Prediction of day ahead prices

I. Gutierrez, E. Fredrixon, F. Holmberg, J. Watson

EC Utbildning

Projektkurs

202410

Abstract

In the current report day ahead prices of electricity are modelled using linear, CART, Random Forest and Ranger models. Weather, inflation and price data was processed in an extraction, transformation and loading (ETL) pipeline in python and modelled using R. It was possible to predict prices with relatively low RMSE (30-40 %) using a Ranger model. This result could, arguably, improve using more complete data from the entire European energy market. Still, it demonstrates how little means in data can be used to get an idea of the daily day ahead price.

Table of content

1 Introduction

Electricity was discovered in 1752 and in today's modern world people and industries rely heavily on electricity to function. The price of electricity is essentially governed by demand and supply. In Sweden, the demand is principally made up by households and industry who demand 54 % and 34 % of the total consumption, respectively [11]. Supply is various, where electricity is produced (by 2022) by nuclear fuel (29 %), wind (19 %), solar (1 %), thermal (9 %) and hydro (41 %) power.

In recent years inflation has increased globally, and money's worth has been decreasing. Most would agree that economies have been struggling. With that said, households and industries also get impacted by this change in weather conditions, as higher utility bills and increased costs of everyday necessities. Therefore, cutting costs and risks are sought after. One such post is electricity, and a model to predict prices would be useful.

The energy market is complex, making modelling a non-trivial matter. Still, it is possible to achieve *a result* with little effort with today's infrastructure of data (API and other open sources) and widespread availability of machine learning (ML). Variables that are easily accessible and may, at least, rationally describe prices are inflation and weather (i.e. temperature). Of course, there are other outside influences that can also affect electricity prices, such as world conflicts and rising global climate change. However, such data are difficult to quantify and incorporate into models.

The main purpose of this report is to model electricity prices (day ahead) using weather and inflation data. To achieve this, the following list of tasks are relevant,

- Create an extraction, transformation and loading pipeline for the data.
- Inspect and attempt to model the data using suitable models.
- Predict prices using the best model.

(Saunders, 2023) (Bergqvist, 2023)

2 Theory

2.1 Day ahead Prices

The day before electricity is sold there are closed auctions between producers and consumers. Consumers bid for the next day, balancing supply and demand for each hour of the day. With day ahead Prices you can ensure efficient electricity trading and grid stability.

2.2 Regression model

A regressions model is a method for analysing the relationship between a dependent variable and one or more independent variables. It also has the capability to predict, given the ability of new independent variables.

2.2.1 Random Forest Regressor

The Random Forest Regressor is an algorithm used for regression tasks which means that it is designed to predict continuous values such as price or temperature. By using multiple decision trees that are trained on different subsets of the data, each tree in the forest output a prediction. It is this "collective wisdom" that yield the final prediction of the Random Forest Regressor. (Breiman, 2001)

2.2.1.1 Ranger

Ranger is a package used in R studio that acts as a fast implementation of Random Forest and works particularly well if you have high dimensional data. (Wright, Wager, & Probst, 2023)

2.2.1.2 Law of Large Numbers

The Law of Large Numbers states that the more samples you have the truer sample mean is. (Grimmett & Stirzaker, 2001, ss. 325-331)

2.2.2 Linear Regression

Linear regression models a straight line that reduces the difference between the predicted value and the actual output values. It is common practise to use a linear model as a first step when exploring predictive modelling. (IBM, 2024)

2.3 CART

The CART algorithm works by recursively partitioning data into subsets based on feature values, aiming to create branches and nodes that result in the best possible prediction of the target variable. (Guild, 2021)

2.4 Machine Learning

By using data and algorithms, computer scientists try to enable AI to imitate the way humans learn,

"Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed". (Samuel, 1959)

The more training data and the more features that are useable for accurately describing the label, the more reliable the prediction.

2.4.1 Model Evaluation

Model validation is the process of assessing how well a machine learning model generalizes to unseen data by evaluating its performance on a separate test set or using techniques like cv It helps ensure that the model is not overfitting to the training data and can make accurate predictions in real-world scenarios.

2.4.1.1 K-fold Cross-Validation

When training a model, a common alternative is to use cv which splits a training set into k distinct subsets called folds. If it splits the training set into 5 folds it will train and evaluate the model 5 times choosing a different fold for evaluation every time and training on the remaining 4 folds. (rishu_mishra, 2021)

2.4.2 Model Performance Metrics

In every machine learning pipeline, there will be performance metrics (i.e. RMSE, R^2 and adjusted R^2). Performance metrics indicate if the model is making progress and quantify it.

