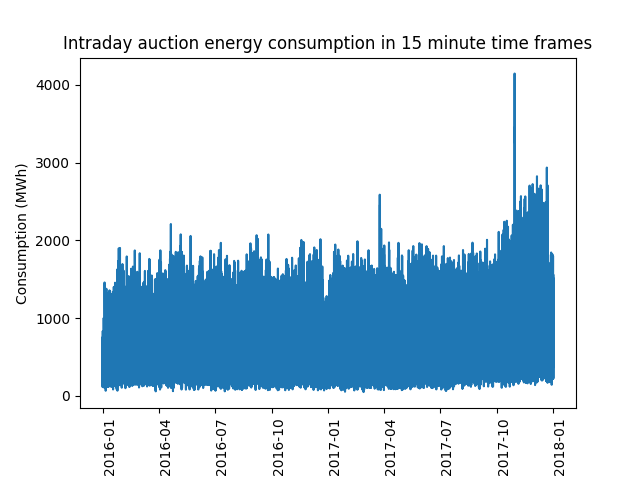
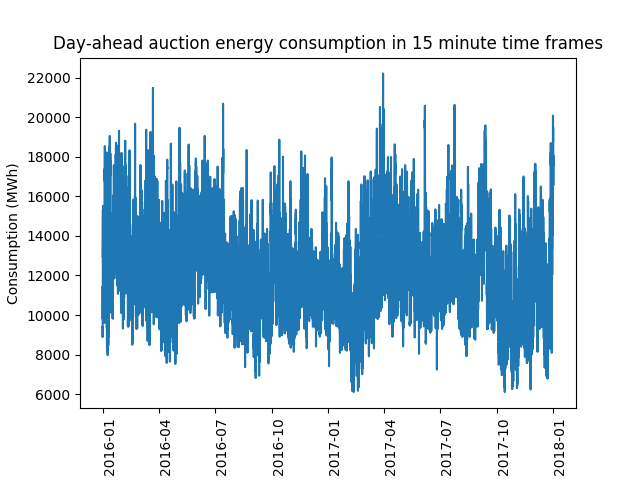


Fig. 3-3



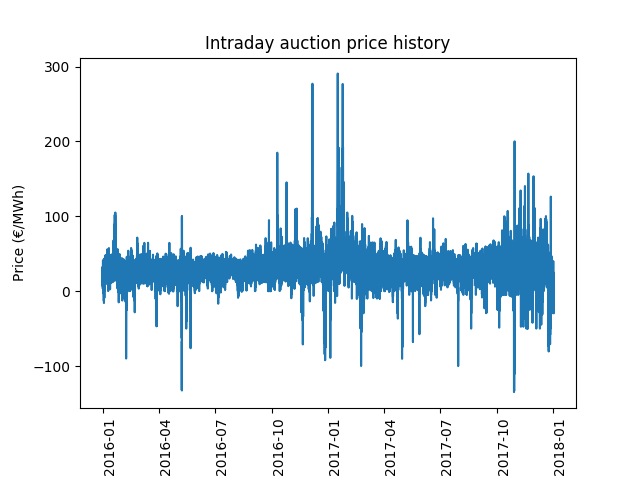
Fig. 3-4

Fig. 3-5

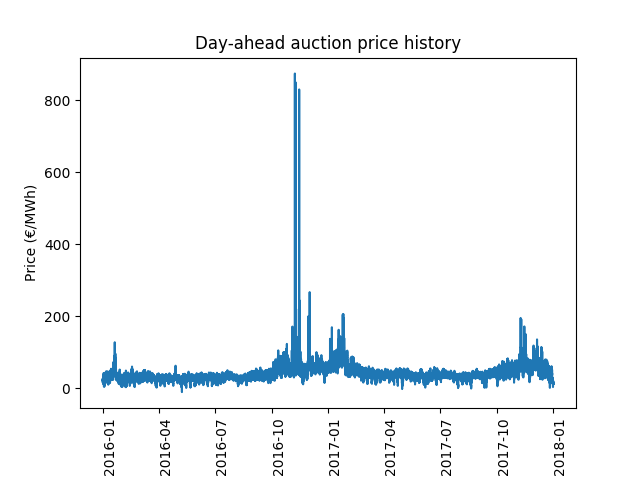
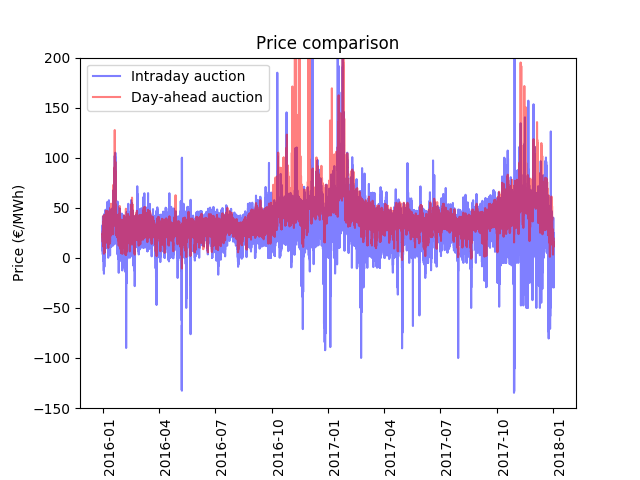
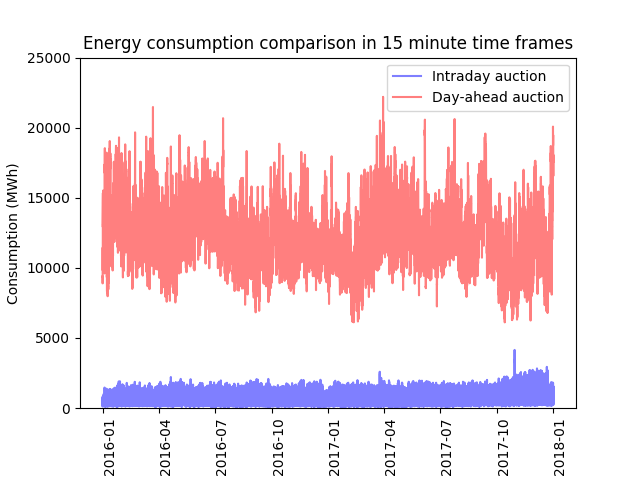


