# Prediction of electricity consumption and production among certain prosumers in Estonia

Machine Learning competition launched by Enefit on Kaggle

**BOULLIAT Kevin - August 2024**

# Table of content


1

## INTRODUCTION

1. Introduction to the competition
2. Estonia's energy situation
3. Estonia's geopolitical situation
4. Presentation of Enefit

2

## MODELS PERFORMANCE RESULTS

5. Model used
6. Models MAE score results
7. Comparison of models' predictions

3

## BUSINESS STUDY

8. Methodology used
9. Cost of different models

4

## TECHNICAL DETAILS

10. Workflow and LGBM parameters

5

## CONCLUSION

11. Project conclusion
12. Areas of improvement
13. Use of AI and project limitations
14. About myself

6

## APPENDIX

15. Electricity market data analysis
16. Gas market data analysis
17. Solar weather data analysis
18. Focus on bad predictions

# INTRODUCTION

1. Introduction to the competition
2. Estonia's energy situation
3. Estonia's geopolitical situation
4. Presentation of Enefit

# Introduction to the competition

*Click here to access the competition on Kaggle*

The goal of the competition is to **create an energy prediction model of prosumers to reduce energy imbalance costs** in Estonia.

1

### What is a prosumer?

A prosumer is an individual or entity that **both produces and consumes goods or services**, such as energy.

2

### Why is predicting prosumer behaviour crucial?

**Prosumers**, who both consume and generate energy, **contribute a large part of the energy imbalance**. Despite being only a small part of all consumers, their unpredictable energy use **causes logistical and financial problems** for the energy companies.

3

### Which metric is used?

**MAE is used in this competition** because it provides a clear measure of average error in predictions, which is useful for **understanding performance in real-world scenarios**. It is also less sensitive to outliers.

# Estonia, a country still heavily dependent on fossil fuels...

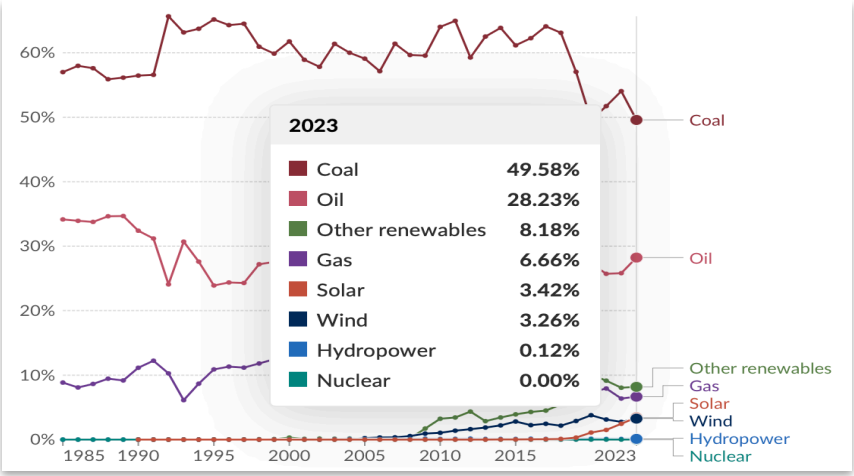


*Fig.1: share of energy consumption by source in Estonia*

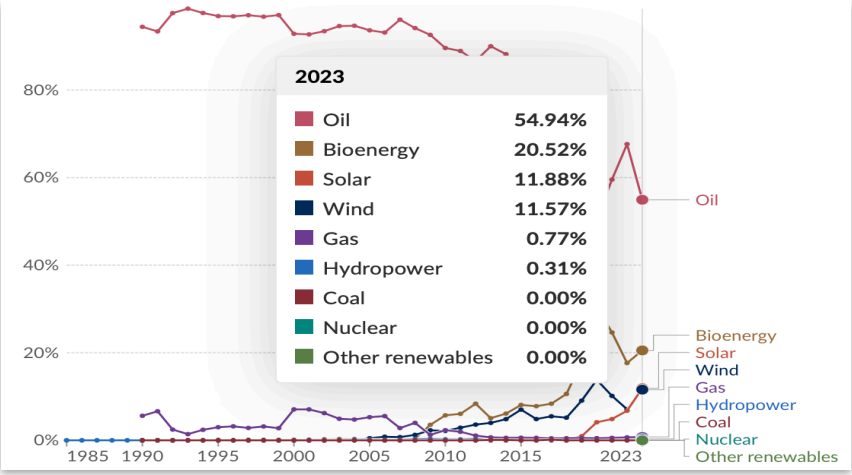


*Fig.2: share of electricity production by source in Estonia*

- Estonia relies heavily on fossil fuels in its energy mix (see Fig. 1) and for electricity production (see Fig. 2). **The lack of nuclear energy, which would provide abundant and flexible power, combined with the country's harsh climatic conditions, makes it crucial to depend on fossil fuels to accurately balance electricity production**.
- In the context of this competition with prosumers, the only source of energy production considered is solar energy. Looking at Figure 2, we can see that **solar energy accounts for only 10% of electricity production sources**. **This provides better context for this project**'s energy landscape.

*Source: ourworldindata.org, hikerbay.com*

# ... and strongly impacted by Russia's invasion of Ukraine

*Fig.3: Estonia's geographical location in Europe*

*Fig.4: Electricity prices in Estonia between 2019 and 2024*

- Estonia is a country in Eastern Europe that shares a border with Russia. The Russian invasion of Ukraine in 2022 led to significant disruptions in the global energy market, particularly affecting natural gas supplie in Europe. **This caused electricity prices to spike due to increased demand for alternative energy sources** and the uncertainty surrounding energy security.
- Figure 4 shows that electricity prices have almost returned to normal. However, price instability remains due to the ongoing conflict in Ukraine. **This highlights the importance of developing predictive models for companies in the energy sector.**

*Source: statista.com*

# Eesti Energia, a major energy player in Estonia

| Company | Activities | Market share | % of Estonia's electricity prod |
|---|---|---|---|
| Eesti Energia | Electricity production (oil shale, renewables), energy sales, and energy services.<br>***Enefit, Enefit Power and Enefit green are subsidiaries*** | ~60% | ~80% |
| ALEXELA | Electricity and gas supply, focusing on sustainable energy solutions | ~10-15% | *Not significant* |
| elektrum | Electricity supply and energy services | ~10% | *Not significant* |