2.4.2.1 RMSE (Root Mean Squared Error)

By measuring the average difference between the model's predicted values and the correct value, RMSE provides an indicator of how well the model fits the data, with smaller values indicating better predictions. (Frost, 2023)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Equation 1 RMSE

2.4.2.2 R²

R² measures the overall accuracy of a model and is measured between 0 and 1 where 1 indicates a model that perfectly predicts values in the target field. However, when adding more predictors R² cannot get a lower value. (IBM, 2024)

$$R^2 = 1\frac{SS_{res}}{SS_{tot}}$$

SS_{res} = Sum of Squared Residuals

SS_{tot} = Total Sum of Squares

$$SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$$

Equation 2 R²

2.4.2.3 Adjusted R²

Since R² cannot get lower values when adding more predictors it is common to use Adjusted R² which penalizes the addition of unnecessary predictors. (IBM, 2024)

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Equation 3 Adjusted R²

2.4.2.4 Variable Importance

Variable Importance is a measure used to rank individual predictor variables in a random forest model based on their contribution to predicting the outcome. A mean MSE is first calculated on the train set for a model. When variable "sunshine minutes" is permutated the model is re-examined on the same data set, resulting in a new mean MSE. The lowering in mean MSE is reported as a factor between 0 and 100. If sunshine minutes significantly lowers the MSE, it is reported as 100. If there is no lowering in mean MSE the variable has no importance, which is reported as 0. (Wybron, 2019)

3 Method

Data used for the current report was gathered (part 1) and used for modelling (part 2). Python and R were used for building a pipeline and to gain insights from data, respectively. The reader is encouraged to inspect the code, which can be found elsewhere.

3.1 Extraction, Transformation and Loading (Part 1)

Data was processed in an extract, transform and load (ETL) pipeline, executed by a main script. Data was downloaded through API or csv files. These sources are shown in Table 1. The steps in the main script can be seen in Figure 1.

Sources of data provided different time periods and formats. Different transformations were made (rounding, conversions and formatting) for each dataset. In the step of merging the three dataframes, only rows with complete data were allowed. Finally, the dataframe was stored in SQLite 3 database table.

Table 1: Data from various sources.

Data	**Source**	**Extraction**
Day ahead prices	transparency.entsoe.eu	API
Weather data	open-meteo.com	CSV
Inflation	ec.europa.eu	CSV

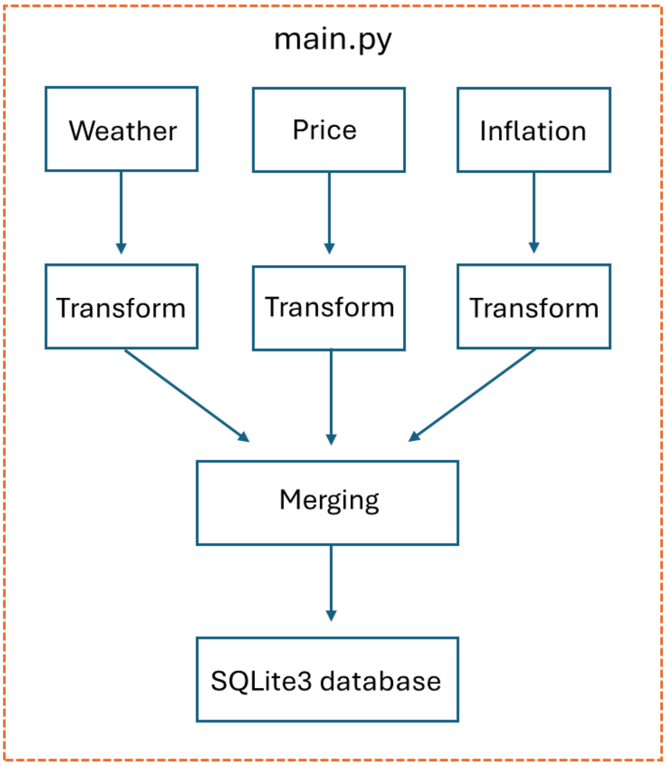


Figure 1: Layout of main.py use for data pipeline.

3.2 Data and modelling using R (Part 2)

A table of data was loaded from an SQLite 3 database and used for analysis. Before modelling the following steps were made:

- Evaluate missing data.
- Set random seed for reproducibility.
- Data was split into train (0.8) and validation (0.2) sets. A total of 2830 observations were used.
- Two observations not included in the data were collected separately, for the sake of having test observations.