Fig. 3-6

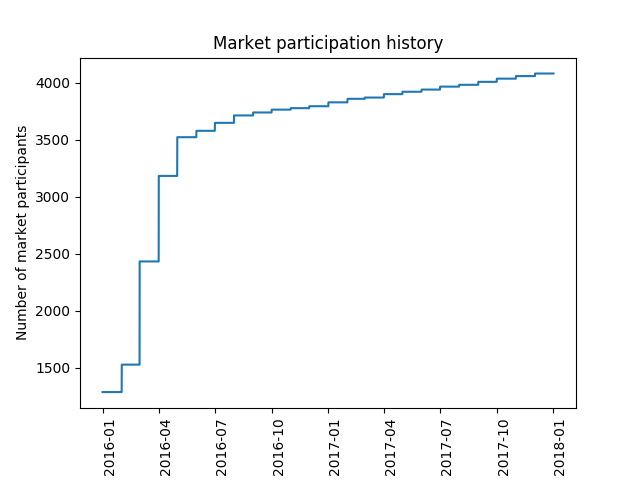


Fig. 3-7

Section 4-0

In addition to the scraped auction data, some feature values needed to be added. As one of these relevant features the German holidays were scraped from the page <https://feiertage-api.de>. This data later allows conclusions to be drawn as to whether electricity prices and consumption are dependent on public holidays. In our dataset the holidays are reflected by a boolean value, which indicates whether a day was a holiday or not. Furthermore, two other features specify the days distance of the observed day from the last and the next holiday. Another added feature is the number of participants in the market, the electricity providers. The data for the participants was scraped from <https://www.acer-remit.eu/portal/register-download?fileType=XML&euregId=> on a monthly level. A plot that shows the number of participants over time can be found in Figure 4-1. Last but not least, various weather data was scraped, as the weather also has an enormous influence on the electricity price of the market, especially concerning renewable electricity generation. The monthly data was scraped from the website <https://www.weatheronline.de/Deutschland/> for all available German locations and then averaged for each month. From this weather data the following features were added to the dataset: Daily average temperature in Celsius, daily maximum temperature in Celsius, daily minimum temperature in Celsius, monthly number of freezing days, monthly number of icy days, monthly rain volume in mm, monthly number of rainy days, the average amount of daily sunshine hours per month, average monthly wind speed in km/h, monthly number of snow days and the average amount of daily snowfall per month in cm. Since not all listed locations had a monthly data availability of 100%, the country wide average values were adjusted to account for the data availability of each location. This was accomplished by dividing the weighted average of a feature (which is the sum of each location’s value multiplied by the data availability of that location) by the sum of data availability across all locations. Several plots visualizing the weather data can be found in the appendix, ranging from figure 4-2 to figure 4-7.