*Tab.1: Major energy players in Estonia*

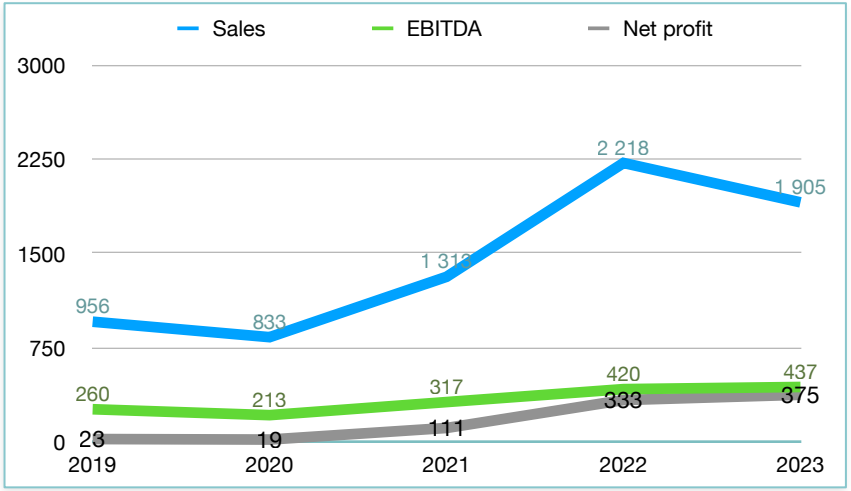


*Fig.5: Key financial indicators for Eesti Energia between 2019 and 2023*

- Eesti Energia is the largest energy company in Estonia. It owns several subsidiaries, including **Enefit, which encompasses Enefit Power and Enefit Green, specializing in the energy sector's transition to renewable energy sources**. The Estonian company dominates the energy market with a 60% market share.
- By looking at the financial indicators, we can see that **2022 was a significant year for the company**. This can be explained by the conflict between Russia and Ukraine, which **allowed many energy players in Europe to make substantial profits** due to the increase in electricity prices.
- **Artificial intelligence**, along with consumption and production predictions, **represents major topics for enabling these companies to remain competitive** in the years to come.

*Source: Enefit.com, wikipedia.com*

# MODELS PERFORMANCE RESULTS

1. Models used
2. Models MAE score results
3. Comparison of models' predictions

# 7 different models were used in this project

**BASELINE MODELS**

**1. « Dummy » model:** simple baseline that predicts the mean of the target values from the training data, allowing comparison with more advanced models.

**2. « Engineer » model:** a step up from the baseline, this model uses simple, intuitive assumptions:

- Consumption prediction: assumes the same electricity consumption as the previous year on the same date (e.g., April 5, 2023, predicted from April 5, 2022).
- Production prediction: based on a strong correlation with « direct_solar_radiation" feature, it models production using a simple ratio, reflecting solar panel dependency.

**BASIC ML MODELS**

**3. Linear Regression model:** I started with linear regression because it's simple and interpretable. It helped me understand the basic relationship between variables. Although it's not the most powerful model, it's a great baseline that provides insights into which features might be influential.

**4. Random Forest model:** I used a random forest model to capture more complex interactions between features. Indeed, Random forest models are great for handling a mix of numerical and categorical data and they can model non-linear relationships. It's a very popular model to try in a data science project. It's particularly useful when there's a lot of noise in the data because it's less likely to overfit compared to other models.

**ADVANCED ML MODELS**

**5. XGBoost model:** to push the accuracy further, I decided to use a XGBoost model. This model was very popular in other competitor projects. It's a powerful boosting algorithm that usually outperforms random forests by iteratively improving the predictions. In this project, XGBoost was useful because it handled the large amount of data well and was effective at capturing intricate patterns in the electricity consumption data.

**6. LGBM model:** finally, I added LGBM (LightGBM) to the mix. It's similar to XGBoost but faster and more efficient, especially with large datasets. LGBM was particularly helpful in speeding up the training process without sacrificing performance, which was important given the size of the data I was working with. This model was slightly better than the XGBoost model.

**7. LGBM model optimized :** since the LGBM model was the most effective, I decided to optimize it using grid search. The process was guided by minimizing the MAE (Mean Absolute Error) on the validation set.

# An optimized LGBM model that performs very well compared with the other models used.

| # | Models | MAE Score |
|---|---|---|
| 1 | « Dummy model » | 444.12 |
| 2 | « Engineer model » | *Only used for graphic visualization* |
| 3 | Linear regression | 414.55 |
| 4 | Random forest | 101.92 |
| 5 | XGboost | 109.25 |
| 6 | LGBM | 92.93 |
| 7 | LGBM optimized 1 | 88.34 |
| 8 | **LGBM optimized 2** | **44.72** |
| | | |
| 9 | Competition winner | 52.30 |

*Tab.2: Different models used - MAE scores comparison*

*« A MAE score of 44.72 means that, on average, the model's predictions are off by 44.72 MWh from the real values »*

- **Advanced models vs. baseline:** it is evident that the more advanced models perform significantly better compared to the baseline.
- **Linear regression performance:** the Linear Regression model did not perform well compared to the baseline. It is likely that the data scaling was not precise enough, which may have negatively impact for that kind of model.
- **LGBM model optimization:** the LGBM model, which performed the best, was optimized twice. First, using grid search, which slightly improved the score. The second improvement came when the code was refined, specifically by adding the function y_pred = y_pred.clip(0) to set all negative predictions to zero. Since negative consumption or production values are not feasible, this adjustment greatly improved the model's accuracy.
- **Comparison with winner's model:** while the winner's model might appear less performant than my LGBM optimized 2 model, the winner's model was evaluated on recent 2024 data connected directly to the company's API. Due to the competition being closed, I was unable to replicate this. My evaluation was limited to the Test set (30% of the final dataset), which may explain some overfitting on the training data and therefore a better score than the winner of the competition.

