Exploring Key Characteristics Influencing Hybrid Heat Pump Performances in the *Demoproject Hybride* through Regression Modelling

MSc Thesis Applied Data Science

Author

J. D. Beunk (5999448)

j.d.beunk@students.uu.nl

First Supervisor

Laurens Stoop

Second Supervisor

Ad Feelders

Submitted in fulfilment of the requirements for
the degree of Master of Science

Department of Information and Computing Sciences
Utrecht University
Netherlands
07/07/2023

Preface

This master's thesis was undertaken as a part of the Applied Data Science program at Utrecht University. Two teams of master students collaborated with Inversable BV and Intergas Verwarming BV, which provided the teams with the dataset collected for the *Demoproject Hybride*. This project, initiated in November 2021, aims to gain valuable insights into the practical performance, savings, and applicability of hybrid heat pumps.

Two distinct research questions were analysed using the *Demoproject Hybride* dataset. The first research question, explored by Jordi Beunk, Johanna Lems, and Abdulhakim Özcan, examined the differences in hybrid heat pump performances between houses. The second group, consisting of Sahar Pourahmad and Ruben Groot, focused on uncovering human patterns in the data.

The cleaning and preparation of the dataset for analysis was a collaborative endeavour. Subsequently, individual students conducted their respective analyses based on their research question and chosen methodology. The introduction and data sections were written together.

This thesis documents the outcomes of our collective effort, shedding light on the factors influencing saving performance scores and revealing valuable trends in the time series data. By delving deeper into these research questions, we aim to contribute to the existing knowledge of real-world performance of hybrid heat pumps.

Jordi Beunk, Johanna Lems, Abdulhakim Özcan, Sahar Pourahmad, and Ruben Groot collectively dedicated their time, skills, and efforts to this study, and we are proud to present the findings of our research in this master's thesis.

Table of Contents

1. Introduction

The transition to sustainable energy resources is a pressing imperative due to the impacts of carbon emissions on the environment. Consequently, various countries, including the Netherlands, are proactively exploring alternative energy sources and technologies to reduce their carbon footprint and mitigate climate change. The Dutch government, in line with its Climate Act and the National Climate Agreement, has set ambitious goals to reduce carbon emissions by 49% before 2030 and by 95% before 2050 (Government of the Netherlands, 2019, p. 191)

In particular, the built environment plays a significant role in these emission levels, with residential areas alone accounting for 12.5% of the country's total carbon emissions resulting from operational energy consumption (Dutch Green Building Council, 2021, p. 9). To reduce carbon emissions from buildings, the Dutch Government has taken several measures, including mandating the adoption of (hybrid) heat pumps starting in 2026 as highlighted in a parliamentary letter (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2023).

The Dutch Heating Industry has estimated that installing 1.7 million hybrid heating systems can help achieve the Netherlands' carbon reduction targets for the built environment by 2030 (De Nederlandse Verwarmingsindustrie, 2021). In line with these projections, the Dutch Government launched the *Demonstratie project Hybride warmtepompen* (i.e. *Demoproject Hybride*) to assess and promote the use of hybrid heat pumps in residential buildings (Demonstratieproject Hybride Warmtepompen, n.d.).

This is a large-scale undertaking that collects real-time data on the performance of hybrid heat pumps in residential buildings throughout the Netherlands. The primary objective of the *Demoproject Hybride* is to evaluate the performance of hybrid heat pumps under real-world conditions.

There are different types of heat pumps: air-to-air, air-to-water, water-to-water and ground-to-water heat pumps. They differ in terms of the thermal energy source (air, water or geothermal) or the medium that is used to transfer heat (air or water) (Wolf, 2023). A hybrid heat pump is a heating system that makes use of an electric air-to-water heat pump, combined with a gas condensing boiler.

A schematic of an air-to-water heat pump is depicted in Figure 1. The heat pump consists of two units, the inside and outside coil, that are connected by two semi-flexible copper pipes. The system contains a refrigerant with a very low boiling point (e.g. R410A: -48.5 °C), which is used to transfer thermal energy from the outside air to the water in the heating system of the house.

The steps to achieve this are shown in Figure 1. These include:

1. The refrigerant heats up in the outside coil and evaporates; the refrigerant is pumped through the outside coil. Since it's colder than the outside temperature, heat is extracted from the outside air and the refrigerant warms up. The temperature is still relatively low, but high enough to evaporate the refrigerant.
2. The evaporated refrigerant is heated up by compression; the refrigerant vapour goes into the compressor component. Due to compression, the temperature of the vapour increases considerably.
3. The evaporated refrigerant heats up water in the inside coil and condensates; the high temperature refrigerant vapour is transferred to the inside coil, where it loses its heat to the water that is used to heat up the house, due to the temperature gradient. Because of this cooling, condensation occurs.
4. The refrigerant cools by expansion; the high temperature liquid is transferred to the expansion device. Here the pressure is decreased allowing the liquid to expand, causing the temperature to

drop. The low temperature refrigerant is transferred to the outside coil and the process repeats itself.

Generally, the heat pump is used to heat up the house. A boiler assists when the demand is high or when the performance of the heat pump is low. Both the demand and the heat pump performance are affected by outside temperatures. The amount of thermal energy in the air decreases with lower temperatures, which results in less efficient use of the heat pump. Additionally, when the outside temperature drops below 0 °C, ice formation on elements of the outside unit can decrease the performance of the heat pump. Outside temperature is one of many factors that affects the performance of hybrid heat pumps.

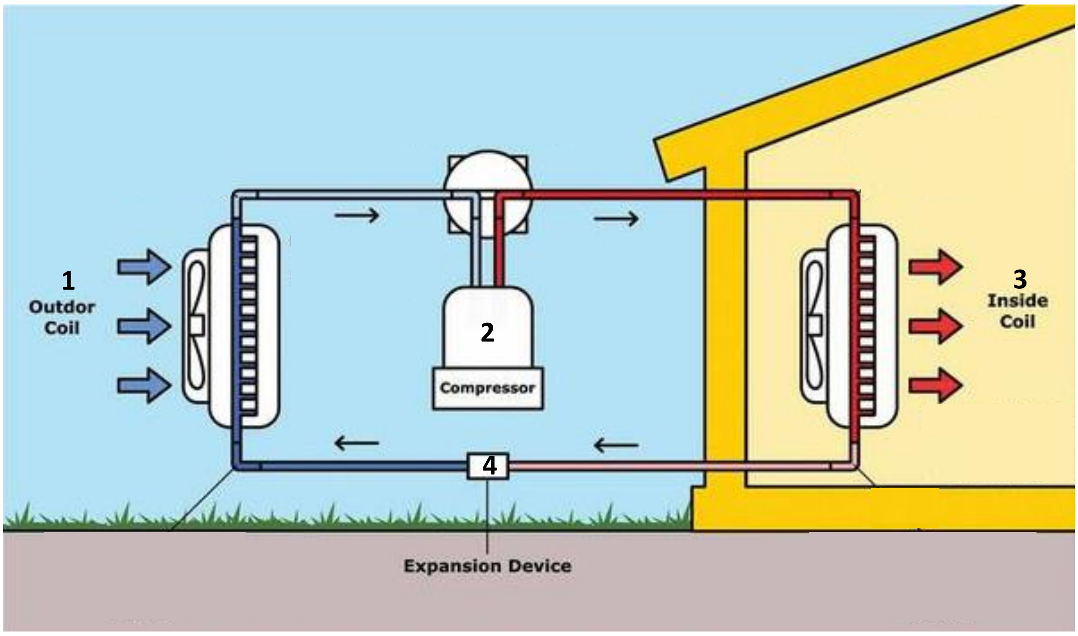


Figure 1: Schematic of an air-to-water heat pump. It depicts the outside coil (step 1) and inside coil (step 3), connected by copper pipes. Additionally, the compressor (step 2) and expansion device (step 4) are shown. The blue and red colours indicate the temperature of the refrigerant in the copper pipes. Source: https://www.energy.gov/energysaver/air-source-heat-pumps

There are several ways to measure the performance of a heat pump. One commonly used indicator is the coefficient of performance (COP), which quantifies the ratio between the energy produced by a heat pump to the energy required to operate it. It provides a measure of a heat pump's efficiency in converting electrical energy into thermal energy. However, the COP is typically determined under specific operating conditions, disregarding seasonal fluctuations or changes in operating conditions.