Fig. 4-1



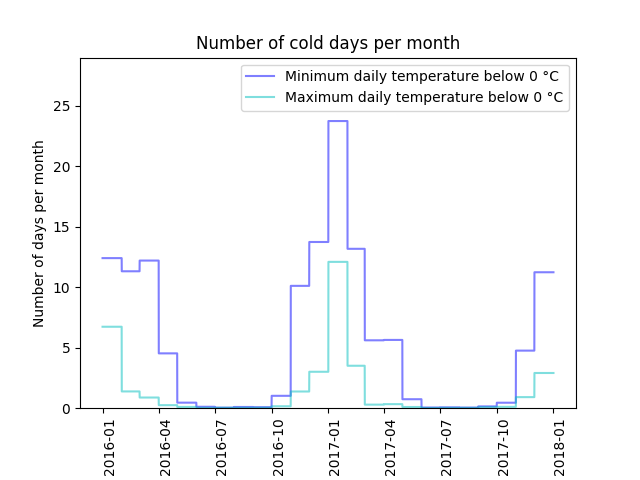
Fig. 4-2

Fig. 4-3

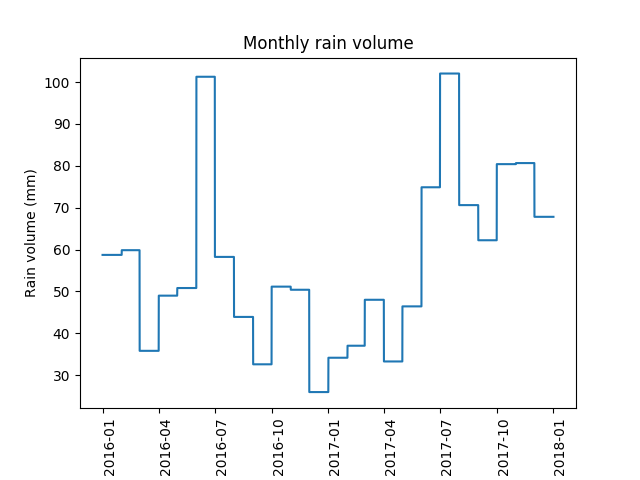


Fig. 4-4

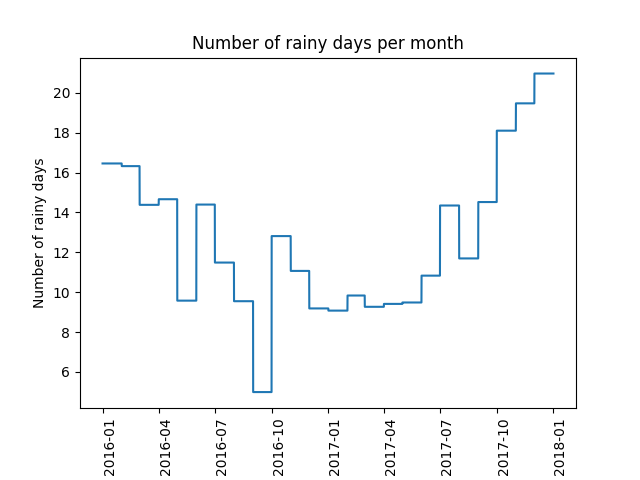


Fig. 4-5

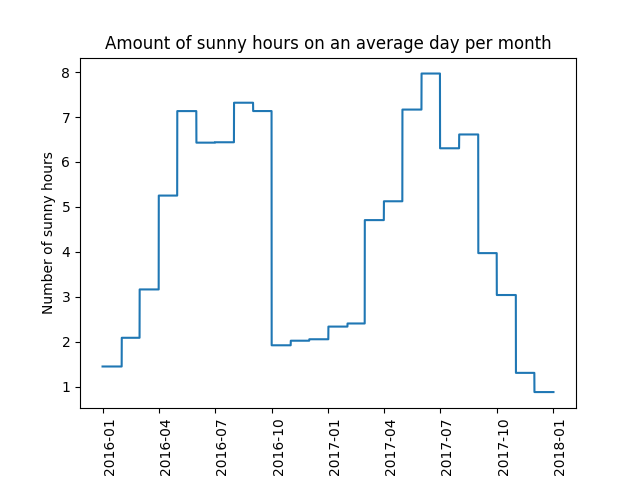


Fig. 4-6

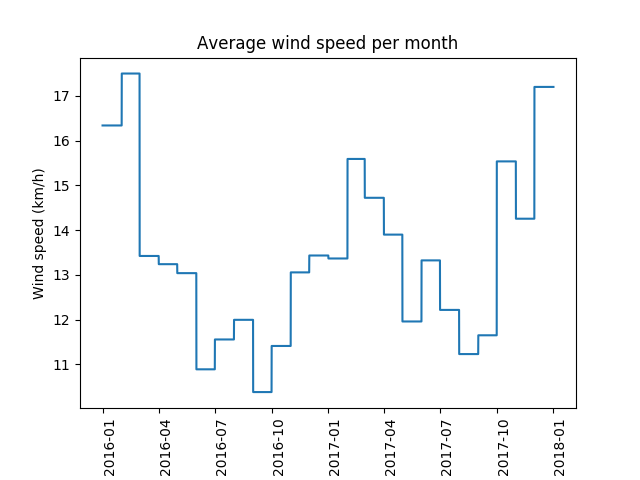


Fig. 4-7

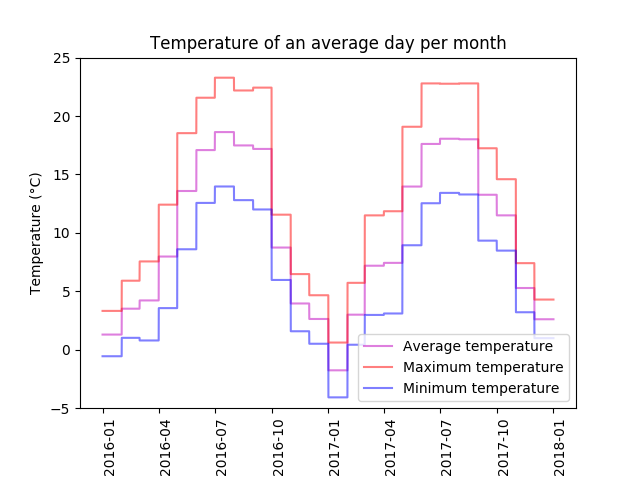


Fig. 4-8

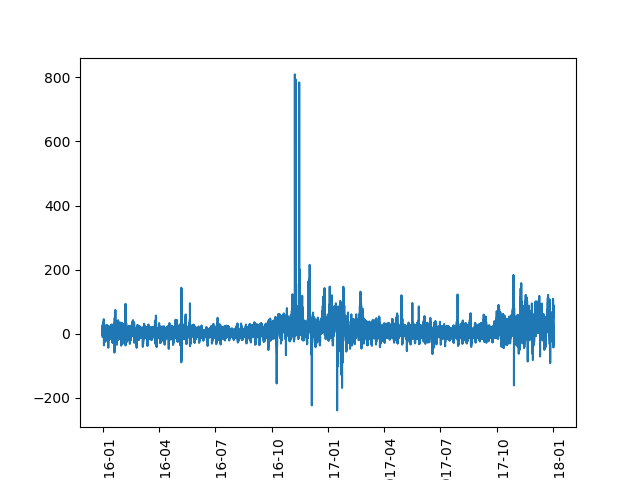
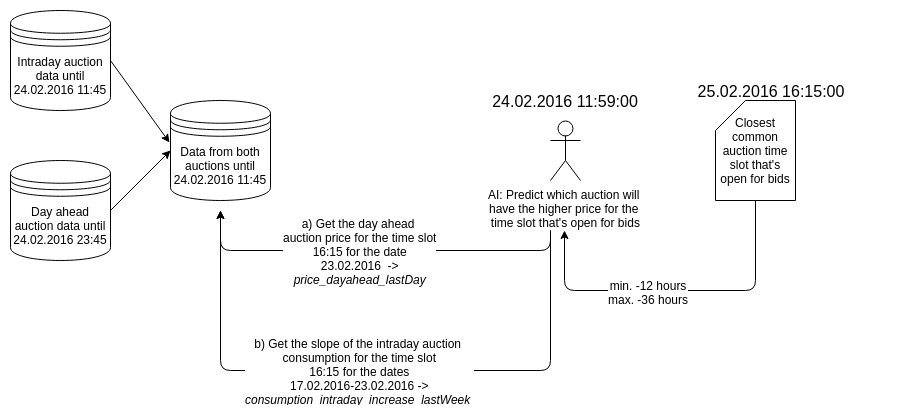