# Visually accurate production and consumption predictions

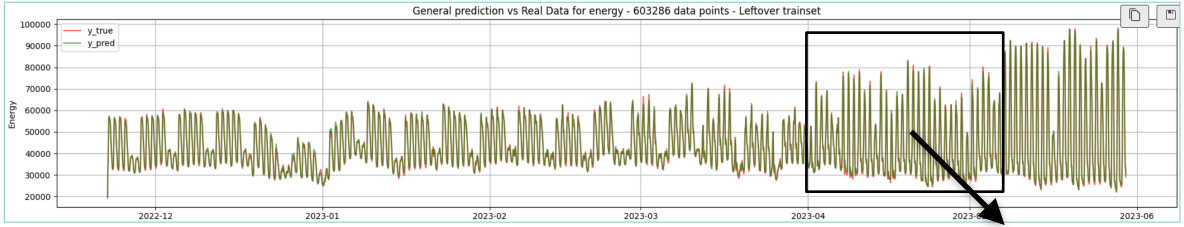


*Fig.6: 6-months global prediction on the test set*

- The predictions on the test set cover the 6 months, resulting in **603,000 data points that have been forecasted**.

- Zooming in on a month, specifically April, provides a **clearer view of the predictions and actual values** for electricity production and consumption.
- It is evident that the **predictions are very accurate, closely matching the real values.**

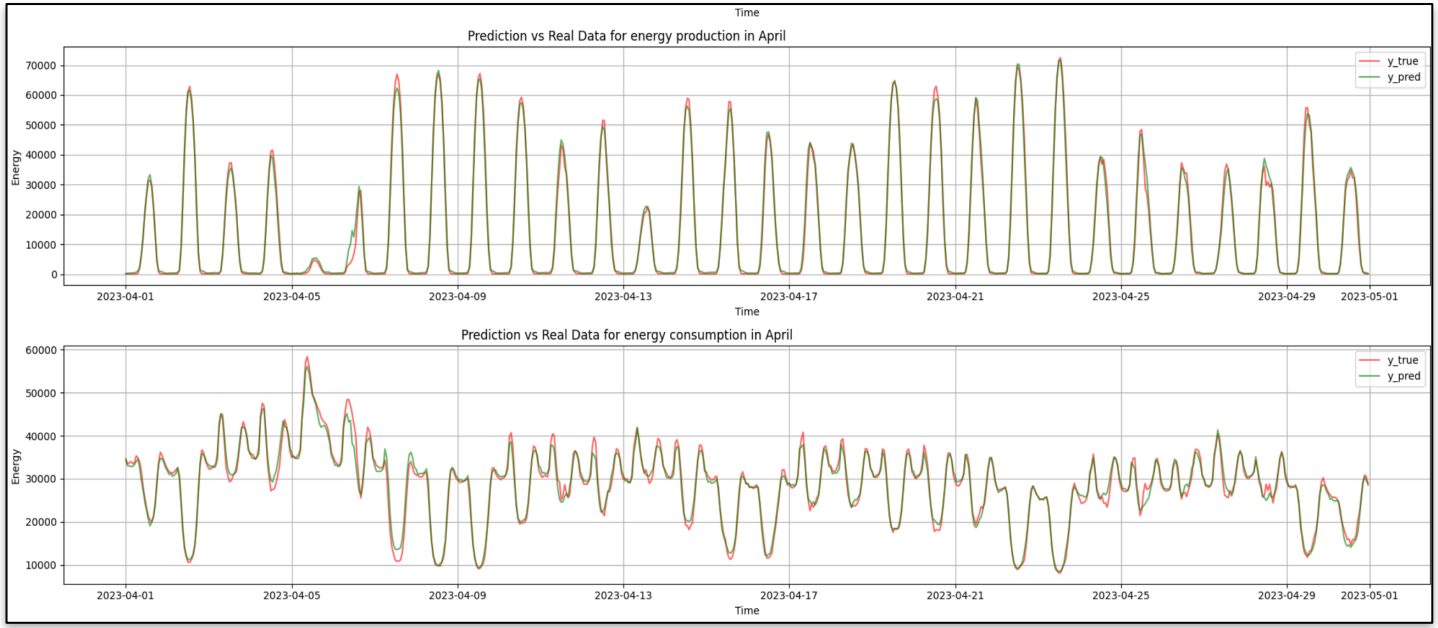


*Fig.7: zoom on the predictions for April 2023*

# Comparison of different models for electricity production

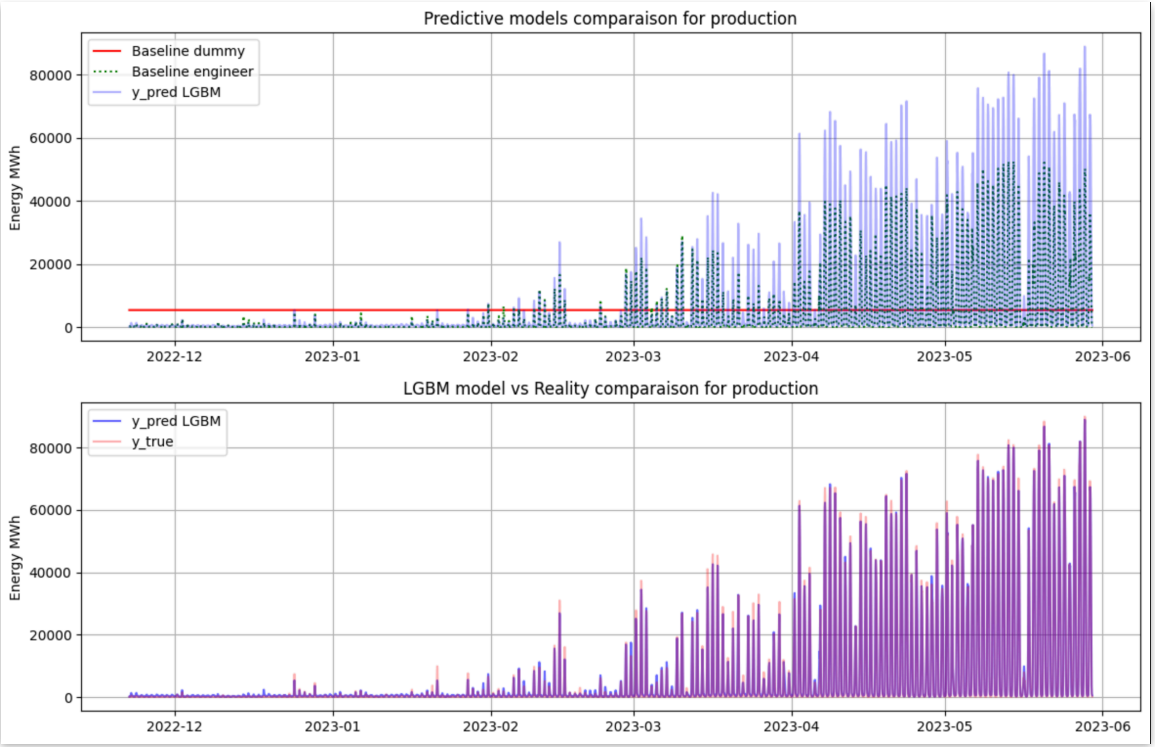


*Fig.8: models comparaison for electricity production*

- It is evident that daily electricity production logically increases as summer approaches.
- The **dummy model logically provides poor predictions** that do not align with reality.
- The « **engineer » model shows decent quality**, following real production trends. However, it appears **slightly under-designed** due to the assumptions made.
- The optimized **LGBM model provides accurate predictions compared to actual production** values.