Random Forest and Ranger models were evaluated using k-fold cross validation. The label variable was "öre/kWh" and the feature variables are listed below:

- Max temperature (°C)
- Min temperature (°C)
- Average temperature (°C)
- Precipitation (sum, mm)
- Rain (sum, mm)
- Snowfall (cm)
- Wind speed (max 10 km/h)
- Wind gust (max 10 km/h)
- Sunshine duration (min)
- Inflation (%)
- Month

A few design choices were made for the inflation and weather features. The level of inflation was only provided once per month. Hence all days per month were filled with the same monthly inflation level, for the sake of modeling. Weather data was collected in SE3 region from 8 larger cities (Göteborg, Karlstad, Jönköping, Stockholm, Gävle, Falun, Mora, Uppsala). The rationale for this sample selection was that weather in large cities would affect the consumption (hence the price) most. These 8 cities are also geographically spread out over SE3 and would also generally describe the weather in this area. Finally, each feature (weather) was extracted as a mean of these 8 observations. (Mosquera-López, 2024)

4 Results and discussion

4.1 Training set results

4.1.1 Exploratory data analysis

Data processed with the ETL pipeline was explored by plotting some aspects. This was done to better understand the data set variables and statistics (e.g. min, mean and max values). In Figure 1 the daily electricity and day ahead prices are shown. General notes are that these prices are related and data is not linear. Other plots where also inspected (time vs rain, sunshine, etc) and can be seen in Appendix. As expected, some variables are similar to periodical waves (i.e sunshine ours and temperature) through the year, while others were not. Curiously, inflation does fit the day ahead prices well. All in all, the exploration in data was proven useful to better understand results in modelling.

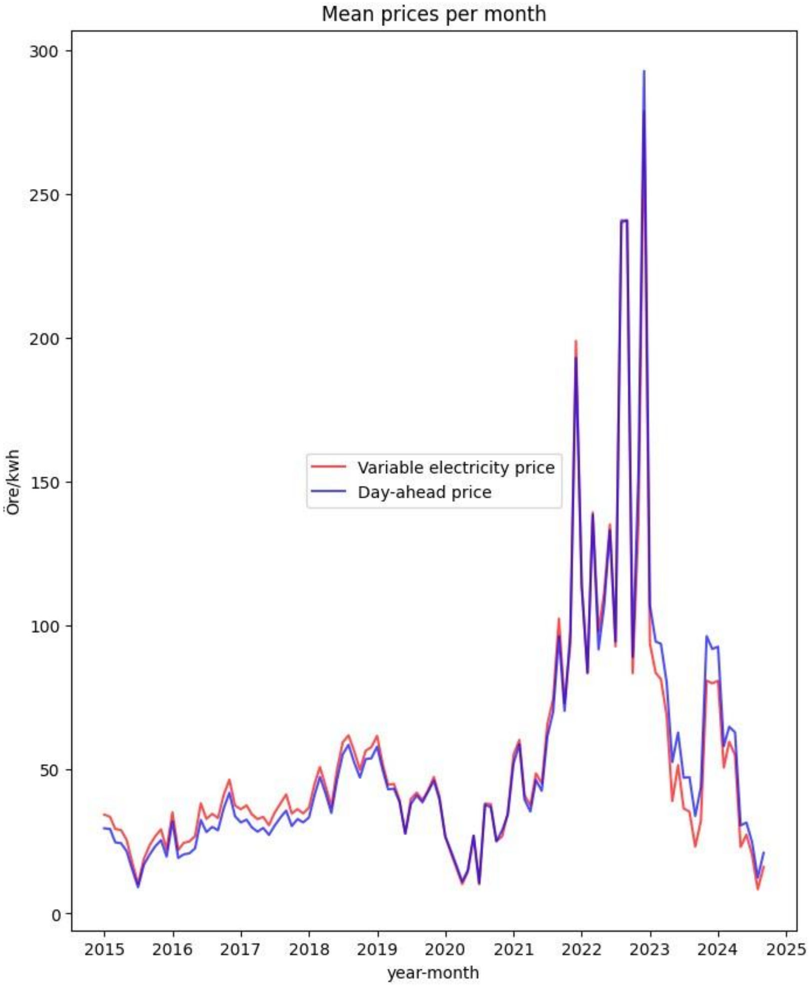


Figure 2: Mean electricity prices (actual and day ahead) from 2015 to 2024.