Fig. 4-9



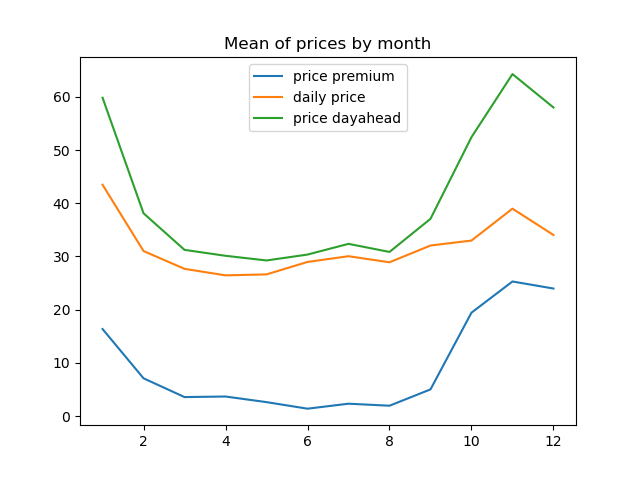
Fig. 5-1-1

Fig. 5-1-2

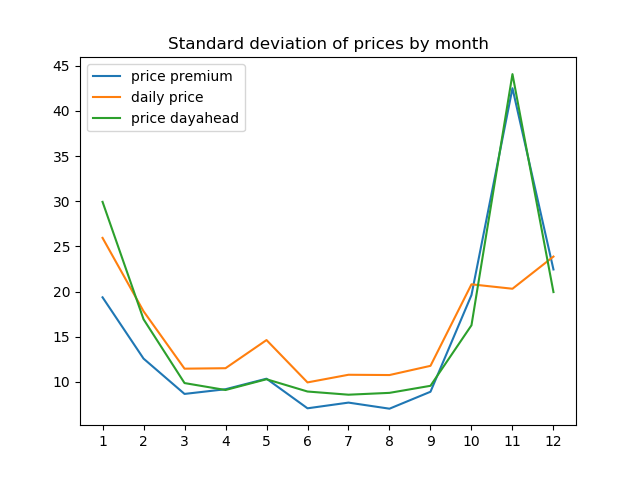


Fig. 5-1-3

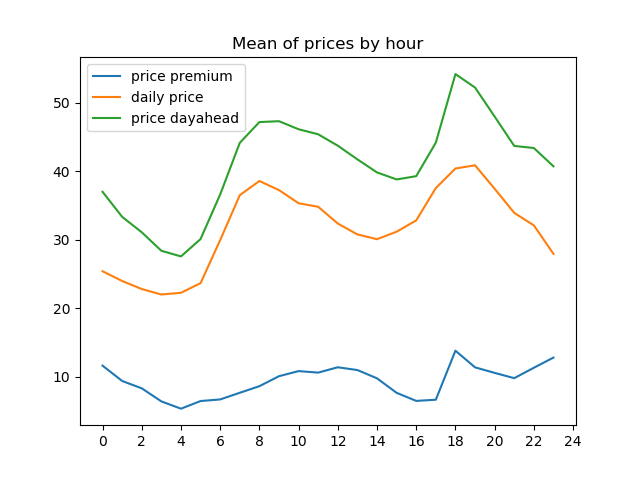


Fig. 5-1-4

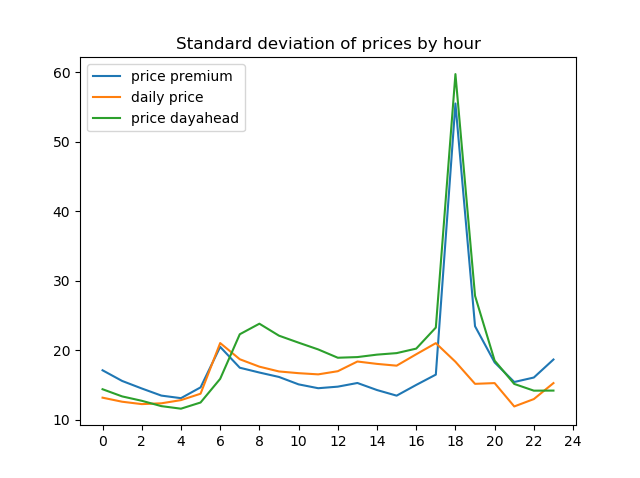
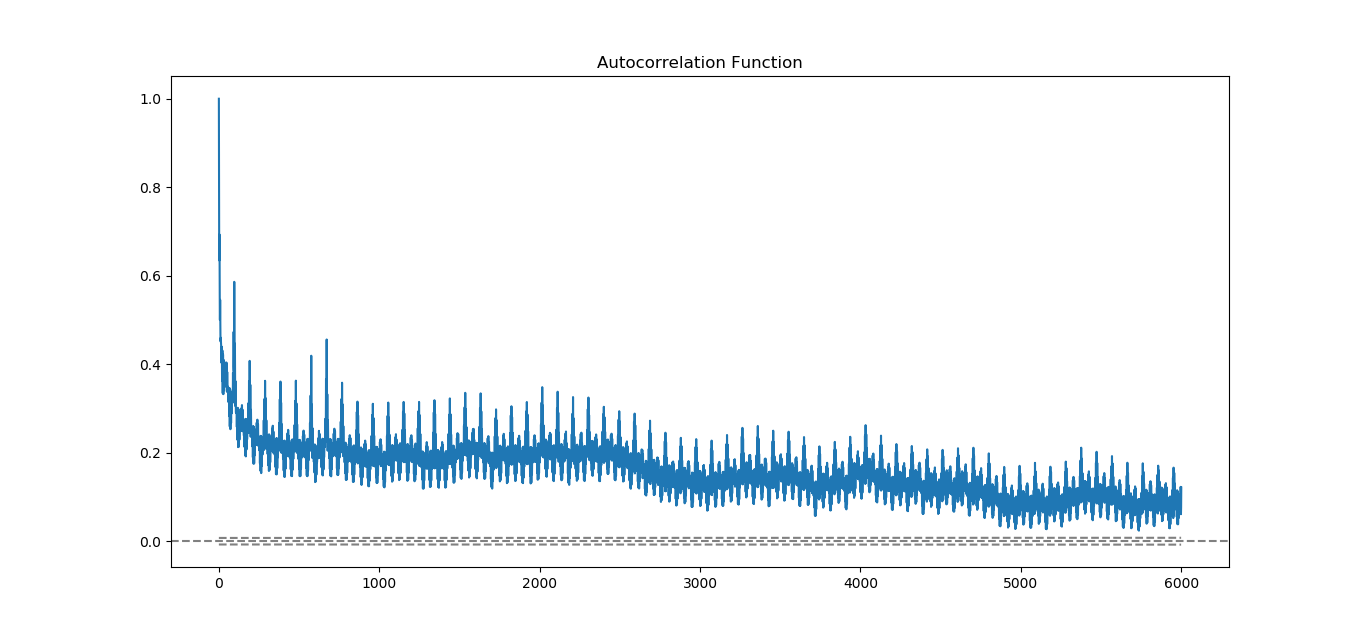