A more accurate representation of a heat pump's efficiency in real-world scenarios is the seasonal coefficient of performance (SCOP). It represents the average COP throughout a heating season, weighted by the operating hours at different outdoor temperatures. Therefore, it considers the performance under different operating conditions. One shortcoming of the SCOP is that it only considers the energy produced by the heat pump and not how effectively this energy is used to heat up a house. In order to measure the performance of a heat pump in a more holistic way, the saving performance score is used in this research.

The saving performance score is the ratio between the amount of energy saved and the amount of energy consumed by the heat pump. It compares the savings in gas with the increase in electricity usage due to the heat pump.

To compute the saving performance score (SP), first the savings in gas consumption are calculated by subtracting the current yearly gas consumption (*gas*$_{now}$) with the yearly gas consumption prior to the installation of a heat pump (*gas*$_{prior}$). For the latter, instead of taking the gas consumption from a single year, the mean of two separate historical years is used as a representative value for the gas consumption before the installation. The savings in gas is converted from gas use (in m^3) to energy use (in kWh) by multiplying it with a conversion factor of 9.8. Finally, the gas savings, expressed in terms of energy (kWh), are divided by the electricity consumption of the heat pump (*electricity*$_{hp}$) to end up with the saving performance score. This is shown in Equation 1.

$$\mathrm{SP} = \frac{9.8 * (\mathrm{gas}_{\mathrm{historical}} - \mathrm{gas}_{\mathrm{now}})}{\mathrm{electricity}_{\mathrm{hp}}} \tag{1}$$

There is a scarcity of studies examining the real-world performance of hybrid heat pumps, particularly in relation to Dutch households and the local climate. However, a notable study conducted in the Netherlands is the *Installatiemonitor* project.

The *Installatiemonitor* project looked into the relationship between house characteristics and energy usage of heat pumps in the Netherlands (Installatiemonitor, 2022). They found that release temperature of the heating system was positively correlated with energy usage of the heat pump. This is related to the type of heating system (i.e. low-temperature vs high-temperature). Similarly, the surface of energy loss, which encompasses a combination of the house type and surface of the house, was found to be positively correlated. Conversely, no consistent relationships were found regarding energy label, year of installation and type of insulation. Furthermore, the research showed that heat pumps fulfil around 60% of the annual heat demand, with a substantial contribution occurring primarily at outdoor temperatures above 0 °C. On average, the SCOP for the households considered in the *Installatiemonitor* project was determined to be 3.8. However, variations between households are expected, since the heat pump performance relies on a multitude of factors ranging from heat pump type to behavioural patterns of the residents (Miara and Krame, 2011; Oikonomou et al., 2022).

Despite providing valuable insights into the real-world performance of heat pumps in the Netherlands, the *Installatiemonitor* does not consider variations in the heat pump performance among households and the specific characteristics that contribute to these differences. Understanding the factors influencing the divergent performances of heat pumps between households is crucial for gaining a comprehensive understanding of their overall efficiency and effectiveness.

This study aims to identify key characteristics that account for the substantial variation in the real-world saving performance scores among the houses in the "Demoproject Hybride". The primary objective of this study is to find data-based answers rather than relying solely on model-based answers. Three distinct machine learning techniques will be employed: classification, clustering and regression. Specifically, this study will focus on the outcomes derived from regression techniques.

This study was carried out in collaboration with Inversable BV and Intergas BV and examines the real-world performance of hybrid heat pumps in a subset of houses from the "Demoproject Hybride". As the adoption of hybrid heat pumps as a sustainable heating solution is a relatively recent development, there is a lack of in-depth studies on the performance of these systems in real-world settings.

The rest of this thesis is structured as follows. In Chapter 2 we describe the dataset and outline the steps taken to prepare the data for our analysis. The methods that are used to find the key characteristics that determine the saving performance score are discussed in Chapter 3. In Chapter 4 the results of the model selection are shown and we describe the performance and most important features of the best regression model. A reflection on the data and methods is discussed in Chapter 5. The conclusion of the report and an answer to the research question can be found in Chapter 6.

2. Data

The data used in this thesis was obtained from the *Demoproject Hybride*, which includes fully anonymised house metadata (169 houses), house time series data and weather time series data (both 168 houses). For 167 participating houses all three datasets where available, only these houses are considered here.

The house metadata encompasses crucial house characteristics such as the type of house, its construction year, the level of insulation, and the number of inhabitants. Whereas the house time series data offers real-time measurements on the boiler, heat pump, heating system, as well as the overall gas and electricity usage of each house. The weather time series data was extracted specifically for each house based on the nearest weather station.

These three datasets are described in Section 2.1, followed by Section 2.2 which goes into the preprocessing of the datasets.

2.1 Description of the Data

2.1.1 The house metadata

The house metadata contains information on various characteristics of the houses, as summarized in Table 1. The metadata was provided by the homeowners through a survey and is therefore susceptible to human errors. Additionally, it is worth considering that certain house characteristics, such as the number of inhabitants, may have changed since the time of the survey.

To protect the privacy and confidentiality of the homeowners' information, the provided dataset has undergone anonymization techniques before it was made available for this thesis. As a result, specific details such as the exact build year and area of the houses were transformed into broader classes. Furthermore, the geographical information pertaining to the houses was restricted to the level of municipality only. Finally, data associated with the manufacturers of the heat pumps was excluded from the dataset provided for this analysis as they did not consent to the use of their data in a 'product comparison analysis'.

Table 1: Overview of the house metadata, including the variable names, descriptions, and types of data used.

Variable name	**Description**	**Type**
Id	Anonymized ID for each house	Nominal
Type	Type of house	Nominal
Municipality	Municipality in which the house is located	Nominal
Energy label	Energy label of the house	Ordinal
Area class	Categorized area of the house for anonymization purposes	Ordinal
Build year class	Categorized build year of the house for anonymization purposes	Ordinal
Number of inhabitants	Number of people living in the house	Continuous
Historic annual gas consumption	The average annual gas consumption of two years prior to the installation of the hybrid heat pump, in m^3	Continuous
Historic annual electricity consumption	The average annual electricity consumption of two years prior to the installation of the hybrid heat pump, in kWh	Continuous
Roof insulation	Whether the roof is insulated	Nominal
Wall insulation	Whether the walls are insulated	Nominal
Floor insulation	Whether the floors are insulated	Nominal
Double glass	Whether the window frames have been fitted with double glazing	Nominal
Additional heating	Whether a house had any additional heating systems, such as a gas fireplace	Nominal
Average days measured	Number of days with measurements of the time series data	Continuous
Saving performance score	Saving performance score	Continuous

The majority of the houses are detached (59%), followed by 26% being semi-detached and 14% being terraced. There is only one apartment among the houses.

Regarding the age and energy efficiency of the houses, a significant majority (85%) of them have energy labels ranging from A-C, indicating a relatively favourable energy performance. Around 38% of houses were built before 1975; in this period, insulation was not yet mandatory during construction. The houses built between 1975-1999 (40% of them) were moderately insulated during the building process. Those houses built after 2000 (22%) were constructed with sufficient insulation measures in place (Kuijeren, 2021). However, it is worth mentioning that many older homes have undergone renovations to incorporate insulation. The inhabitants provided information regarding the insulation status of their walls, floors, ceilings, and windows. This additional data helps to assess the overall energy efficiency of the houses beyond the initial construction standards, and is further discussed in Section 2.2.1.1.