# Comparison of different models for electricity consumption

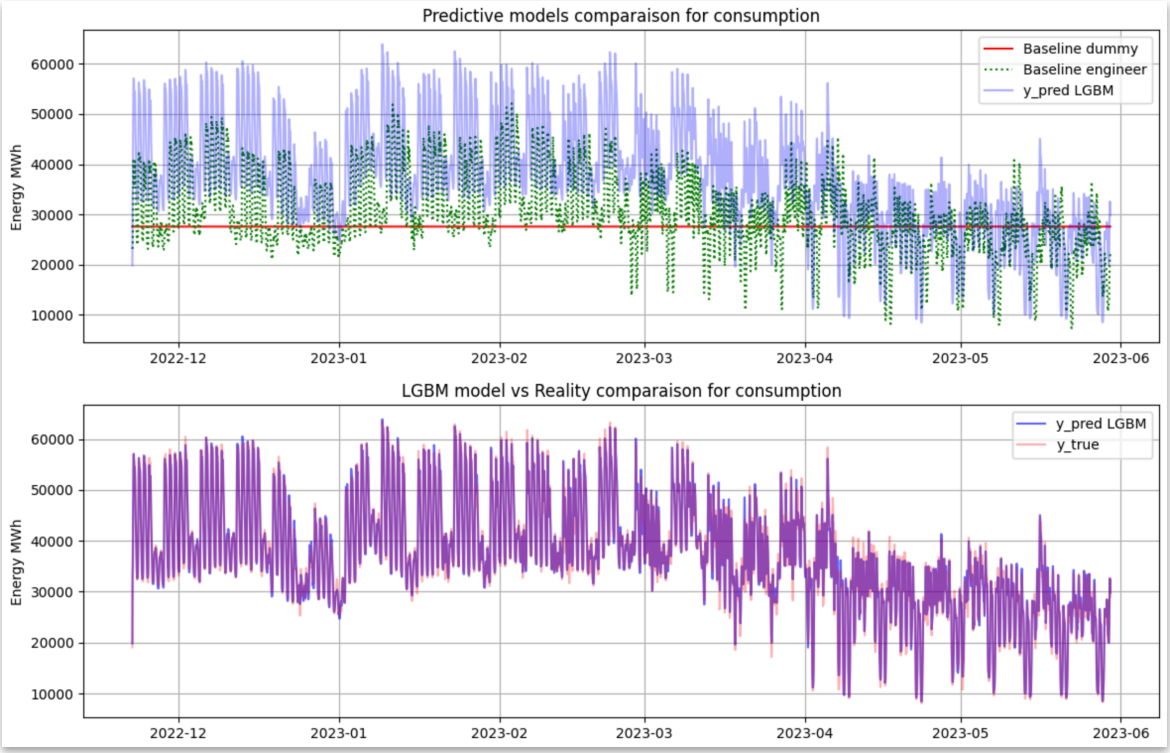


*Fig.9: models comparaison for electricity consumption*

- There is a strong seasonality in electricity consumption with both daily and weekly periodicity. Consumption is lower at night and also during weekends due to reduced industrial activity.
- The **dummy model logically provides poor predictions** that do not align with reality.
- The **« engineer » model is interesting but underperforms**. It effectively captures the daily and weekly seasonality but underestimates the values compared to reality. This suggests that electricity consumption in 2022 was lower than in 2023, potentially due to the Russia-Ukraine conflict.
- The optimized **LGBM model appears to have good accuracy compared to actual production values**. The actual and predicted curves from the model are almost identical.

# BUSINESS STUDY

1. Methodology used

2. Cost of different models

# Methodology for calculating the gains of the models

**Hypothesis 1:** the energy produced by prosumers is immediately consumed by them according to their instantaneous needs.
**Hypothesis 2:** it is assumed that the buying and selling of electricity from prosumers are done at the same rate, which is not the case in reality.
**Hypothesis 3:** it is assumed that all prosumers have the same energy buying and selling contract, which is not true in reality.

## 1 Calculate the difference between production and consumption

$$Diff_{modelx} = prod_{modelx} - cons_{modelx}$$

**Case 1:**
**Diff > 0 :** this indicates a surplus of energy being fed into the grid, meaning more energy is being produced than consumed.

**Case 2:**
**Diff < 0** : more energy is consumed than produced, requiring energy from the grid.

This calculation is performed for all the models to be compared together.

*Note: this calculation is made for each hour of the day.*

## 2 Calculate the residuals between the true values and the model predictions.

$$Diff_{global} = Diff_{true} - Diff_{modelx}$$

**What does this difference represent?**

This difference indicates the expected deviation from reality if the company Enefit had actually adopted one of the models as a reference for its consumption and production predictions.

*Note: this calculation is made for each hour of the day.*

## 3 Multiply the difference from step 2 by the price of electricity to obtain the cost.

$$Cost = Diff_{global} * Price_{electricity}$$

**What does this number represent?**

This number indicates the cost by Enefit incurred due to the difference between a model's predictions and the actual values. It represents the cost Enefit would have paid if the company had adopted one of the models as a reference.

This calculation is made for each hour of the day, and then the cumulative sum is taken to observe the evolution of the cost for the company Enefit.

