Gathering and preparing data, building and evaluating a predictive model for heatpump energy consumption

Viktor Berggren
Sollentuna, Sweden
viktor.berggren@outlook.com

***Abstract*— This report details the methodology and findings of a capstone project for the course "Introduction to Applied ML." Data from multiple sources, pertaining to heat pump energy consumption and relevant weather parameters, was collected, cleaned, and processed within a Jupyter Notebook environment. Subsequently, two predictive models, Linear Regression and Random Forest, were trained and evaluated to forecast heat pump energy usage.**

Keywords—Artificial Intelligence, Machine Learning, Linear Regression, Random Forest, heatpump energy consumption

I. Introduction

This report documents the final project (examination type 2) for the course "Introduction to Machine Learning" (ITM600) at Högskolan Väst.

II. The Assigment

The project objective was to apply Machine Learning (ML) techniques to a real-world problem. The findings and solutions were presented through an online presentation accompanied by this report (examination type 2). The dataset was self-sourced (BYODS).

III. The challange

The challenge was to predict future heat pump energy consumption for a geothermal heat pump installation in a 120 m² house (plus equally sized basement) built in 1945. The heat pump was installed in June 2024, and 308 days of detailed consumption data were available, covering almost a full heating season and including domestic hot water energy consumption. While a time series model might be ideal, linear regression and random forest models were recommended and utilized for consistency with the course content.

IV. IV. DATA GATHERING

Heat pump energy consumption data was retrieved from a local Home Assistant installation. Weather data, comprising temperature, wind speed, and solar radiation, was deemed most relevant for predictive modeling. Hourly data for temperature, wind speed, and solar panel energy production (used as a proxy for solar radiation's impact on house heating) were obtained from a solar energy production web service.

V. Data cleaning

Data was initially organized using Excel to achieve hourly granularity. Missing data for three days in November was imputed by manually aligning hourly consumption to match the total consumption of 73 kWh for that period, mirroring the consumption patterns of adjacent days. The two data sources were merged into a single table and exported as a CSV file for further processing in a Jupyter Notebook.

Within the Jupyter Notebook, an extraneous column introduced during import was removed, empty cells were filled with 0, and time data was formatted into day and hour components. Heat pump data integrity was verified. Descriptive statistics and distributions of the data frame variables were examined (Table 1).

	time	temp	wind	solar_cells	heatpump
count	7416	7416.000000	7416.000000	7416.000000	7416.000000
mean	2024-11-16 11:30:00.000000256	8.056088	12.174939	557.779935	0.698429
min	2024-06-15 00:00:00	-10.650000	1.100000	0.000000	0.000000
25%	2024-08-31 05:45:00	2.300000	8.150000	0.000000	0.100000
50%	2024-11-16 11:30:00	7.550000	11.300000	0.000000	0.600000
75%	2025-02-01 17:15:00	14.350000	15.200000	476.500000	1.000000
max	2025-04-19 23:00:00	24.900000	39.900000	5154.000000	5.200000
std	NaN	7.299909	5.749457	1097.277775	0.641918

Table 1 Input data statistics

To enhance model training, new features were engineered: hour of day, day of week, week of year, and month. Furthermore, a rolling 4-hour average temperature was calculated to account for the thermal inertia of the house (Figure 1).

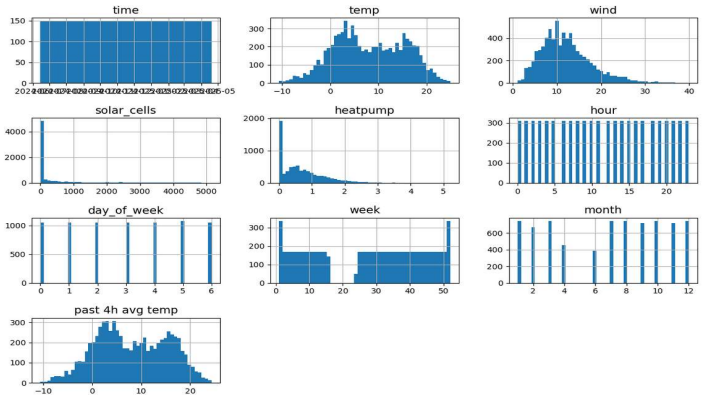

Figure 1 Data frame histograms

DATA ANALYSIS AND TRAINING TEST DATA SPLIT

Feature correlation with heat pump energy consumption was analyzed (Table 2), revealing strong correlations with the 4-hour average temperature and week of year. Scatter plots were utilized to visualize these relationships. The dataset was

split into training and testing sets using stratified sampling based on month to maintain consistent monthly representation in both sets (Table 3)

Feature	Coefficient
past 4h avg temp	-0.524271
week	-0.061857
wind	0.035754
month	0.034573
hour	-0.017804
temp	0.010686
day_of_week	-0.004880
solar_cells	0.001940

Table 2 Feature correlation

Stratified train set

```
month
1     0.100303
10    0.100303
8     0.100303
12    0.100303
3     0.100303
7     0.100303
9     0.097100
11    0.097100
2     0.090695
4     0.061531
6     0.051753
Name: proportion, dtype: float64
```

Stratified test set

```
month
7     0.100404
3     0.100404
10    0.100404
8     0.100404
12    0.100404
1     0.100404
9     0.097035
11    0.097035
2     0.090296
4     0.061321
6     0.051887
Name: proportion, dtype: flo
```