On average, the houses in the dataset consisted of 3.0 inhabitants. This is higher than the average household size reported by the Central Bureau for Statistics (CBS) in 2022, which was 2.13 people (CBS, 2022). The households in the metadata set comprised 48% singles and couples (1-2 individuals), 40% small families (3-4 individuals), and 12% large families (5 or more individuals).

Regarding the liveable space of the houses, a small proportion (5%) has an area below 100 m^2, while 25% of the houses have an area ranging from 100 to 150 m^2. The majority of the houses have an area between 150 and 250 m^2 (56%), the final 14% of houses have an area greater than 250 m^2. It should be noted that, on average, the houses have a large living space of 72 m^2, compared to a normal Dutch house of 53 m^2 (CBS, 2022).

With respect to the gas and electricity consumption prior to installing the hybrid heat pump, the average gas consumption across the dataset was approximately 1890 m^3 per year. The house with the highest gas usage, consumed an average of 4775 m^3 per year before the installation of a hybrid heat pump. In contrast, the lowest-consuming house had an average gas consumption of only 452 m^3 per year.

The average electricity consumption prior to the installation of the hybrid heat pump was 2468 kWh per year. It is worth noting that the accuracy of electricity consumption may be affected by the fact that some participants reported their total electricity consumption, while others reported their net consumption (subtracting electricity consumption and electricity production from solar panels). The house with the highest electricity consumption consumed an average of 10098 kWh per year before installation of the heat pump. Conversely, the lowest consuming house had an average negative value of -4054 kWh per year, indicating that the homeowners produced more energy with their solar panels than they consumed.

According to the CBS, the average gas consumption per house was 1239 m^3 per year, and the average electricity consumption was 2741 kWh per year in 2019 (Centraal Bureau voor de Statistiek, 2021). The high gas and electricity consumption in our dataset is likely due to the larger house sizes and higher number of inhabitants.

The average saving performance score in the dataset was 5.3. The scores ranged from 1.6 to 13.6. Figure 2 provides an overview of the saving performance score distribution.

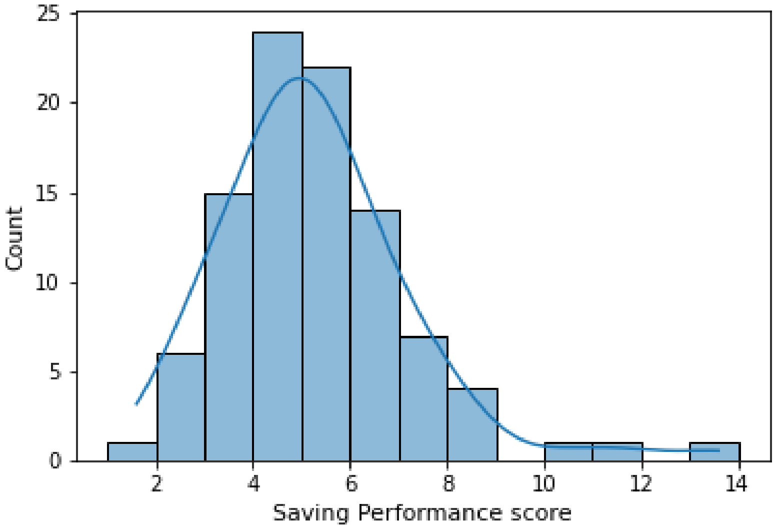


Figure 2: The distribution of the saving performance scores in the Demoproject Hybride.

The houses included in our dataset may not be entirely representative of the majority of houses in the Netherlands. The houses in our dataset are larger than the average house, accommodate more people, and exhibit higher energy and gas usage. Consequently, caution must be exercised when generalizing the findings to the broader population of households in the Netherlands.

2.1.2 The house time series data

The house time-series data consists of real-time measurements. On average, there are 7 months of measured data per house, ranging from 20 days for the shortest duration to 15 months for the longest.

During this period, various sensors were deployed in the houses, including a Smart meter and a thermohygrometer, and sensors at the heating systems, heat pumps and boilers. A summary of these sensors is provided in Table A1 in Appendix A.

Heating system sensor
The heating system sensor is responsible for monitoring the heating system within the house. It measures various parameters such as flow rate and the supply and return temperature within the heating system. The energy and power of the thermal system are derived from this flow rate and supply temperature. Measurements are collected every 5 seconds, providing detailed insights into the heating system's performance.

Heat pump sensor
The heat pump sensor is positioned at the heat pump unit. It specifically measures the energy and power supplied to the heat pump. However, it does not directly measure the amount of energy converted by the heat pump and transferred to the heating system.

Boiler sensor
The boiler sensor monitors how much energy and electricity the boiler uses. It should be noted that a small portion of this energy could be used to heat water for showers or other purposes besides heating systems.

Smart meter
The Dutch Smart Meter Requirements (DSMR) or smart meter is responsible for reading the meter readings of the house. It transmits a "telegram" every 10 seconds for power readings and every 60 seconds for energy and gas readings. This enables real-time monitoring and analysis of the energy consumption patterns.

Please note that the power consumed and delivered values represent the combined total for all three phases of the network. Similarly, the energy consumed and delivered values account for both high and low tariffs. Energy can be delivered by a house when it generates energy itself, such as through the use of solar panels.

Thermohygrometer
The thermohygrometer sensor is typically placed in the living room and is used to measure the temperature and humidity of a room. It provides valuable data for understanding the indoor climate conditions within the house.

2.1.3 The weather time series data

For each house, a number of relevant weather conditions where provided by the K*oninklij*k *Nederlands Meteorologisch Instituut* (KNMI) weather stations. The weather data was extracted from the nearest weather station based on the location of each house, ensuring that each house has weather information specific to its municipality. Temperature, rainfall, relative humidity, wind speed, sun hours and ice formation were collected. Table A2 in Appendix A provides an overview of the collected weather data.

2.2 Preprocessing of the data

The metadata and the house and weather time series are utilized to determine the variables that exert the most significant influence on the saving performance score.

Before conducting the analysis, various data preparation steps were undertaken. The house metadata (Section 2.2.1) and the time series datasets (Section 2.2.2) were processed independently, and then merged into a final dataset (Section 2.2.3). Detailed explanations of the preprocessing steps are provided in their respective sections.

2.2.1 Preprocessing the house metadata

The preprocessing of the house metadata was done in four steps. The application of this cleaning and pre-processing procedure reduced the dataset to 86 houses.

First, multiple metadata files were merged to create a comprehensive dataset comprising information on the 167 houses used. As part of this merging process, some data was combined to represent the average annual gas and energy usage, prior to the installation of the hybrid heat pumps. Two different metadata files were used to acquire this information: one obtained from the homeowners' survey and the other generated by manually reviewing participants' utility bills. The manually reviewed file was deemed to be more reliable, although it had more missing data compared to the survey-derived file. In instances where the manual file had missing data while the survey file did not, the survey data was used.

In the second step, several data cleaning operations were conducted, including the removal of unwanted characters and translating non-numerical Dutch content into English.

The identification of outliers was done in the third step. One house was identified with an outlier value for the household size (9 people), and was removed from the dataset to prevent any potential skewing of the analysis. In addition, to maintain the focus and reliability of the research question, houses without a saving performance score (82 houses) and those with unrealistic saving performance scores (SP > 9.0, 3 houses) were excluded from further analysis.

In the fourth step, the reliability of the saving performance scores was carefully examined. Preliminary analysis revealed a strong correlation between the duration of data collection and the saving performance score. It was observed that houses with a shorter time series of data tended to have higher saving performance scores. For houses with less than a year of available time series data, certain assumptions had to be made when calculating the saving performance scores. However, these assumptions led to an overestimation of the saving performance scores, particularly when the available data spanned less than 100 days. Consequently, six houses with less than 100 days of data were excluded from further analysis to ensure the reliability of the results.