# The LGBM model provides significant financial gains compared to more basic models

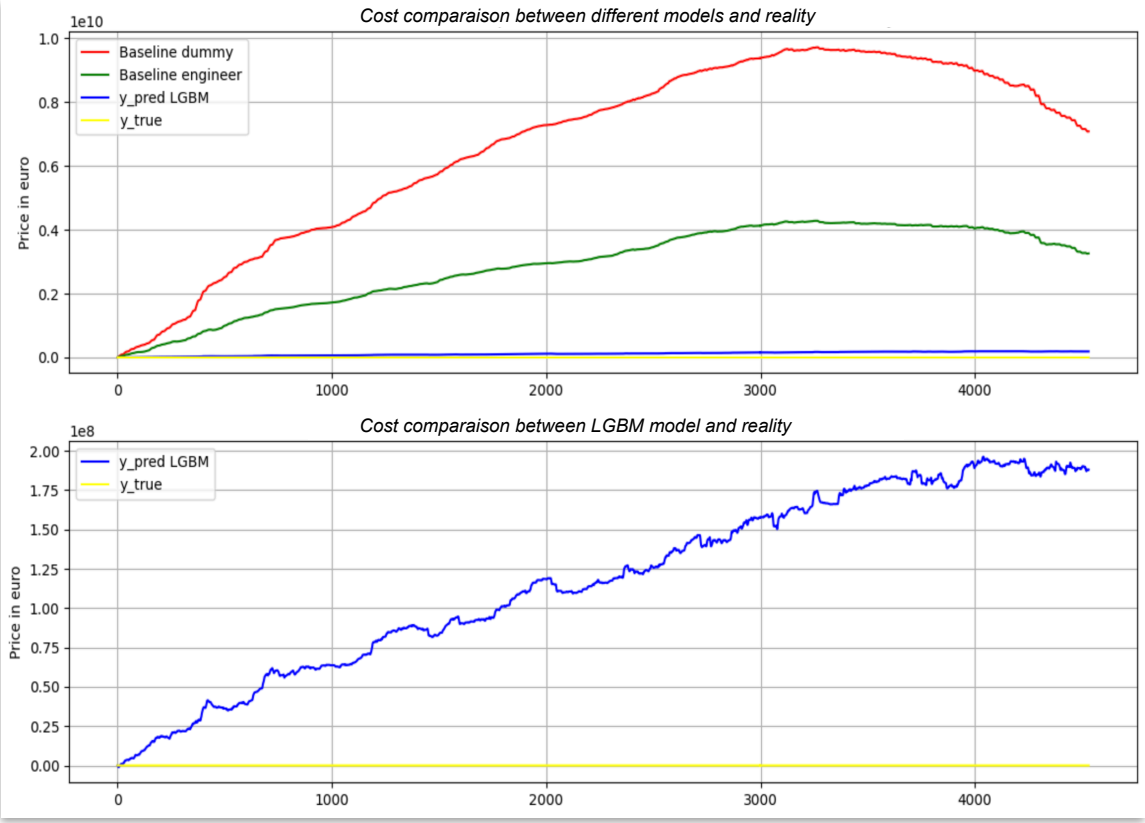


*Fig.10: Potential cost of using different models*

As a reminder, many strong assumptions were made in the context of this project.

- The different **curves represent the cost Enefit would have incurred if the company had used one of the models as a reference** for its consumption and production predictions. Consequently, the yellow curve, which represents the actual values, shows zero cost.
- One **strange thing to note here is the price scale, which seems excessively large**. Despite thorough investigation, I couldn't pinpoint the source of this issue. However, it's useful to keep in mind the relative differences between the various curves.
- The top graph shows that the LGBM model significantly reduces costs for the company compared to the dummy and engineer models. **This highlights the importance of implementing predictive models in energy companies to minimize costs**.
- Zooming in on the LGBM model in the second graph, we see that costs consistently increase. This is because **solar panel production never fully offsets prosumer consumption** over the long term.

# TECHNICAL DETAILS

1. Workflow and LGBM parameters

# Workflow and LGBM parameters

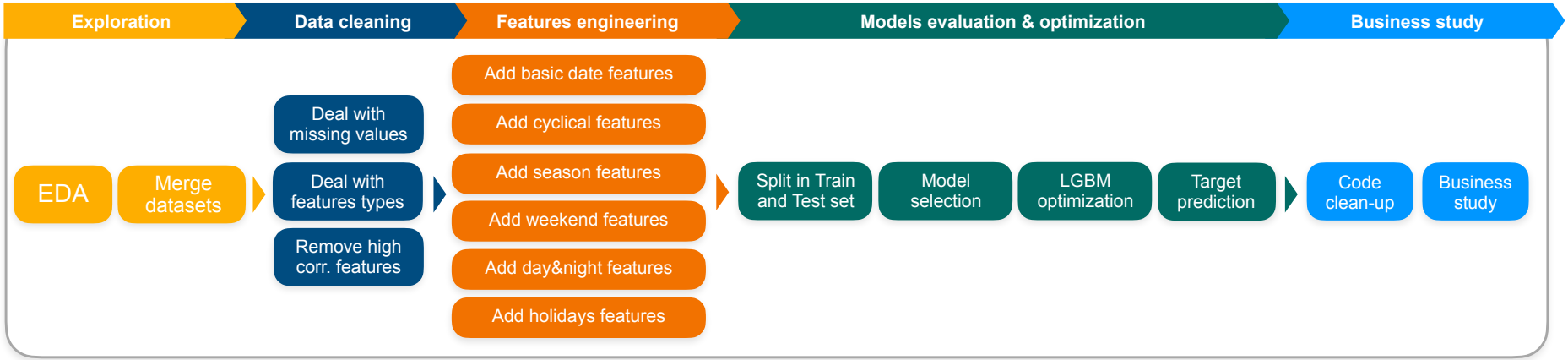


*Fig.11: Global workflow used in this project*

*7 different models were used. To better understand why I chose these models, please refer to slide 9.*

**LGBM model params used**

1. **Objective: 'regression'** - Defines the task as regression to predict continuous values.
2. **Metric: 'MAE' -** Uses Mean Absolute Error (MAE) to assess model accuracy, reflecting the average prediction error in real-world units.
3. **Boosting_type: 'gbdt' -** Gradient Boosting Decision Tree captures complex relationships with low variance, crucial for accurate predictions.
4. **Num_leaves: 31 -** Limits tree complexity to prevent overfitting while capturing important patterns.
5. **Learning_rate: 0.05 -** A lower learning rate fine-tunes predictions, reducing overfitting risk and improving model generalization. 0.05 is a classic learning rate.
6. **Feature_fraction: 0.9 -** Uses 90% of features for each training iteration, enhancing model robustness and generalization.
7. **Early_stopping_round: 100 -** Stops training early if no improvement after 100 rounds, avoiding overfitting and saving computation time.