Fig. 5-1-7



daily

weekly

Fig. 5-1-8

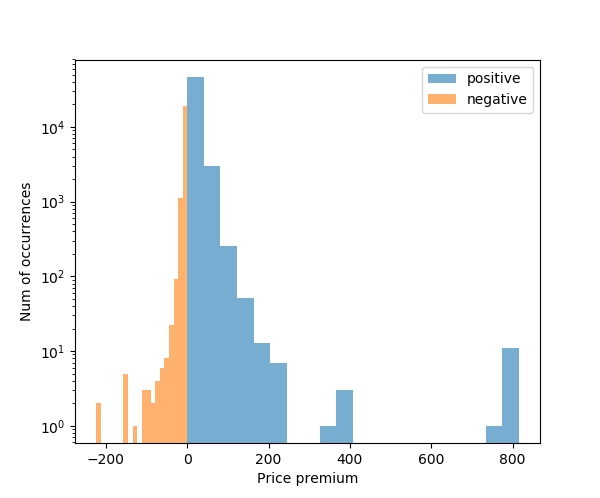


Fig. 5-1-9

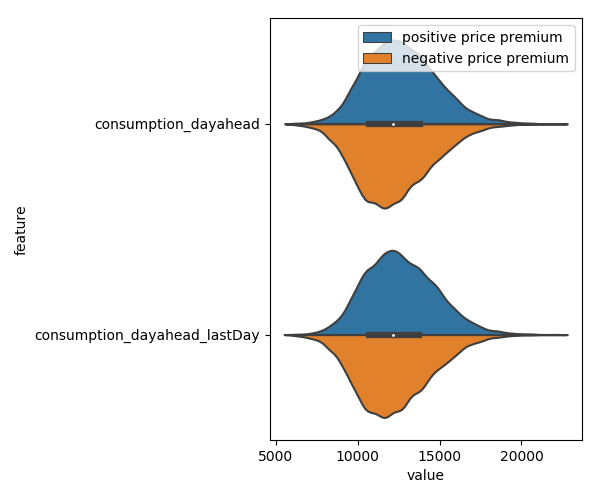


Fig. 5-1-10

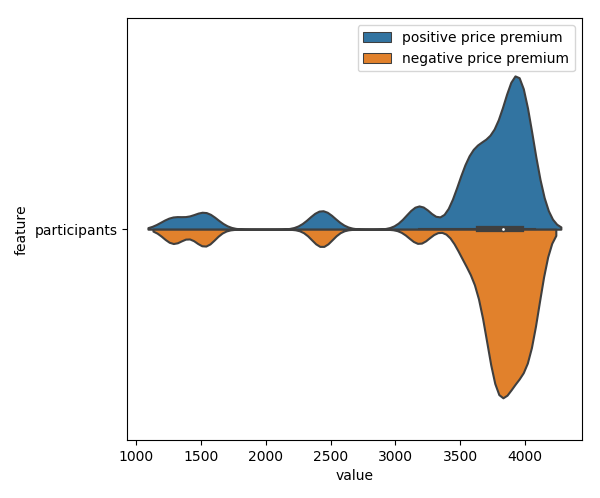


Fig. 5-1-11

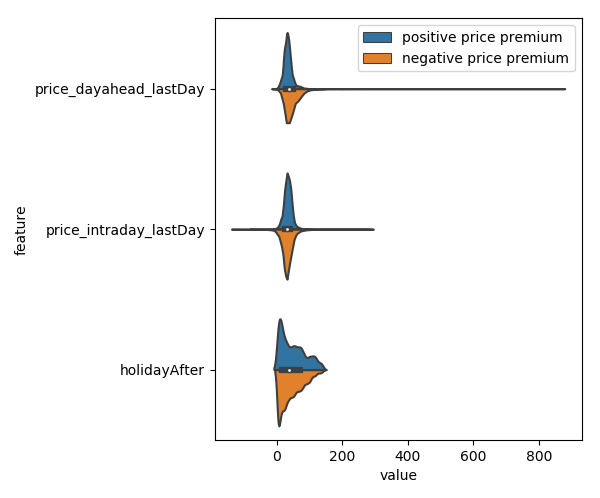


Fig. 5-1-12