2.2.1.1 Definition of additional classes in the metadata

To enhance the discriminative capability, certain variables are transformed into categorical form. These include the insulation count, number of inhabitants and energy labels, as detailed below.

Insulation count

The insulation count for houses is calculated based on the presence of different types of insulation such as roof insulation, wall insulation, floor insulation, and double glazing. Like all metadata, this information is obtained from the homeowner's survey. If a homeowner is unsure or unaware of the presence of a particular type of insulation, that specific insulation is marked as NaN. To determine the insulation count, we count the number of insulation types for each house. For instance, if the house had roof, wall and floor insulations but lacked double glazing, the insulation count of this house would be set to 3. To make the calculation more accurate, we leveraged an article from VK Makelaars, a rental agency in the Netherlands (van Kuijeren, 2022). This article serves as a valuable resource, allowing us to make assumptions regarding insulation counts based on the year of construction for each house. By incorporating such informed assumptions, we enhance our ability to estimate the insulation counts for houses accurately.

Number of inhabitants

Due to the large range in number of inhabitants, classes are created for this variable. The following three classes are made: singles and couples (1-2 individuals), small families (3-4 individuals) and large families (5 or more individuals).

Energy Labels

The energy labels were divided into two distinct classes. The first class comprises energy labels D-G, which are considered poor. The second class includes energy labels A-C, which are generally considered average or good.

Missing data in the number of inhabitants and energy label were imputed using logistic regression based on features they had a high correlations with.

2.2.2 Preprocessing of the house and weather time series data

The house and weather time series data were extracted and aggregated per day. The daily aggregates were created for two reasons. Firstly, aggregating the data for a specific period prevented that it was updated in real-time as the *Demoproject Hybride* is ongoing. Among other things, this ensured that the dataset was kept constant throughout the research project. Secondly, aggregation by day reduces the computation time considerably, while keeping the right level of detail for our analysis.

The weather time series data was preprocessed by the KNMI and was available for each house at an hourly resolution. The weather time series dataset was aggregated using appropriate methods. Instantaneous variables were aggregated by calculating the daily mean value. Variables such as sun hours and rain were computed as cumulative values per day. Additionally, for outdoor temperature, the minimum, maximum, and mean values were collected

As for the house time series dataset, specific aggregation approaches were applied. Instantaneous variables, such as power and temperature, were aggregated by calculating the daily mean value. On the other hand, cumulative variables like energy and gas consumption were aggregated by selecting the maximum value per day.

The house time series data contained a significant amount of missing data (12%). For cumulative variables, the missing data was imputed using linear interpolation for a maximum of 3 consecutive days. This limit was set to assure the reliability of the interpolated values. This resulted in negative cumulative values for the boiler energy of one house (017QC6A9), likely because a new sensor was installed during the period when data was missing. For this house, missing values were not imputed.

Several outliers were observed in the time series data, as an extreme drop in one or more cumulative variables. These drops were related to the instalment of a new energy meter (DSMR) in the house, which caused the cumulative value to reset to 0. Additionally, one house (wl7RZ9xP) contained a small drop in energy delivered and consumed for a specific day and another contained extreme peaks in gas usage (pF6hF6FW). The reasons for the outliers of these devices are unknown. The instances where outliers occurred were removed from the timeseries. Hereafter, all cumulative variables were differenced to obtain the day-to-day change in the measurement.

Finally, two new variables were computed based on time series data. Firstly, the difference between the indoor and outdoor temperature is included, as it provides additional information on the ability of the house to retain heat. This combines several metadata variables like insulation count, energy label and area class, while overcoming the problem of inaccuracies due to human error. Secondly, the difference between the supply and return temperature of the heating system is computed. This provides insight into the capability of the heating system to lose heat to the house, and is thus a measure of the efficiency of the heating system. This value is normalized by dividing it with the mean of the supply and return temperature, to differentiate between heating by the boiler and heating by the heat pump.

2.2.2.1 Binning of time series based on outdoor temperature

Comparison of households in terms of time series data was obstructed by the fact that not many houses had continuous time series for the same time period. In order to solve this problem, the data was binned on outdoor temperature. This allowed us to make optimal use of the available data, even when data was missing for certain time periods.

The outdoor temperature was split into 5 bins: <0, 0-5, 5-10, 10-15 and ≥15 °C. This is motivated by the fact that the same temperature ranges are used in the saving performance calculation. For the calculation of the saving performance, the gas usage of the last 12 months is used (See Equation 1). Since gas consumption contained missing data, the data is binned on outdoor temperature. The average gas usage for each temperature range is multiplied by the number of days this temperature range occurred to get an estimation of the gas consumption for the last year.

For consistency, the same temperature ranges are used to bin the time series data. Note that we did not estimate missing values based on the number of days temperatures within a range occurred, but only used days with data available. The mean of the time series variables was computed for every house and each bin, which allowed us to compare the time series data of different household within certain temperature ranges with the saving performance score. This technique significantly increased the number of features available for analysis.

After applying the binning, all houses had data for all the bins, apart from four houses who did not have data for the temperatures higher than 15 degrees. However not all the bins had the same amount of data measured. As depicted in Table 2, there are only a few colder days in the analysed period. This is partly due to the climate in the Netherlands and partly due to the relatively warm winter of 2022-2023.

Table 2: Overview of data availability per bin.

Bin (˚C)	Amount of data (%)
<0	4
0-5	19
5-10	34
10-15	25
≥15	18

2.2.3 Combining metadata and time series data

After all the preprocessing steps described above, the metadata and time series data are merged into one dataset. Since the methods used in this study do not allow for the presence of missing data, all houses that contain missing data in any of the columns are removed. This resulted in a final dataset that consists of 61 houses and 151 features.

3. Methods

In this Chapter, the methods used to find the key characteristics that account for the variations in the saving performance scores among houses in the *Demoproject Hybride* are discussed.

This thesis will focus on the use of regression models to find an answer to the research question. The motivation for using regression models is the fact that the saving performance is a continuous prediction variable. However, clustering and classification models are discussed in the theses of the other group members.

All parts discussed in this Chapter have the combined goal of obtaining a regression model that is best able to predict the saving performance score. However, it is important to note that we are not necessarily interested in the prediction of the saving performance score. By finding the best performing regression model, we aim to find the most important features that are used to predict the saving performance score. The performance of the model is only used to provide an indication of the reliability of the most important features. Thus, the question remains: how do we obtain a regression model that is able to predict the saving performance score most accurately?

First, we elaborate on the feature creation and selection methods that are used (Section 3.1). Using the most important features, several regression models are trained and their performance assessed. This allows us to select *Elastic Net regression* as the best performing model (Section 3.2). Hereafter, the hyperparameters of this model are tuned with the goal of improving the model performance (Section 3.3). Finally, the methods used to obtain the most important features are discussed (Section 3.4).

3.1 Feature creation and selection

The features that are directly used to calculate the saving performance score are removed, since their effect on the score is already known (shown in Equation 1). These features include: historical gas usage, real-time gas usage and both heat pump power and energy.

Furthermore, multicollinearity is observed in the dataset, which means that the independent variables are highly correlated (see Appendix B). Since most regression models are not able to deal with multicollinearity, manual feature removal of highly correlated variables is performed based on their correlation with the saving performance score; only the variables that are most strongly correlated with the saving performance score are selected. For temperature, the minimum indoor and minimum outdoor temperature, and the difference between the indoor and outdoor temperature are included. Both the supply and return temperature of the heating system are removed, as their influence is captured by the normalized difference between the two. Finally, the boiler power is removed, since it contains similar information as boiler energy, and boiler power has more missing data. After removal of the variables described above, there were 102 features.

The data is standardized to ensure that all variables are on the same scale of magnitude. Two standardization techniques are examined, one based on *z-score standardization* and one based on *min-max standardization*. Both techniques showed similar performances, but *min-max standardization* is used in further analysis. This is motivated by the fact that this method ensures that the ranges of all variables are equal, which increases the interpretability of the regression coefficients of our models.