4.1.2 Model evaluation

Models were evaluated on the training set and the results can be seen in Table 2. The linear and did not perform well ($R^2_{adjusted}$ < 0.5). The linear model is useful when that data fits well around a linear (e.g. linear combination) curve. As seen in Figure 2, the data is not linear. This notion suggests the opposite, that a model (i.e. CART) suitable for non-linear data would yield better metrics. This was true for tested CART model. By increasing the number of trees (Random Forest and Ranger), the

RMSE (%) dropped from 60 % to 21-25 %. This improvement in RMSE (%) may be explained by that the CART model consists of 1 tree, whereas the Ranger consists of 500 trees. Probably, the entire ensemble could make better predictions due to the “wisdom of the crowd” effect. All in all, the results of tested models are what to be expected. This is also verified by the general increase of $R^2_{adjusted}$ which is gained from switching form linear model to CART, and subsequently Random Forest/Ranger model. (Yadav, 2024)

Table 2: Results metrics of models on training set.

Regression model	RMSE	RMSE (%)[1]	$R^2_{adjusted}$
Linear	47.5	80	0.42
CART	36.0	60	0.67
Random Forest	14.9	25	0.96
Ranger	12.2	21	0.97

[1]Calculated from mean of train data set of 59,50 öre/kWh. Min and max were -7,8 respectively 551,41 öre/kWh.

4.1.3 Variable importance

Variable importance was calculated for Random Forest and Ranger models, which are shown in Table. The results show that (1) similar metrics are obtained for both models and (2) inflation appears to be most important (100). Other weather data also affect (up to 22) the model prediction. The mean decrease in MSE is therefore most affected by inflation, hence inflation is the most important variable.

The similarity of results is coupled with that the Random Forest and Ranger are closely related. Overall, the Ranger model is an optimized version of the Random Forest model. Thus, it is to be expected that both models yield similar results. However, it is also interesting to notice that with these data dimensions it is indicated that it is difficult to optimize a random forest model to improve prediction capacity.

Table 3: Variable importance for Random Forest and Ranger models.

Variable	Random Forest	Ranger
Max temperature (°C)	18	18
Min temperature (°C)	10	11
Average temperature (°C)	14	15
Precipitation (sum, mm)	3	4
Rain (sum, mm)	40	40
Snowfall (cm)	0	0
Wind speed (max 10 km/h)	17	18
Wind gust (max 10 km/h)	21	22
Sunshine duration (min)	8	8
Inflation (%)	100	100
Month	10	9

4.2 Validation set results

Random Forest and Ranger model performed best on the train set and was used for predictions on the validation set. In Table it can be observed that both models performed equally, in terms of RMSE and $R^2_{adjusted}$. In agreement with previous results (see 4.1.3) it difficult to optimize Ranger to exceed the Random Forest performance. The Ranger model is notably faster to fit data, which is a preferable practical aspect of this model over the other. It also performed slightly better (i.e. lower RMSE) on the training set.

A general observation of the model's performances on the validation set is that RMSE increased and $R^2_{adjusted}$ decreased. A few reasonable suggestions to this observation are:

- Overfitting.
- Too few datapoints (training set).
- Unbalanced spread in data between training and validation set.

Here, design choices need to be made to devoid or reduce impacts of these three aspects. Random Forest/Ranger usually reduce the risk of overfitting, through the effect of "strong law of large numbers" [3][4]. We try to use these models to predict real day ahead prices, and with the goal to also include unusually high prices (due to turbulence in market). A solution to improve the models would therefore to include more "unusually high prices" so that they become commonly observed in the sets of data. This would address the two remaining issues of too few datapoints and unbalanced spread between data sets. As a result, the RMSE and $R^2_{adjusted}$ should not increase and decrease, respectively. (Chuprunov & Fazekas, 2009) (Bernoulli, 1713)

Table 4: Validation results of Random Forest and Ranger model.

Regression model	**RMSE**	**RMSE (%)**[1]	$\boldsymbol{R^2_{adjusted}}$
Random Forest	32.2	54	0.75
Ranger	32.3	54	0.75

[1]Calculated from mean of test data set of 60,13 öre/kWh. in and max was -9,43 respectively 503,64 öre/kWh.

4.3 Test observations results

Two observations (2024-09-19 and 2024-10-12) had been collected separately and was used for prediction of unseen data. In Table the predictions and actual day ahead prices are presented. The difference is slightly lower compared to performance on the validation set (see Table). The results are at least verified by having the same order of magnitude. However, the model is probably not suitable from a practical perspective as 30-40 % error would impose high financial risks. Still, it is an interesting result that with little effort a model could be produced that does not give unreasonable predictions (e.g. 80 % RMSE).

Table 5: Predictions of unseen data.