# CONCLUSION

1. Project conclusion
2. Areas of improvement
3. Use of AI and project limitations
4. About myself

# 3 ideas to take away from this project

## 1 An efficient LGBM model developed

The **LGBM model performed well by making accurate predictions** (with an average error of 44 MWh).

Graphically, **the LGBM model successfully reproduced trends and seasonality and almost perfectly matched the actual values** for both production and consumption of prosumers.

## 2 An imperfect but interesting model for costs saving

As we have seen previously, **the scale of the costs** shown on slide 16 **seems far too large to be realistic**.

Nevertheless, it can be noted that a well-performing predictive model, like the one developed, can **save significant costs compared to simpler, non-predictive models**.

## 3 A real business challenge for companies

This very interesting project highlights a **real business challenge faced by companies in the energy sector** related to grid stability.

This instability results in high costs for these companies. It **demonstrates the importance** for energy companies **to work on predictive models**.

# Taking a step back and identifying areas for improvement

Many areas for improvement could have been explored. I decided to highlight 4 of them below:

1. **Creating separate models**
I think it might have been beneficial to develop two distinct models for this project. One focusing on predicting consumption and the other on production. Since production and consumption are independent, exploring this approach could have potentially improved the model's performance.

2. **Ensemble method**
Many top competitors on Kaggle have used sequential or ensemble methods to enhance prediction accuracy. Although this approach is technical and time-consuming, it could have been explored in my project to achieve better results.

3. **API testing**
I would have liked to test my model via the company's API to get an exact score for the project and obtain an official ranking on Kaggle. Unfortunately, this was not possible when I started the competition as it was closed.

4. **Code improvement**
The final code could have been improved. Many highly-ranked competitors used classes in their projects, which can elevate code quality and make it more professional. I am currently not proficient with classes, but learning to use them could enhance the project.

# Use of AI and project limitations

## Artificial intelligence as a senior personal data scientist:

In this project, the use of artificial intelligence was a great help to me. I used the GPT-4 model from OpenAI. In particular, **I utilized this model as an assistant, as if I were working with a senior data scientist**. The main goal of this project for me was to improve my skills in handling and predicting time series.

Artificial intelligence in this project **allowed me to better understand time series, elevate my code to a more professional level, and gain a deeper understanding of the models I was using**, just as a senior data scientist would have done in real life.

## More data would have allowed for further research and deeper exploration in this project:

To take this project further, **it would have been interesting to obtain data on the production of other types of energy in Estonia.** Having access to this data would have allowed me to compare my results and provide more nuanced insights.

It would also have been valuable to **have access to a longer historical dataset**. In fact, only the year 2023 was complete, and due to the war in Ukraine, the available data does not reflect the trends from the years before the conflict.

## Peer review cannot yet be replaced by AI:

Although artificial intelligence has been a great help, acting as a senior data scientist, **the inability to discuss and share my ideas for this project with a real person was a significant drawback**. Indeed, throughout my work, **it is possible that coding or reasoning errors may have been made.**

# Know more about myself

Kevin BOULLIAT

Kevinbllt

## Kevin BOULLIAT

**Email:** k.boulliat@gmail.com
**Address:** 2323 Lake Shore Blvd W, Etobicoke

**PR card - Do not need sponsorship**

**Driving licence**

- **French:** mother tongue
- **English**: bilingual
- **German**: intermediate

### About my interest in data and energy topics:

As a data enthusiast, **I thrive on making data speak and uncovering insights** through meticulous analysis and manipulation. I am passionate about bridging data with business needs to **make complex information accessible and solve real-world problems**. My goal is to **leverage data to drive meaningful decisions** and contribute to impactful business solutions.

I am also **passionate about the energy sector,** a field in which I have specialized through education and professional experiences. I firmly believe that **data plays a pivotal role in the sector's success**. Therefore, I am determined to advance my career in energy, focusing on advanced analytics topics, where I **can bring my technical and business-oriented approach** to drive value.

### About my professional experience:

My past experiences in engineering, first as a project manager and then at ENGIE as a consultant in transformation and strategy, followed by a role as a Data and Business Analyst, have provided me with a **versatile skill set.** This allows me to be **effective across a wide range of topics, both technical and business-related**. I am able to **handle complex issues and work efficiently with diverse stakeholders**, all while delivering substantial added value.

### About my mindset:

In my daily work, I always strive to bring qualities that I consider essential: **a high standard of work, personal excellence, strong interpersonal skills, and the achievement of tangible results**.