Additionally, cross-terms are included to analyse the combined effect of two variables on the saving performance score. It allows the models to uncover potential complex interactions between predictors that may not be evident when considering them individually. This is done by taking the second degree polynomials of the variables, and only selecting interaction terms. An example of the second degree polynomials of two variables, x1 and x2, is shown in Equation 2. Their interaction term is shown in red.

$$y = x1 + x2 + x1^2 + x2^2 + \boldsymbol{x1 * x2} \tag{2}$$

It is important to note that only cross-terms of variables with the same temperature bin are included. This is motivated by the fact that different temperature bins contain data from different days, and we are only interested in the combined effect of variables that occur simultaneously.

Having included cross-terms to our dataset, feature selection is performed. The motivation for this is twofold:

1) Only using a subset of the available variables reduces the risk of overfitting, which generalizes to model to unseen data and therefore improves its performance.
2) By selecting relevant variables, we simplify the model and increase the interpretability.

Since interactions are included by adding the cross-terms, a univariate feature selection method is used. This consists of a linear regression model that tests the effect of each variable on the saving performance score by using the F-statistic. It adds the variable with the largest effect on the saving performance score to a list. This process is repeated until the list contains a predefined number of variables (*k*).

Initial runs showed that the model performances were sensitive to the value of *k*. Therefore, each model is trained using multiple *k*-values, ranging from 1-30 features. There are two reasons for selection 30 features as the upper limit. Firstly, a relatively low number of features improves the interpretability of the model results. Secondly, it reduces the computational intensity of the model runs.

3.2 Model performance metrics and selection procedure

After finding the most relevant features to use in our analysis, 6 different regression models are employed. These included: *Linear, Lasso, Ridge, Elastic Net, Decision Tree* and *Random Forest regression*. Three performance metrics are used to find the best performing model: the coefficient of determination (R^2), root mean squared error (RMSE) and the Bayesian Information Criterion (BIC). Their use in measuring model performance is as follows:

R² measures how well the model fits the data. It represents the proportion of variance in the dependent variable that is explained by the independent variables.

RMSE measures the average distance between model predictions and the actual value. Thus, it is a measure of the magnitude of the error between the predicted and actual values.

BIC also measures how well the model fits the data, but adds a penalty for model complexity. This penalty is based on the amount of features that are used to predict the dependent variable.

The BIC metric did not provide any insight, since the penalty for adding more features to the model was too strict. This is likely the result of the large amount of features, due to the binning on outside temperature and adding cross-terms. Therefore, only the R^2 and RMSE are used to select the best model.

For every model, the training process contained the following steps:

1. Split the data into a training and test set

2. Standardize the independent variables
3. Train the model and predict the saving performance scores of the test set
4. Save and plot the R^2 and RMSE scores of the training and test data

To determine the optimal size of the test set, several runs are performed using the commonly used values 20%, 25% and 30%. These showed that a size of 30% resulted in the best model performances, which is therefore used in the analysis.

As mentioned above, each model is trained using several *k*-values, ranging from 1 to 30 features. Thus, steps 1 to 4 are repeated 30 times, for each value of *k*. Since the model performance was very sensitive to the exact split in training and test data, 50 iterations for each *k*-value are performed, all with a different combination of training and test data. For model comparison, the mean value of R^2 and RMSE over these 50 iterations are used. Additionally, the standard deviation is computed to assess the sensitivity of the models.

3.3 Hyperparameter tuning

Using the methods described above, *Elastic Net regression* was found to be the best performing model. This is a linear model that combines two regularization techniques (*L1* and *L2*). The hyperparameters of this model are tuned to improve the performance.

A grid search is used to find the best combination of hyperparameter values. It trains the model using all possible combinations of specified parameters using cross-validation, and finds the combination with the highest performance. The hyperparameter include the *alpha, L1-ratio, maximum iterations* and *selection* parameter. The *alpha* parameter determines the strength of *L1* and *L2* regularization, whereas the *L1-ratio* controls the mixing proportion between the two. The *maximum iterations* defines the upper limit on the number of iterations that the model will perform and the *selection* parameter determines the method used to update the coefficients in every iteration. The parameter ranges used in the grid search can be observed in Appendix C. The model accuracy is used as the scoring function, which means that the highest performance is defined by the highest R^2-value.

In addition to the parameter grid and the scoring function, the number of folds in the cross-validation step of the grid search were defined. This parameter determines into how many subsets the training data is split. The model is then trained and tested on every combination of these subsets. In choosing the amount of folds, it is important to keep a balance between the amount of samples in the test set and the amount of splits. Since we have a relatively small amount of samples in our dataset, a relatively small number of 5 folds is used. This decision is based on the theory discussed above, and substantiated by several trial runs with different numbers of folds.

Since the model performance is sensitive to the exact split in training and test data, 50 iterations are used for the grid search. The combination of hyperparameters that occurs most frequently in these 50 iterations is chosen as the optimal model setting.

3.4 Feature importance

The tuned *Elastic Net* model is trained on different subsets of variables. The steps outlined in Section 3.3 are applied, again using 50 iterations with varying training and test set combinations.

The following subsets are used to train the best performing model:

1. The prepared dataset including cross-terms
2. The prepared dataset excluding cross-terms
3. The prepared dataset excluding cross-terms and weather data

The motivation for excluding cross-terms in these subsets is the increased interpretability of the most important features. Additionally, this allows us to assess the effect of cross-terms on the model performance. One subset excludes weather data in addition to this, since participants have no control over weather conditions that affect the performance of their hybrid heat pump. This makes these variables less relevant in optimizing an installed heat pump.

As mentioned above, the performance of the model varies with different combinations of training and test splits. However, we are only interested in the most important features in model runs that perform well. To achieve this, we stored the features and regression coefficients of the best performing model runs. For every run of *k*, ranging from 1 to 100, and every of the 50 iterations for each *k* values, we store the features and regression coefficients if the run has an R^2-value above 0.6. This threshold is chosen such that there is a balance between the performance level of 'good' runs and the amount of runs that reach this threshold.

To find the features that occur most often in these 'good' runs, features with a regression coefficient of 0 are excluded, since they are not relevant to the model prediction. A higher threshold is not preferred, since the data comprises of continuous and categorical features with different distributions, which affects the regression coefficients.

Furthermore, the house ids that are in the test set of 'good' model runs are obtained. This enabled us to compare the mean values of important features of these house ids with the other house ids. However, this method did not provide any insights and the results are not included in this report.

4. Results

In this Chapter, the relevant findings are shown. Firstly, the performances of the 6 different regression models are described (Section 4.1). This section includes a likely explanation on why *Elastic Net* is this best performing model. The results of hyperparameter tuning of the *Elastic Net* model are shortly discussed (Section 4.2). Then, the performance of the tuned *Elastic Net* model is shown (Section 4.3). The most important features using the total dataset (Section 4.4) and using the data subsets are given (Section 4.5). This includes an interpretation of their effect on the saving performance score. Finally, the results are compared with group members who used clustering and classification methods (Section 4.6).

4.1 Model selection

Minor hyperparameter tuning resulted in the following 6 regression models:

- *Linear regression*
- *Lasso regression* (α = 0.1)
- *Ridge regression* (α = 0.1)
- *Elastic Net regression* (α = 0.1)
- *Decision Tree regression* (*max_depth* = 4, *min_samples_leaf* = 4)
- *Random Forest regression* (*max_depth* = 4, *min_samples_leaf* = 4)

The highest performing *k* and the mean value (50 iterations) of the performance metrics (R^2 and RMSE) is shown in Table 3. A detailed plot of the performances over all values of *k* can be found in Appendix D.

Table 3 shows that *Elastic Net regression* has the highest performance, with a mean R^2 of 0.29 and a mean RMSE of 1.21. Note that the standard deviations of the R^2-values are quite substantial, but lowest for this specific model. This indicates that the *Elastic Net* model is least sensitive to variations in the train-test split.