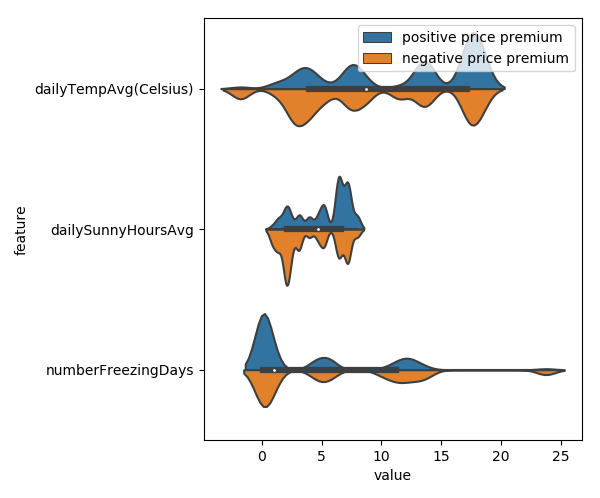


Fig. 5-1-13

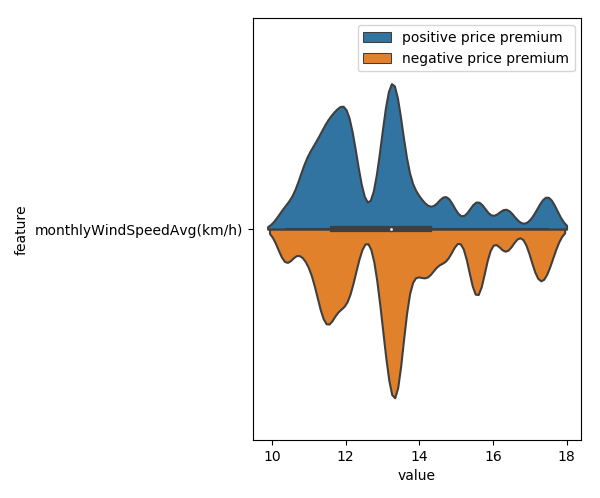


Fig. 5-1-14

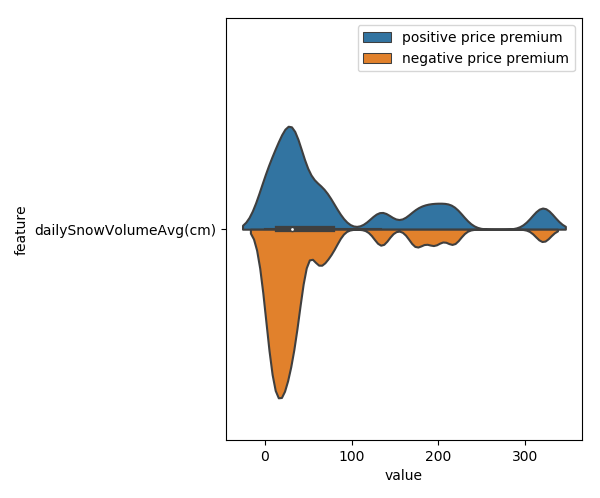


Fig. 5-1-15

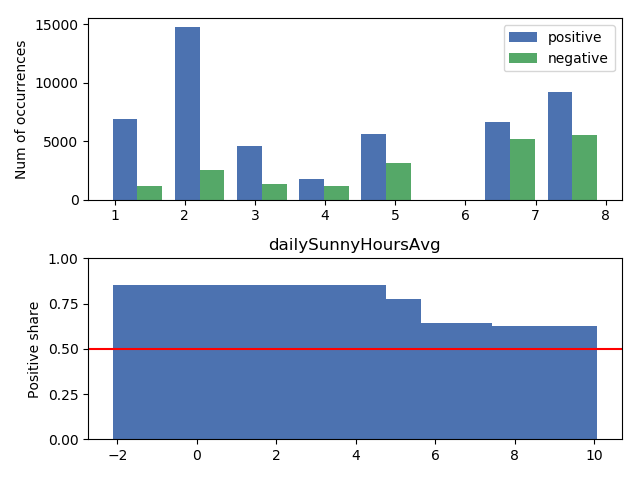


Fig. 5-1-16

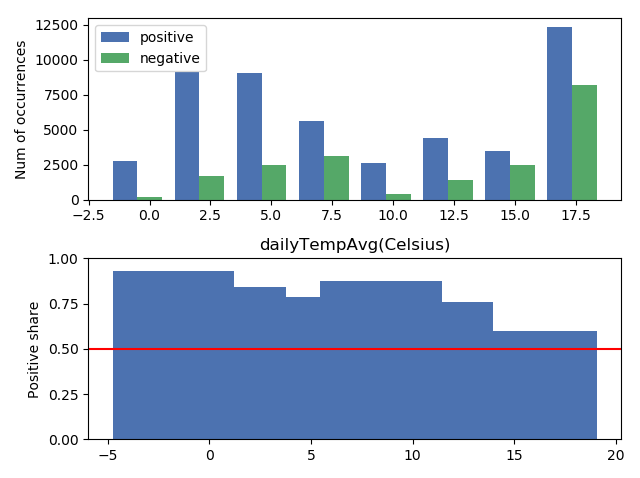


Fig. 5-1-17

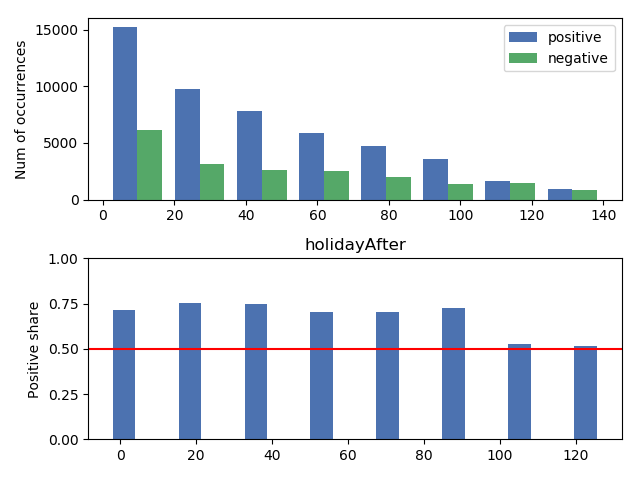