Day ahead price	**2024-09-19**	**2024-10-12**
Actual	19	11
Predicted	43	35
Difference	44 %	31 %

5. Improved model performance with other data features

Future work with improving models to predict electricity day ahead prices should include the following:

- Dynamic exchange rate between EUR and SEK.
- Inclusion of more EU members (and sub-regions).
- More features (electricity production, supply and demand).

In the work for the current report these have been identified as important features that affect electricity prices. The European electricity market is connected, and therefore it is only rational that better models require data that better represent the entire market. (Mosquera-López, 2024)

6. Conclusions

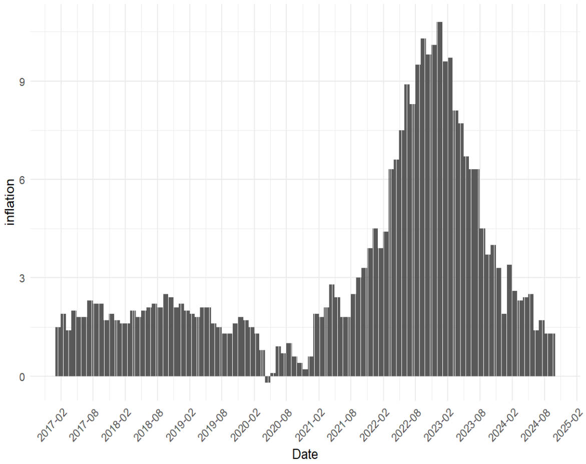
The objectives for the current project were fulfilled and the following conclusions are drawn,

- An ETL pipeline was created for data regarding inflation, weather and the day ahead market. The data was stored in an SQLite 3 database.
- The data was extracted and used to model day ahead prices. Linear, CART, Random Forest and Ranger models were examined, and Ranger model was deemed as best performing.
- The Ranger model did not live up to expectations for practical uses but could still give an approximation of the day ahead price (30-40 %).
- The European market is a network, which could potentially be better modelled by including data that better represent the entire network. Attempts to improve prediction capabilities of models should therefore include more features (i.e. electricity production, supply and demand) and from more regions in the European Union.

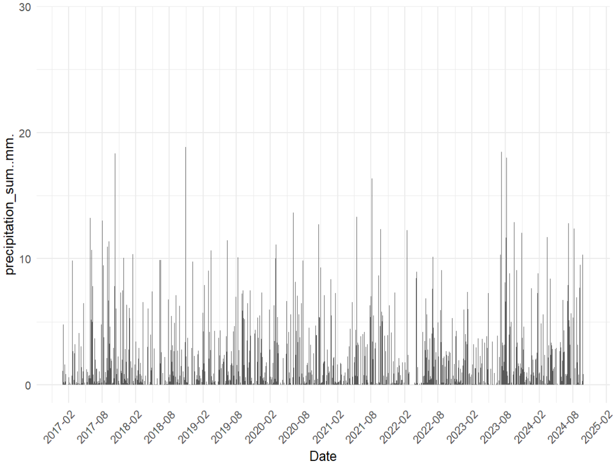
7 Appendix A

7.1 Variable plots

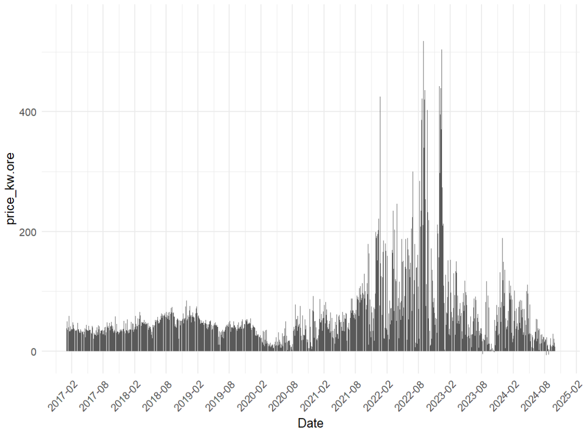
Date vs inflation



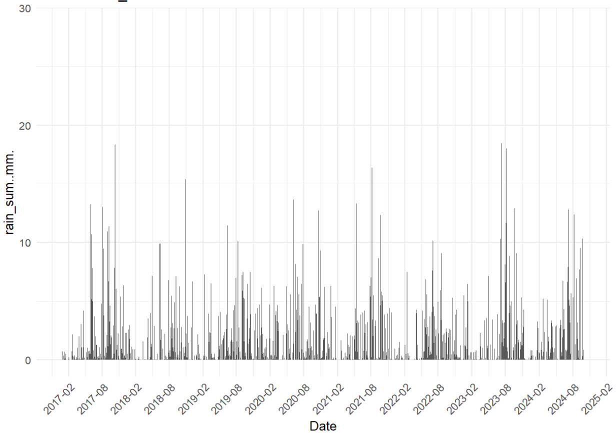
Date vs precipitation_sum..mm.