Additionally, the mean R^2 and mean RMSE vales of the training set are shown for each model. These indicate the amount of overfitting. All models show some degree of overfitting, but this is most dominant in the *Decision Tree* and *Random Forest* models. This substantiates our hypothesis that the model complexity of these models is likely one of the reasons for their lower performance.

Table 3: Highest mean R^2 and RMSE value for every regression model. These value are shown for the training and test set. Additionally, the standard deviations are shown for the performance metrics of the test set.

Model	**Test R^2**	**Train R^2**	**Test RMSE**	**Train RMSE**	***k***
Linear	0.17 (0.26)	0.55	1.32 (0.18)	1.12	9
Lasso	0.13 (0.14)	0.32	1.34 (0.19)	1.29	26
Ridge	0.26 (0.18)	0.54	1.24 (0.19)	1.06	13
Elastic net	**0.29** (0.13)	0.42	**1.21** (0.2)	1.12	21
Decision Tree	-0.01 (0.29)	0.49	1.53 (0.17)	1.05	1
Random Forest	0.2 (0.18)	0.74	1.3 (0.18)	0.85	25

To examine why Elastic Net regression is the best performing model, it is imperative to consider the size of our dataset. The prepared dataset contains a high number of 1812 features (including cross-terms) compared to the 61 houses. Note that only the 30 most important features are used for model selection, which is still substantial compared to the amount of samples. For a model to handle this type of high-dimensionality, mitigation of overfitting is required. This can be achieved by a reduced model complexity or the inclusion of a good regularization technique in the model.

The model complexity of *Decision Tree* and *Random Forest regression* is likely the reason for their lower performance. Of the simpler models, *Linear regression* does not include a regularization technique, which makes it less suitable for our dataset. This remaining regression models, *Lasso*, *Ridge* and *Elastic Net regression*, differ in their regularization method. *Lasso regression* uses *L1* regularization, which penalizes the model based on the sum of the absolute coefficient values. It sets irrelevant coefficients to zero, thereby performing automatic feature selection. *Ridge regression* on the other hand uses *L2* regularization, which penalizes the model based on the sum of the squared coefficient values. It minimizes the size of all coefficients, while preventing removal of any coefficients from the model. This mitigates the effect of multicollinearity in the dataset. *Elastic Net regression* uses a combination of both *L1* and *L2* regularization (Brownlee, 2020). Therefore, it is able to select relevant features and simultaneously deal with multicollinearity. This combination is likely the reason for *Elastic Net* to outperform the other models (Altelbany, 2021).

4.2 Hyperparameter tuning

For *Elastic Net regression*, a grid search is performed in order to find the optimal combination of hyperparameters. The combination with the highest performance is as follows:

- *Alpha* = 0.1
- *L1-ratio* = 0.2
- *Max iterations* = 1e4
- *Selection* = random

All parameters were relatively stable throughout the 50 iterations. Interestingly, both *alpha* and the *L1-ratio* are on the low end of their possible ranges. For *alpha*, this indicates that the strength of regularization is small, implying that the model closely resembles ordinary least squares used in *Linear regression*. As for the *L1-ratio*, this means that *L2* regularization is dominant. Therefore, the regularization is primarily focussed on mitigating the effects of multicollinearity rather than feature selection. This is expected, since feature selection is already performed before training the model.

4.3 *Elastic Net* performance

Using the tuned *Elastic Net* model, runs are performed over a range of 100 most important features. Figure 3 shows the results of these runs. Both the highest mean R^2 (0.38) and mean RMSE (1.11) improved substantially after tuning. However, the performance is still relatively poor. Additionally, the plot shows that the performance on the training set is significantly larger. Due to the large number of features, this level of overfitting is expected.

Increasing the number of features beyond the optimum amount decreases the performance of the model in terms of R^2 and RMSE, while the training performance keeps improving. Up to the optimal value of *k*, the model learns patterns from the training set that are generalizable. Beyond this point, the model learns patterns that are specific to the training set, but are not present in the test set. Even though the optimal amount of features is different for R^2 (46) than for RMSE (39), a similar pattern of overfitting can be observed. This shows that, on average, the model is more likely to overfit the training data when using more than approximately 50 features.

One key finding that is shown in Figure 3 are the extreme variations in model performance due to different train-test splits. This is shown by the standard deviation for the 50 iterations, depicted by the blue shading. Because these variations are very large, only looking at the specific features that are used when the mean performance is highest is not reliable enough. As outlined in the method section, we therefore stored the most important features in model runs with an R^2-value above 0.6. It is important to note that, even though the standard deviation of R^2 does not reach 0.6 for any of the *k*-values, there certainly are model runs in the 50 iterations that reach this performance level.

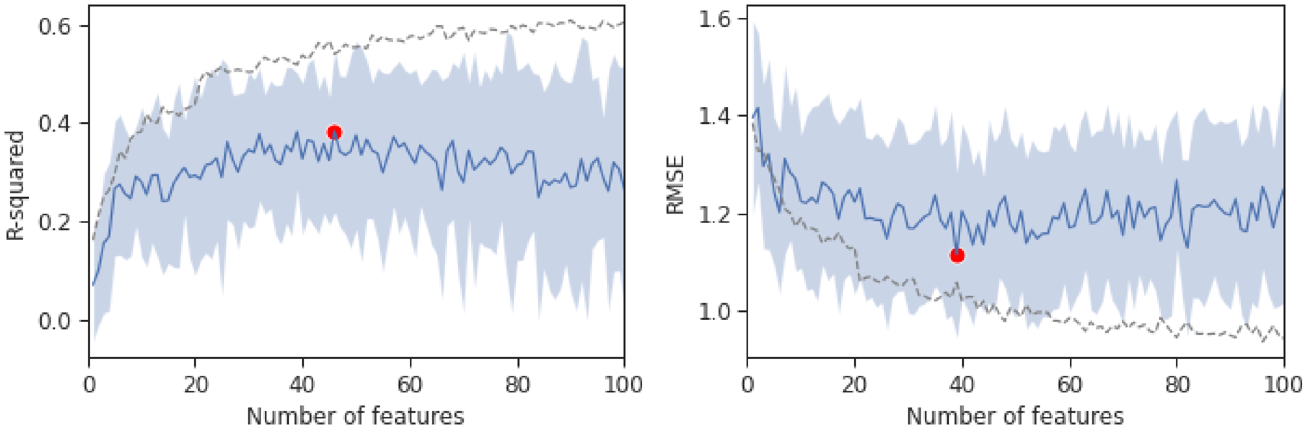


Figure 3: Dependency of model performance on the number of features. Accuracy (R^2, left) and error (RMSE, right) of the Elastic Net model runs is shown. The blue line indicates the mean performance of the 50 iterations, while the shading depicts one standard deviation from this mean. The red dot shows the highest mean value for both performance metric. Additionally, the dashed grey line indicates the mean performance on the training set.

4.4 Feature importance including cross-terms

Since 50 iterations for every of the 100 *k*-values are computed, the model performance of 5000 model runs is obtained. Of these runs, only 69 reached an R^2-value that exceeded the threshold of 0.6. For these 69 runs, the 10 most occurring features with an absolute regression coefficient above 0 are shown in Table 4.

The first thing to note is that all 10 most important features are cross-terms. This makes the interpretation of their effect more complicated. Generally, we see that insulation count and historical electricity usage occur most often in the top 10 cross-terms. Interestingly, cross-terms that include insulation count have a negative regression coefficient. For historical electricity usage, the regression coefficients are positive.