Table 3 Monthly distribution after stratification

BUILDING AND EVALUATION OF ML MODELS

Two models were developed and evaluated:

1. **Linear Regression:** A linear regression pipeline incorporating standardized features (temperature, wind speed, solar energy production, hour of day, day of week, week of year, month, and 4-hour average temperature) achieved an R^2 score of 0.67 on both the training and testing sets (Table 4). Subsequent optimization using radial basis function (RBF) transformation on the temperature feature improved the R^2 score to 0.73 on the test set (Table 5). This early optimization demonstrated the potential of capturing non-linear relationships between temperature and energy consumption.
2. **Random Forest:** A random forest model, trained on the same features (without standardization) achieved a significantly higher R^2 score of 0.81 on the test set (Table 6). This model also demonstrated better performance in predicting peak consumption values, associated with auxiliary electric heating activation during periods of extreme cold.

Residual analysis (Figure 2) and actual vs. predicted plots (Figure 3) confirmed the superior performance of the Random Forest model.

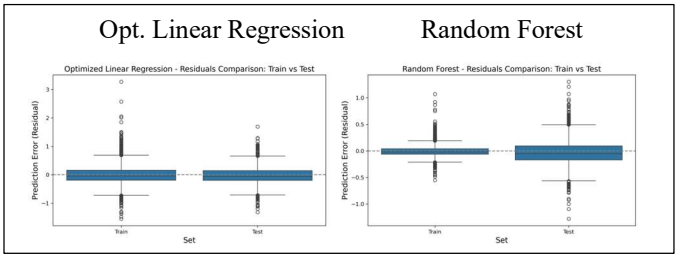

Figure 2 Residuals comparing side by side

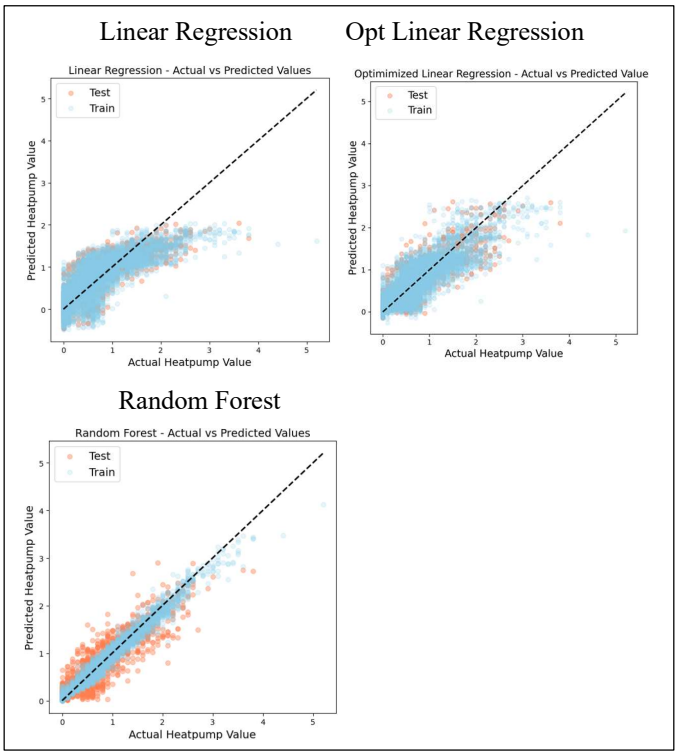

Figure 3 Actual vs predicted values comparing LR, Optimized LR and Random Forest.

Feature importance analysis (Figure 4) indicated the "4-hour average temperature" as the most influential feature for both models.

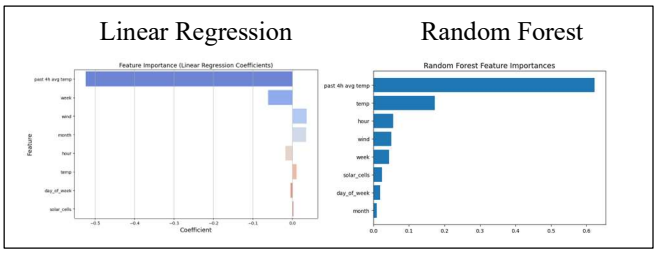

Figure 4 Feature importance comparison

VI. CONCLUSION

This project successfully demonstrated the application of machine learning for predicting heat pump energy consumption using weather and time-based features. The Random Forest model significantly outperformed the Linear Regression model, exhibiting greater accuracy and robustness in capturing complex relationships within the data, including peak consumption events. Early findings in the model development process highlighted the value of addressing non-linear relationships, as demonstrated by the improvement gained through applying an RBF transformation to the temperature feature within the Linear Regression pipeline. This underscores the importance of exploring non-linear modeling techniques for capturing the nuances of heat pump energy consumption dynamics. The superior performance of the Random Forest, however, suggests its inherent ability to handle such complexities without explicit feature transformation.

REFERENCES

[1] Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. O'reilly, 2019

[2] https://www.home-assistant.io – Home assistant open source community

[3] https://monitoring.solaredge.com – site for solarcell production and weather data

[4] https://chatgpt.com/ - GTP-4-Turbo used as external source for inspiration and discussion of alternatives

[5] https://aistudio.google.com/prompts/ - Gemini 1.5 Pro used as external source to improve text readability and grammar.