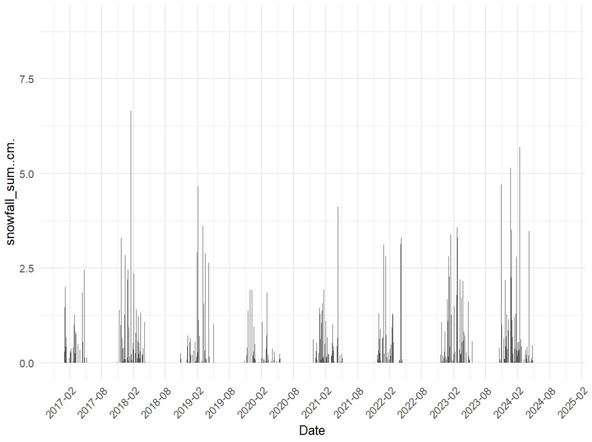
Date vs price_kw.ore



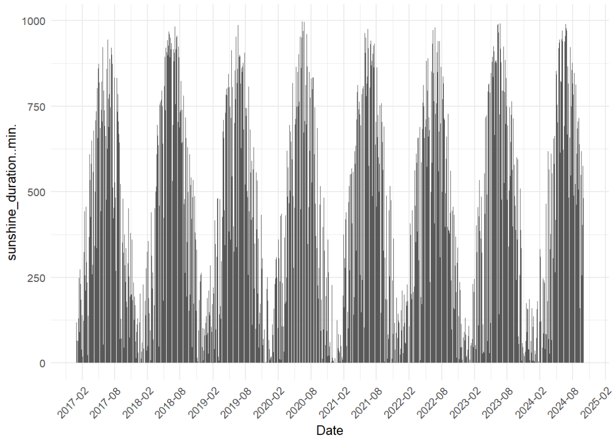
Date vs rain_sum..mm.

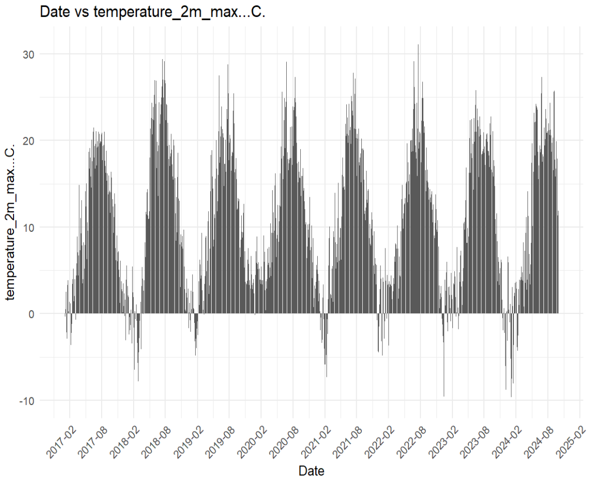


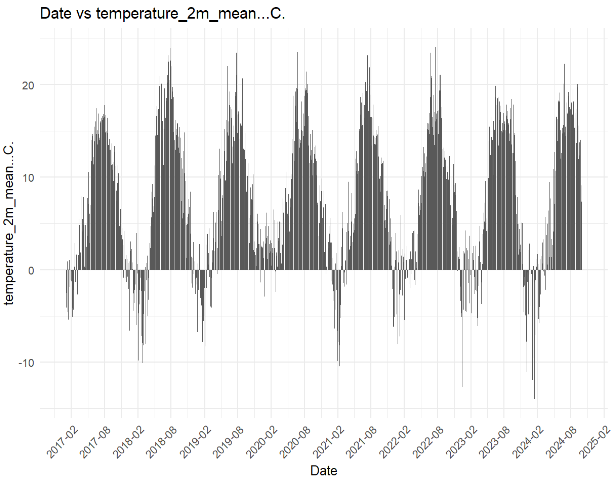
Date vs snowfall_sum..cm.