Furthermore, variables related to weather data occur several times in the top 10 features. Examples are ice formation, rain, sun hours and wind speed. These variables are less relevant in answering our research question for two reasons. First, these variable are highly correlated, which complicates the interpretability of their effect on the saving performance score. For example, sun hours and rain are negatively correlated, and both are in turn related to the outdoor temperature. This makes it difficult to interpret their effect, especially when they occur as cross-terms. Secondly, the relationship between weather data and the saving performance score is less useful in optimizing the performance of hybrid heat pumps, which is an extension to the goal of this project. Simply said, owners of a hybrid heat pump have no control over the weather. Therefore, knowing the relationship between these weather variables and the saving performance score does not help them in optimizing the performance. One exception to this could be ice formation or rain fall at low outdoor temperatures, as it might directly influence the outside unit of the heat pump, which could be controlled for.

Table 4: The 10 most occurring relevant features for model runs with an R^2-value above 0.6. For cross-terms, both features are shown in separate columns (Feature 1 and 2). The temperature bins associated with the most important features is shown as well. Their importance is indicated by the amount of times they occur in the best model runs and their regression coefficients.

Feature 1	Feature 2	T. bin (°C)	Counts	Regr. Coef.
Rain	Historical electricity	≥15	69	0.27
Supply-return temperature	Insulation count	10-15	69	-0.26
Ice formation	Historical electricity	0-5	67	0.19
Wind speed	Insulation count	<0	67	-0.23
Min outdoor temperature	Insulation count	10-15	66	-0.19
Wind speed	Insulation count	10-15	66	-0.3
Min outdoor temperature	Insulation count	5-10	65	-0.16
Supply-return temperature	Insulation count	≥15	65	-0.16
Sun hours	Historical electricity	10-15	65	0.21
Sun hours	Historical electricity	<0	65	0.21

4.5 Feature importance in data subset

To increase the interpretability, two subsets of the data were defined and the model performance on them analysed. The first subset contains the same variables as the one discussed above, but without their cross-terms. The second subset also does not include the weather data from the KNMI, in addition to this. The effect of these subsets on the model performance and the most important features are shown below.

4.5.1 Excluding cross-terms

When excluding cross-terms, the model performance in terms of mean R^2 (0.26) and mean RMSE (1.27) decreased substantially. This is expected, since the number of features is reduced. Due to the decreased performance, the R^2 threshold is lowered to 0.4 in order to obtain enough model runs. The most important features are shown in Table 5.

Generally, the features are similar to those found when including cross-terms. However, there are three meteorological features that are less important when excluding cross-terms: ice formation (0-5 °C), wind speed (10-15 °C) and sun hours (<0 and 10-15 °C). This indicates that these three features mainly affect the saving performance score indirectly. Additionally, rain (<0 °C) and the minimum outdoor temperature (0-5 °C) are one of the most important features when excluding cross-terms. Below, two examples of unexpected and expected effects of important features are interpreted. Note that these interpretations are merely hypotheses, and further research is needed to test their validity.

Table 5: The 10 most occurring relevant features for model runs with an R^2-value above 0.4, when excluding cross-terms. The temperature bins associated with the most important features is shown and their importance is indicated by the amount of times they occur in the best model runs and the regression coefficients.

Feature	T. bin (°C)	Counts	Regr. Coef.
Min outdoor temperature	5-10	44	-0.27
Min outdoor temperature	10-15	44	-0.47
Rain	<0	44	-0.35
Historical electricity		44	0.46
Wind speed	<0	43	-0.28
Rain	≥15	42	0.56
Min outdoor temperature	0-5	41	-0.16
Supply-return temperature	10-15	41	-0.32
Supply-return temperature	≥15	40	-0.26
Insulation count		40	-0.53

Unexpected effects

Interestingly, the effect of rain on the saving performance score depends on the temperature bin. For outdoor temperatures below 0 °C, the effect is negative, likely due to snowfall blocking the airflow through the outside unit. Additionally, rain on cold days possibly increases the energy demand of the heating system due to increased indoor occupancy, which negatively affects the saving performance score. Conversely, the effect of rain on the saving performance score is positive for temperatures above 15 °C. However, limited measurements makes this result less reliable (see Figure E1 in Appendix E).

It is expected that the efficiency of a heat pump increases with increasing outdoor temperatures. However, the minimum outdoor temperature for bins between 0 and 15 °C negatively affects the saving performance score. Notably, the minimum outdoor temperature shows a strong negative correlation with maximum outdoor temperature for these bins (Pearson's r = -0.59). The implications of this are twofold. Firstly, the negative correlation could suggest that the maximum instead of the minimum outdoor temperature affects the saving performance most in these temperature bins. However, maximum outdoor temperature shows a weaker correlation with the saving performance score, which is not in line with the hypothesis. Secondly, this could imply that the difference between the minimum and maximum outdoor temperature affects the saving performance score. Preliminary correlation analysis indeed shows that this difference is more strongly correlated with the saving performance score than minimum or maximum outdoor temperature for bins 0-5 and 10-15 °C (see Appendix F).

Notably, the minimum outdoor temperature for bin <0 °C is not considered important. Since the relative energy supply by the heat pump is lowest for this bin (see Appendix G), it is expected that differences in this relative supply between houses could result in significant variations in gas usage and, with that, in saving performance score. This is substantiated by the fact the heat pump energy is negatively related with gas usage for temperatures below 5 °C (see Appendix H). This effect of the minimum outdoor temperature on the saving performance score is not observed, possibly because this minimum generally occurs at night when the heat pump is not used. However, the maximum outdoor temperature shows a strong positive correlation with the saving performance score for bin <0 °C (see Appendix F), indicating that this variable might be more appropriate to capture the expected relationship in this bin.

Expected effects

The relationship between historical electricity consumption and the saving performance score is positive. Since historical electricity consumption is measured before instalment of a heat pump, it serves as a proxy for historical gas consumption, which positively affects the saving performance score (Equation 1). This is substantiated by the strong positive correlation between historical electricity and gas consumption (Pearson's r = 0.33).

Wind speed negatively affects the saving performance score for outdoor temperatures below 0 °C. Similar to rainfall in this temperature bin, this is likely related to an increase in energy demand due increased indoor occupancy.

4.5.2 Excluding cross-terms and weather data

When excluding weather data in addition to excluding cross-terms, the mean R^2 (0.08) and mean RMSE (1.37) decreased even further. Therefore, the R^2 threshold is lowered to a value of 0.2 in order to obtain enough model runs. Note that this negatively affects the reliability of our results. The most important features are shown in Table 6.

All non-meteorological variables in Table 5 are observed in Table 6 as well. The meteorological variables are replaced by the following variables: number of inhabitants, flow rate (0-5 and ≥15 °C), indoor humidity (5-10 °C), boiler energy (<0 °C) and electricity consumption (0-5 °C).

Note that the boiler energy at temperatures below 0 °C negatively affects the saving performance score, which is in line with the above mentioned hypothesis that the relative energy supply by the heat pump compared to the boiler at these temperatures could impact the saving performance score.

Table 6: The 10 most occurring relevant features for model runs with an R^2-value above 0.2, when excluding cross-terms and weather data. The temperature bins associated with the most important features is shown and their importance is indicated by the amount of times they occur in the best model runs and the regression coefficients.

Feature	**T. bin (°C)**	**Counts**	**Regr. Coef.**
Supply-return temperature	10-15	37	-0.40
Historical electricity		37	0.54
Insulation count		37	-0.50
Supply-return temperature	≥15	36	-0.33
Number of inhabitants		33	0.23
Flow rate heating system	0-5	32	-0.15
Indoor humidity	5-10	32	0.19
Boiler energy	<0	30	-0.27
Electricity consumption	0-5	29	-0.26
Flow rate heating system	≥15	27	-0.24

4.6 Comparison with clustering and classification

Alongside the experiments conducted here, other group members used classification and clustering techniques to analyse the same dataset (see Chapter 1).

To compare the *Elastic Net* performance with the best classification model, the accuracy of the best model run is computed using the ranges of the saving performance classes made by the other group members. This resulted in a balanced accuracy of 69% and 58% when using the dataset including cross-terms and the subset excluding cross-terms and weather data, respectively.