# APPENDIX

1. Electricity data analysis
2. Gas data analysis
3. Weather data analysis
4. Focus on bad predictions

# Seasonality of electricity prices

Here are the data provided by Enefit on Kaggle

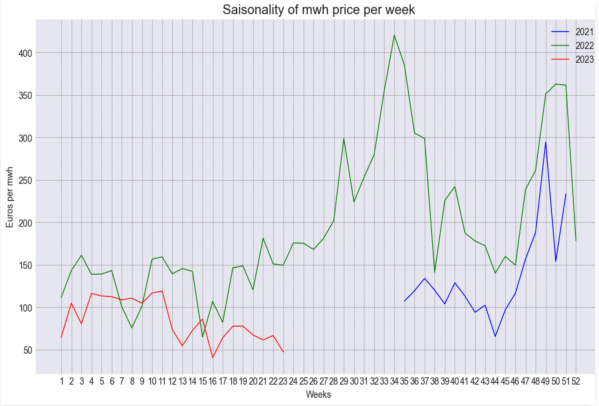


*Fig.12: Electricity price per week per year*

*Fig.13: Electricity price per month per year*

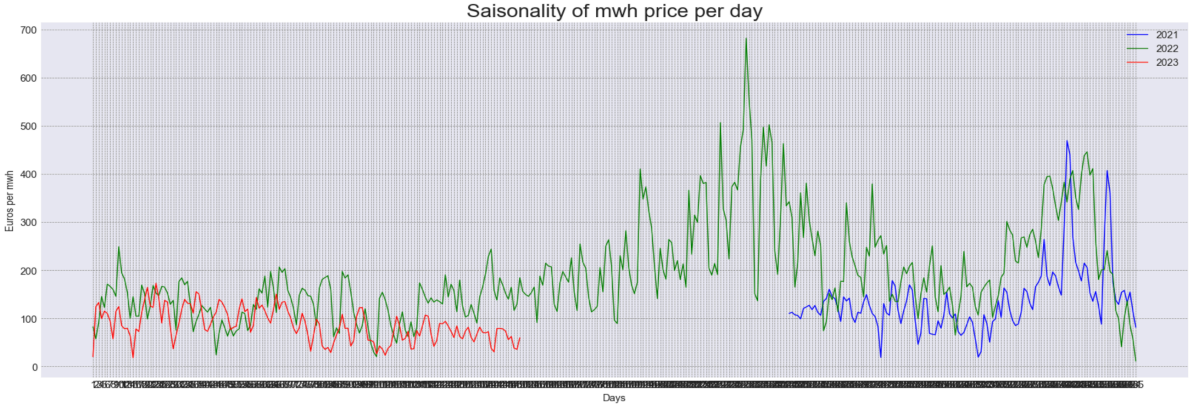


*Fig.14: Electricity price per day per year*

- **There appears to be some seasonality**. Indeed, towards the end of the year, electricity prices seem to be higher, which can be expected assuming higher demand, particularly in an eastern country like Estonia.
- The **overall price trend seems to be rather constant, with prices below 500 euros**. At certain times of the year, there are more significant variations, which can likely be explained by local and possibly political phenomena.
- **The war in Ukraine beginning in 2022 appears to affect electricity prices**. Indeed, prices in 2022 are higher than those in 2021 and 2023, although they remain within a reasonable price range.

# Seasonality of gas prices

Here are the data provided by Enefit on Kaggle

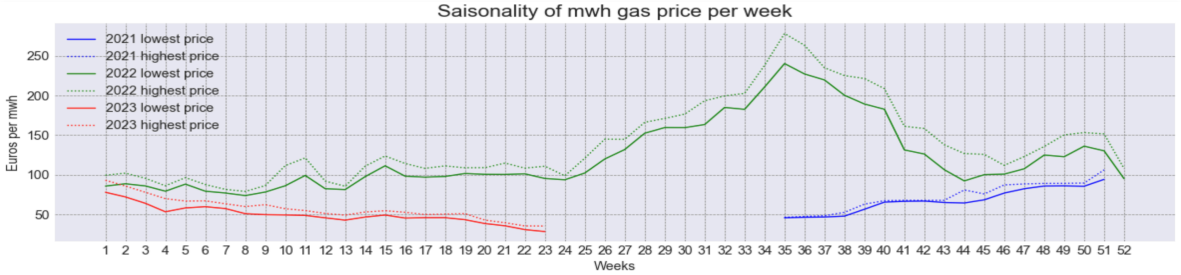


*Fig.15: gas price per week per year*

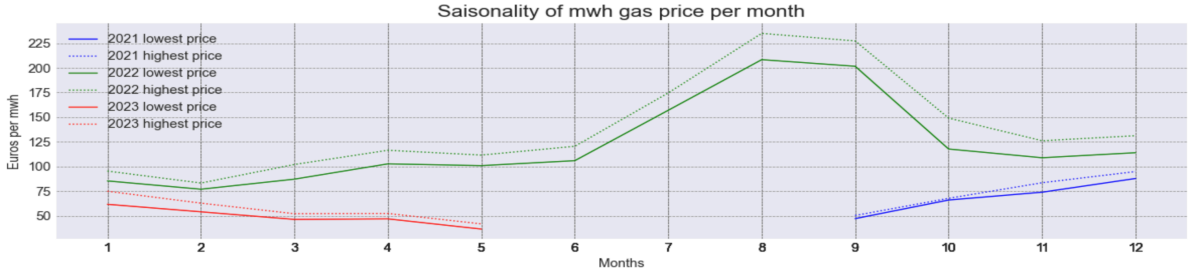


*Fig.16: gas price per month per year*

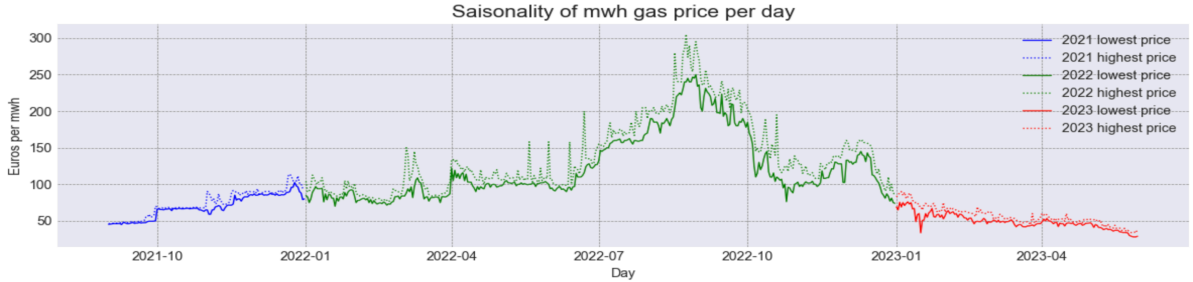


*Fig.17: gas price per day per year*

- It appears that **there is a seasonality effect in gas prices**. Indeed, prices seem to decrease as summer approaches and increase as winter approaches.
- The year **2022 seems particularly more expensive in terms of prices compared to 2021 and 2023**. One can assume that the beginning of **the war in Ukraine heavily impacted the Estonian gas market as it did in Western Europe**. Unfortunately, we do not have enough data for 2021 and 2023 to conduct a complete week-by-week study and compare it to 2022.
- Variations between the high and low prices of gas per day are far from constant; significant variations are recorded between these two prices, indicating that the gas market is tight.
- **Gas prices appear to return to more reasonable levels in 2023** compared to 2022.