Fig. 5-1-18

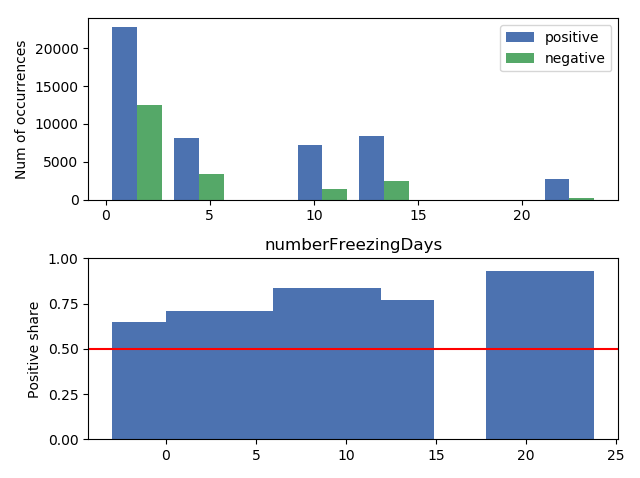


Fig. 5-1-19

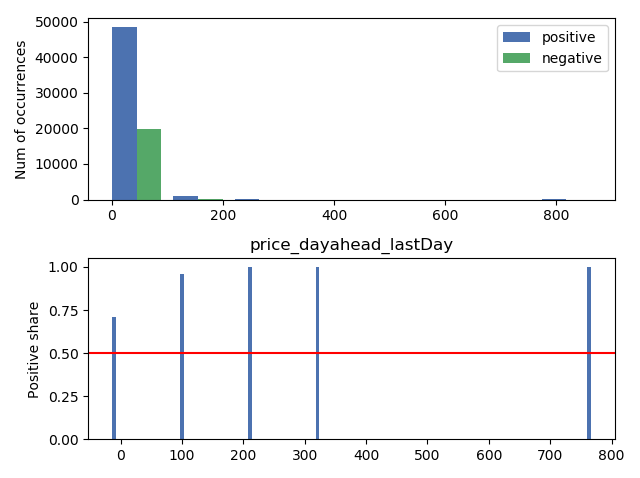


Fig. 5-1-20

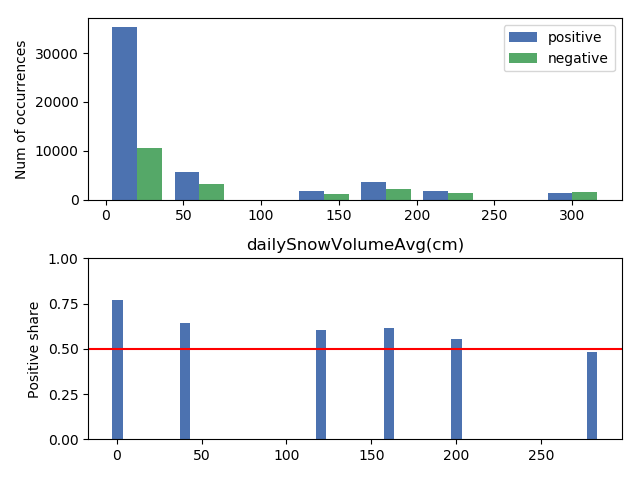


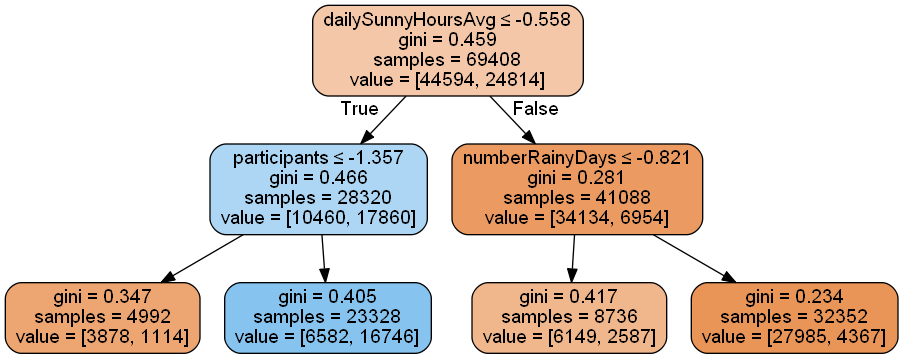
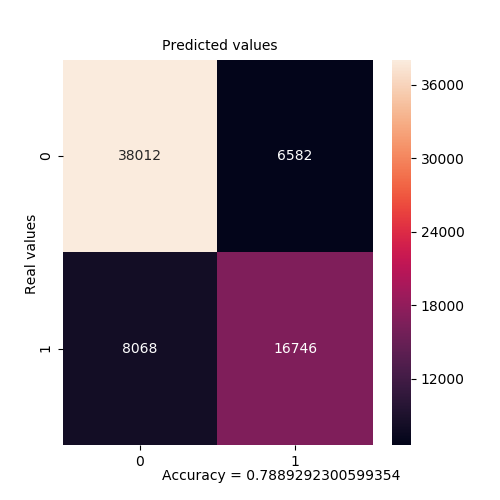
Fig. 5-1-21