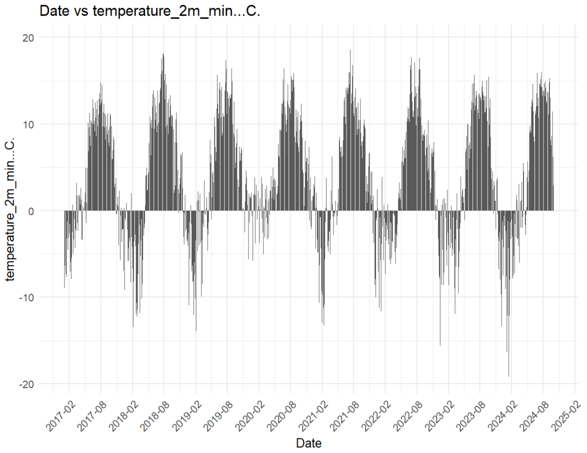


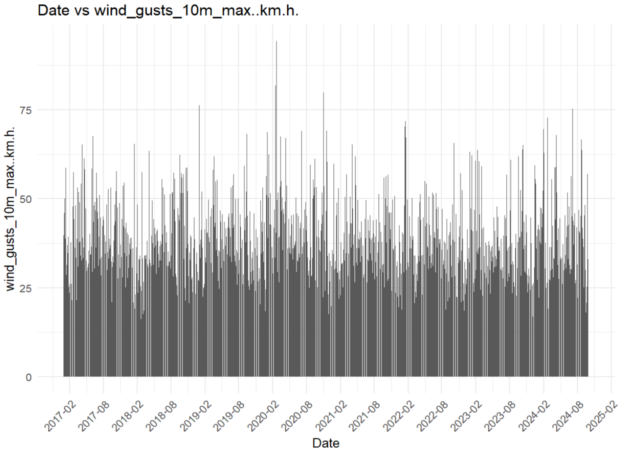
Date vs sunshine_duration..min.