# Solar weather features evolution

Here are the data provided by Enefit on Kaggle

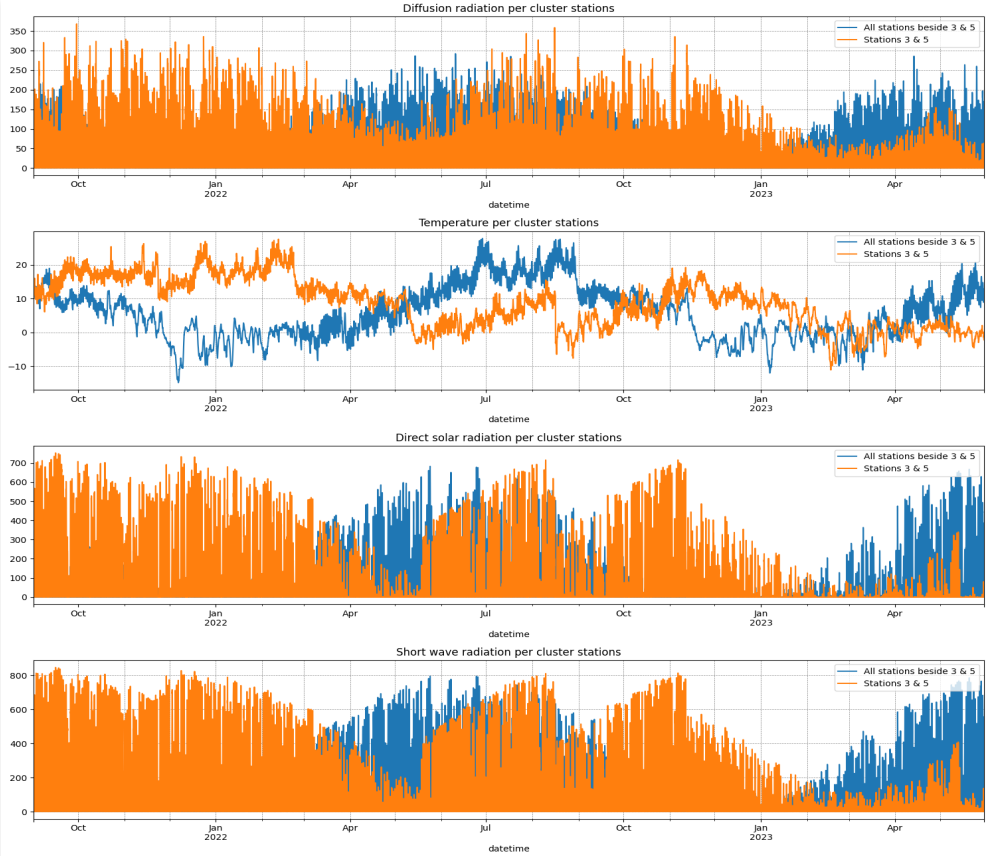


*Fig.18: solar features evolution across different weather station in Estonia*

- **Two weather stations, 3 and 5, appear to be very different in terms of meteorological readings**. For example, in January, these two stations have high solar radiation values compared to other stations. However, on the map, these two stations do not appear to be located in areas more conducive to receiving sunlight.
- The **variations in solar characteristics seem to follow expected patterns**: higher values during summer and lower values towards winter.

# Focus on bad predictions

As a reminder, predictions have been made between December and June (6 months prediction)

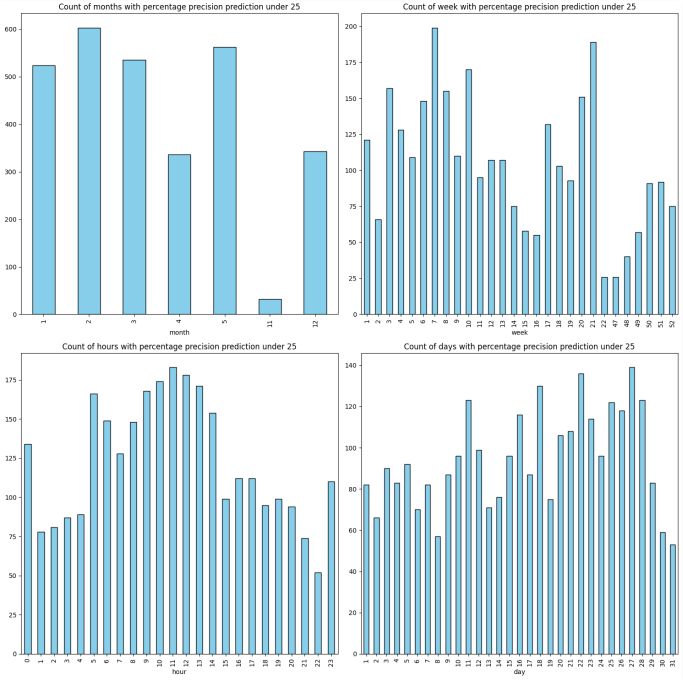


*Fig.19: Detail of the production prediction under 25% of accuracy*

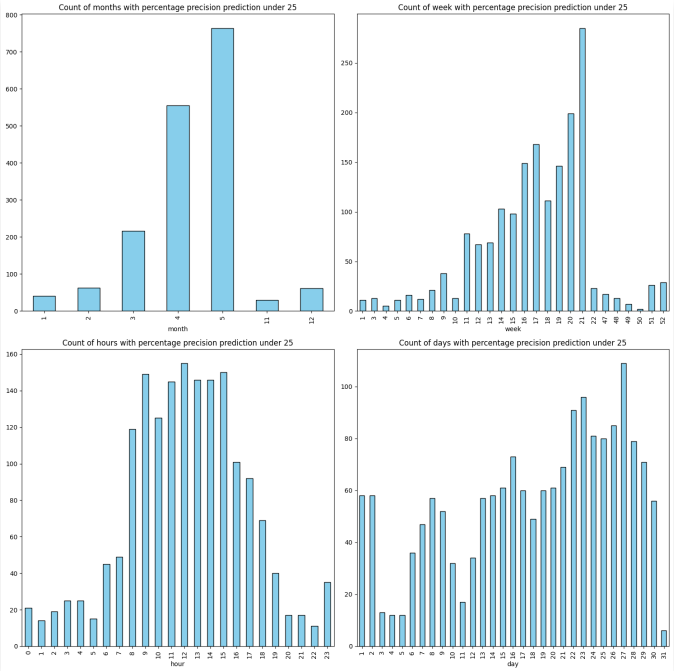


*Fig.20: Detail of the consumption prediction under 25% of accuracy*

- The predictions started at the end of November 2022, **which explains why the weeks and months associated with November have low values in the histograms** at the top of the two figures
- Regarding the **production predictions (Figure 19), we can see that no clear trend emerges from the histograms**. One might imagine that due to the strong correlation between production and meteorological data on sunlight, **it is easier for the model to make accurate predictions**.
- Concerning the consumption predictions (Figure 20), there are **noticeable trends in the "bad predictions."** For instance, in April and May, prediction deviations greater than 25% were quite frequent.
- Similarly, the **midday hours seem more challenging to predict** for consumption.

*Here only the predictions with a deviation of more than 25% from the actual values have been isolated to try to identify trends in these “bad predictions”.*

# Thank you for reading my work!

*Feel free to contact me*