Fig. 5-1-22



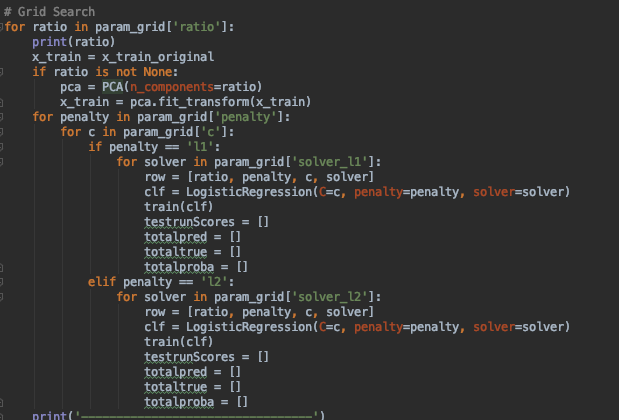
Fig. 5-2-1

Fig. 5-2-2

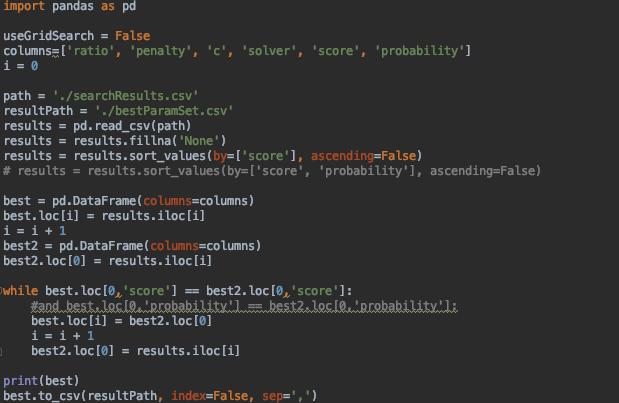


Fig. 5-2-3

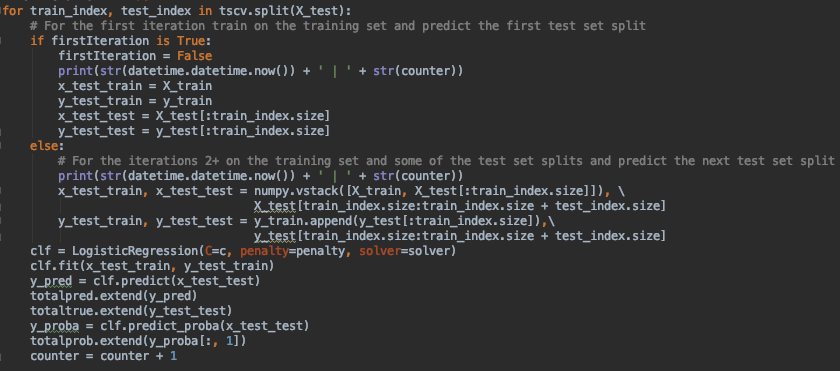


Fig. 5-2-4

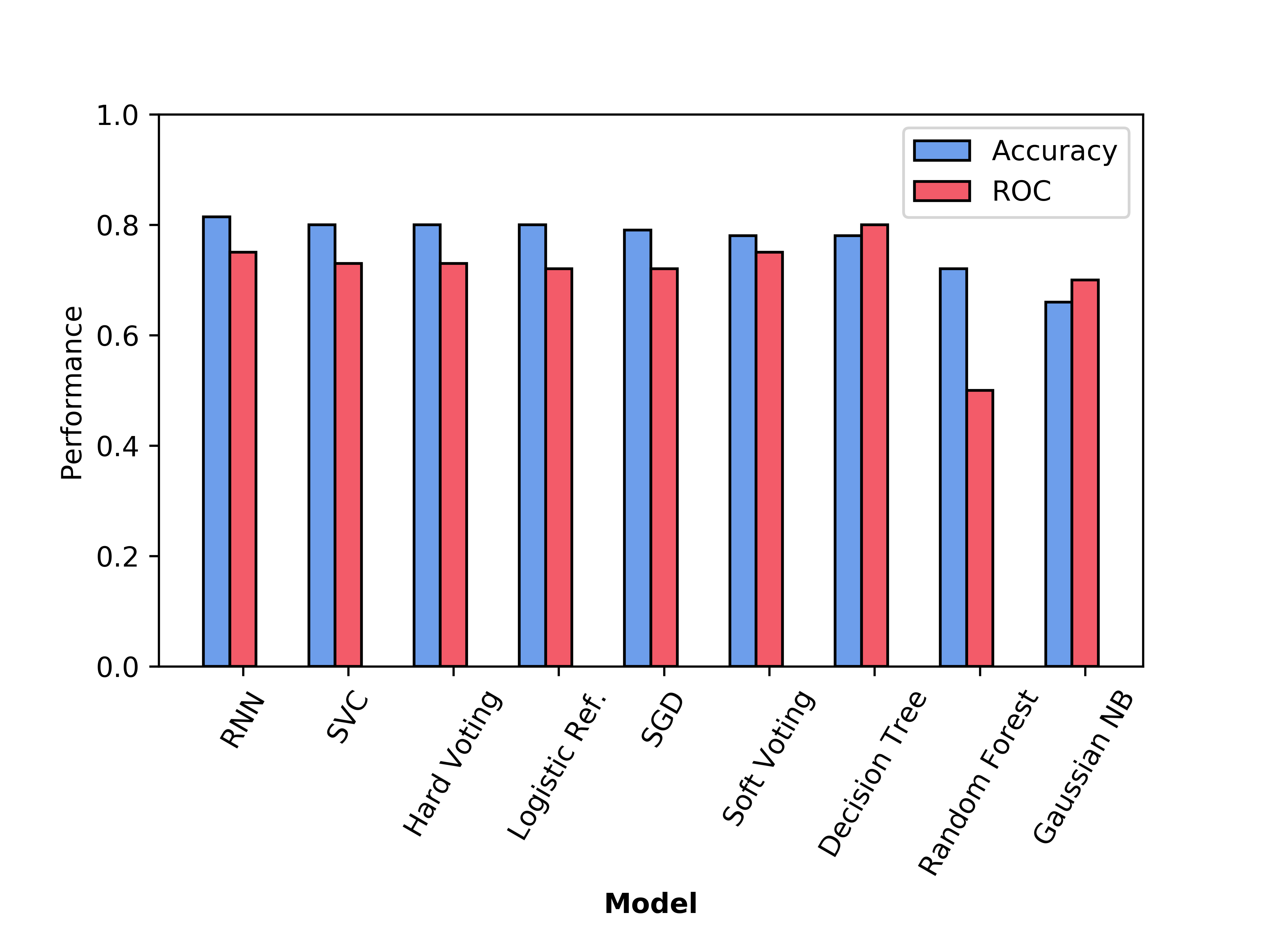


Fig. 5-2-5

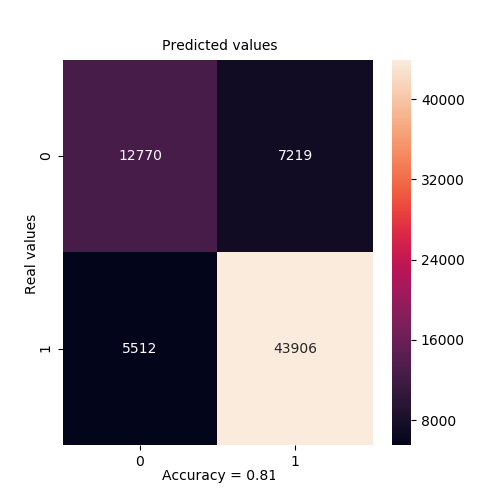


Fig. 5-2-6