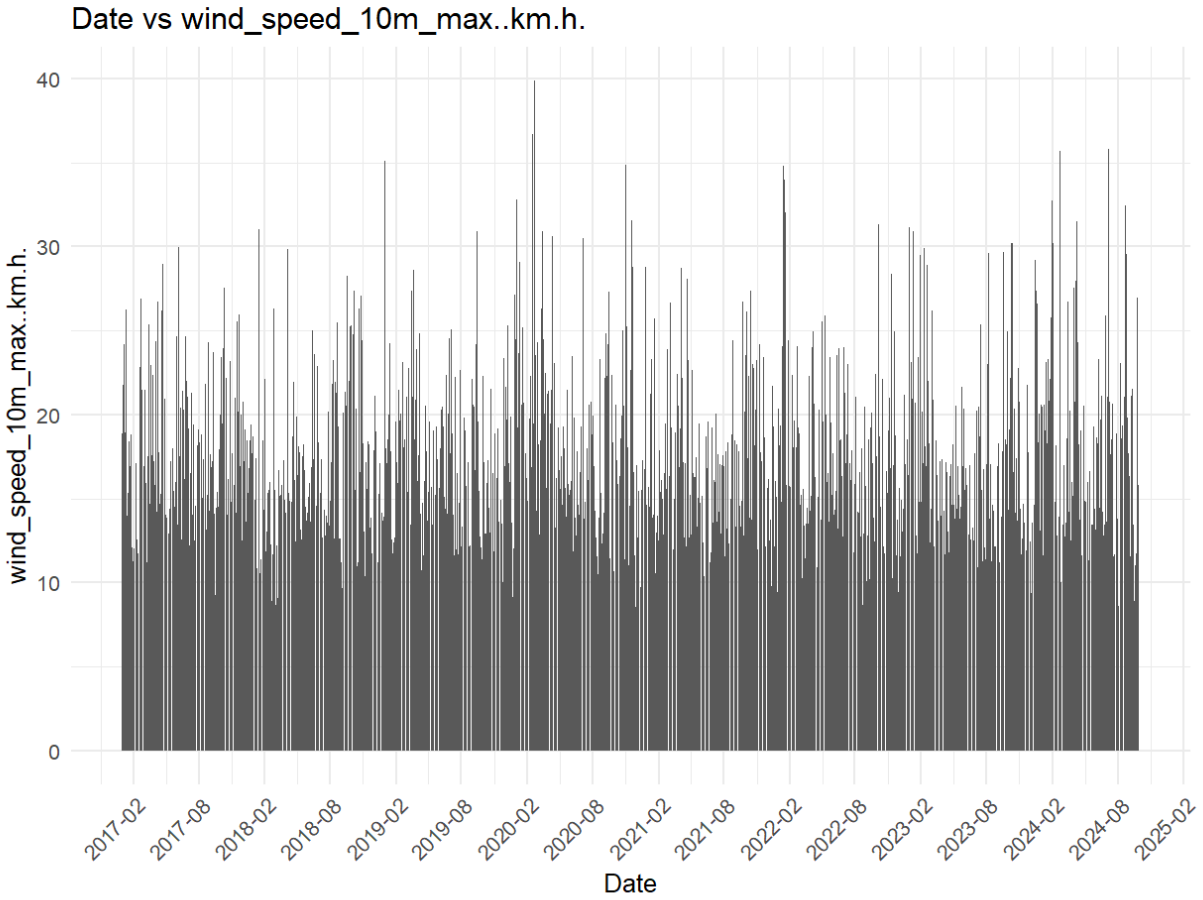



Date vs temperature_2m_max...C.
temperature_2m_max...C.
30
20
10
0
-10
2017-02
2017-08
2018-02
2018-08
2019-02
2019-08
2020-02
2020-08
2021-02
2021-08
2022-02
2022-08
2023-02
2023-08
2024-02
2024-08
2025-02
Date


Date vs temperature_2m_mean...C.
temperature_2m_mean...C.
20
10
0
-10
2017-02
2017-08
2018-02
2018-08
2019-02
2019-08
2020-02
2020-08
2021-02
2021-08
2022-02
2022-08
2023-02
2023-08
2024-02
2024-08
2025-02
Date


Date vs temperature_2m_min...C.
temperature_2m_min...C.
20
10
0
-10
-20
2017-02
2017-08
2018-02
2018-08
2019-02
2019-08
2020-02
2020-08
2021-02
2021-08
2022-02
2022-08
2023-02
2023-08
2024-02
2024-08
2025-02
Date


Date vs wind_gusts_10m_max..km.h.
wind_gusts_10m_max..km.h.
75
50
25
0
2017-02
2017-08
2018-02
2018-08
2019-02
2019-08
2020-02
2020-08
2021-02
2021-08
2022-02
2022-08
2023-02
2023-08
2024-02
2024-08
2025-02
Date


Date vs wind_speed_10m_max..km.h.
wind_speed_10m_max..km.h.
40
30
20
10
0
2017-02
2017-08
2018-02
2018-08
2019-02
2019-08
2020-02
2020-08
2021-02
2021-08
2022-02
2022-08
2023-02
2023-08
2024-02
2024-08
2025-02
Date

8 References

(u.d.). Hämtat från https://open-meteo.com/en/docs/historical-weather-api

Bergqvist, S. (2023). *Vattenfall*. Hämtat från https://www.vattenfall.se/fokus/trender-och-innovation/energianvandning-i-sverige-2022/

Bernoulli, J. (1713). *Ars Conjectandi.*

Breiman, L. (2001). *Random Forests.* Kluwer Academic Publishers.

Chuprunov, A., & Fazekas, I. (2009). *Strong laws of large numbers for random forests.*

Frost, J. (2023). *statisticsbyjim*. Hämtat från https://statisticsbyjim.com/regression/root-mean-square-error-rmse/

Grimmett, G., & Stirzaker, D. (2001). Probability and Random Processes: Third Edition. i G. Grimmett, & D. Stirzaker, *Probability and Random Processes: Third Edition* (ss. 325-331). Oxford University Press.

Guild, C. (den 28 08 2021). *Rpubs.* Hämtat från https://rpubs.com/camguild/803096

IBM. (den 18 01 2024). Hämtat från https://www.ibm.com/docs/en/cognos-analytics/12.0.0?topic=terms-r2

IBM. (den 18 01 2024). Hämtat från https://www.ibm.com/docs/en/cognos-analytics/12.0.0?topic=terms-adjusted-r-squared

IBM. (2024). Hämtat från https://www.ibm.com/topics/linear-regression

Mosquera-López, S. (den 23 07 2024). *SienceDirect*. Hämtat från https://www.sciencedirect.com/science/article/pii/S0140988324004973

mrmishraoofc. (den 29 07 2024). *geeksforgeeks.org*. Hämtat från https://www.geeksforgeeks.org/bayesian-information-criterion-bic/

rishu_mishra. (den 28 12 2021). *geeksforgeeks*. Hämtat från https://www.geeksforgeeks.org/k-fold-cross-validation-in-r-programming/

Saunders, T. (den 12 07 2023). *ScienceFocus*. Hämtat från https://www.sciencefocus.com/science/who-invented-electricty

Wright, M. N., Wager, S., & Probst, P. (2023). *A Fast Implementation of Random Forest.*

Yadav, A. (den 20 07 2024). *Medium*. Hämtat från https://medium.com/@amit25173/linear-regression-vs-random-forest-7288522be3aa

Zajic, A. (den 29 11 2022). *builtin*. Hämtat från https://builtin.com/data-science/what-is-aic