The best performing classification model is a tuned *Random Forest (RF) classifier* with a balanced accuracy of 59% and 51%. The balanced accuracy of the *Elastic Net* model is higher than the *RF classifier* in the comparative experiments.

It is important to note that the classification dataset did not undergo manual feature selection based on correlations. Similar to our results, the flow rate, indoor humidity and boiler energy occur in the top 10 features, but in different temperature bins (see Appendix I). Additionally, the supply and return temperature of the heating system, which are separate variables in the classification dataset, are deemed important. Contrary to our results, the indoor temperature is considered important in the *RF classifier*. Interestingly, house characteristics obtained from the metadata are not important in the classification model, while the insulation count and number of inhabitant do affect the saving performance score in the *Elastic Net* model.

The best performing clustering model is a *K-means* model. The performance of this model cannot be compared to ours, because the silhouette score is used as a metric. However, the most important features can be analysed. It should be noted that the *K-means* model is employed on a dataset that does not include temperature bins. This complicates the comparison with our findings. A similar relationship between the difference in supply and return temperature and the saving performance score is found by the *K-means* model (see Appendix I). Contrary to our findings, the difference between indoor and outdoor temperature instead of the minimum outdoor temperature was considered important by the clustering model.

5. Discussion

In this Chapter, a reflection on the dataset and the methods used to perform the analysis is outlined. Additionally, suggestions for further research are provided.

Even though several model runs showed a high performance in predicting the saving performance score, the variations with changing train-test splits are extremely large. This indicates that the dataset contains a relatively low number of samples compared to the number of features. There are several ways to improve this imbalance.

The number of houses in the analysed dataset could be increased by reducing the amount of missing data. Especially missing data related to historical gas and electricity consumption affected the number of samples, since these variables are used to compute the saving performance score. Furthermore, missing metadata could be obtained by resending the survey to the households.

Additionally, the acquisition of longer measurement time series could increase the reliability of our results in two ways. Firstly, including data over multiple years reduces the effect of weather conditions for one specific winter season. Secondly, it will increase the overlap in time series between houses, resulting in a more reliable comparison.

The binning method used to mitigate the effect of non-overlapping timeseries introduced errors into the dataset as well. When binning on outdoor temperature, it is assumed that days with the same mean temperature can be compared. However, the behaviour of inhabitants is likely different on a 10 °C day in winter compared to summer. Furthermore, binning introduces imbalance into the dataset, since not all temperatures occur with the same frequency (see Table 2 in Section 2.2). As a result, the differences in measurement time are still present in temperature bin ≥15 °C in our dataset. In some cases, this results in inaccurate relationships (see Appendix I), which underlines the need for longer time series to reliably compare households in this temperature bin.

Data on the manufacturers of the heat pumps was excluded from the dataset provided for this analysis (see Section 2.1.1). However, it is likely that the manufacturer plays a role in the efficiency of the heat pump. Therefore, it could affect the results found in this study.

It would be interesting to conduct further research by analysing the data of the previous winter only, as this is when the heating system is predominantly active. This would ensure that the time series of the houses coincide. Additionally, valuable insights could be gained by incorporating variations of variables throughout the day, in addition to their daily mean values. Finally, enriching the dataset could involve providing additional metadata (e.g. on the placement and cover of the outside unit) and increasing the number of sensors (e.g. for the outside and inside unit separately).

6. Conclusions

The findings of this study indicate that *Elastic Net* outperforms other regression models. This can be attributed to the model's capability to select relevant features and effectively handle multicollinearity. However, large variations in the model performance are observed as a result of changes in the split of training and test data. The highest performance is achieved when using the dataset that includes cross-terms, although this complicates the interpretability of the most important features impacting the saving performance score. Therefore, two subsets of data without cross-terms are used to increase interpretability, despite a decrease in model performance.

The key characteristics that account for the variation in the saving performance score among houses in the *Demoproject Hybride* are a combination of house and weather time series data and house metadata. Specifically, the difference between the supply and return temperature of the heating system emerges as an important factor in the *Elastic Net* model. In terms of weather data, the minimum outdoor temperature, rain and wind speed affect the saving performance score. Additionally, the historical electricity usage and insulation level of the house played an important role. Notably, in some cases, the effect of these features on the saving performance score cannot be fully explained.

While these results provide valuable insight into the performance of hybrid heat pumps in real-world scenarios, longer measurements are necessary to increase the reliability of these findings. Particularly for temperatures above 15 °C, a clear divide in houses based on measurement time is observed.

Considering the Dutch Government's mandate for the adoption of (hybrid) heat pumps starting in 2026, this study emphasises the importance of conducting similar research on the characteristic that influence the differences in performance of hybrid heat pumps in real-world conditions in the Netherlands. This has the potential to increase the effectiveness of hybrid heat pumps in reducing carbon emissions of residential areas.

References

Altelbany, S. (2021). Evaluation of ridge, elastic net and lasso regression methods in precedence of multicollinearity problem: a simulation study. Journal of Applied Economics and Business Studies, 5(1), 131-142. https://doi.org/10.34260/jaebs.517

Brownlee, J. (2020). How to develop elastic net regression models in Python. Tutorial in Python Machine Learning on Machine Learning Mastery. https://machinelearningmastery.com/elastic-net-regression-in-python/

Centraal Bureau voor de Statestiek. (2022). Huishoudens nu [Webpagina]. Centraal Bureau voor de Statistiek. https://www.cbs.nl/nl-nl/visualisaties/dashboard-bevolking/woonsituatie/huishoudens-nu

Centraal Bureau voor de Statestiek. (2022, October 14). Woonoppervlakte in Nederland [Webpagina]. Centraal Bureau voor de Statistiek. https://www.cbs.nl/nl-nl/achtergrond/2018/22/woonoppervlakte-in-nederland

De Nederlandse Verwarmingsindustrie. (2021, September 19). Hybride warmtepompen, haalbaar en betaalbaar. Report. https://www.verwarmingsindustrie.nl/publicaties/hybride-warmtepompen-haalbaar-en-betaalbaar/

Demonstratieproject Hybride warmtepompen. (n.d.). Retrieved May 20, 2023, from https://www.demoprojecthybride.nl/

Dutch Green Building Council. (2021). Position Paper Whole Life Carbon. https://www.dgbc.nl/publicaties/position-paper-whole-life-carbon-44

Government of the Netherlands. (2019, June 28). Klimaatakkoord. In Klimaatakkoord.nl. Retrieved June 21, 2023, from https://www.rvo.nl/files/file/2020/05/Klimaatakkoord%20-%2028%20juni%202019.pdf

Hoebergen, A. (CBS), van Middelkoop, M. (CBS) and van Polen, S. (PBL). (2021, February 18). Lagere energierekening, effecten van lagere prijzen en energiebesparing [Webpagina]. Centraal Bureau voor de Statistiek. https://www.cbs.nl/nl-nl/longread/rapportages/2021/lagere-energierekening-effecten-van-lagere-prijzen-en-energiebesparing

Installatiemonitor. (2022). Publieke eindrapportage februari 2022. https://www.installatiemonitor.nl/eindrapportage-installatiemonitor-2/

Kuijeren, F. van. (2021, August 12). Aan het bouwjaar van je woning zien hoe die geïsoleerd is. VK Makelaars. https://vkmakelaars.nl/blog/bouwkundig-advies/aan-het-bouwjaar-van-je-woning-zien-hoe-die-geisoleerd-is/

Miara, M., & Kramer, T. (2011). Heat Pump Efficiency Analysis and Evaluation of Heat Pump Efficiency in Real-life Conditions Abbreviated Version. https://wp-monitoring.ise.fraunhofer.de/wp-effizienz/download/final_report_wp_effizienz_en.pdf

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2023, May 1). Reikwijdte normering verwarmingsinstallaties. https://www.rijksoverheid.nl/documenten/kamerstukken/2023/05/01/kamerbrief-over-reikwijdte-normering-verwarmingsinstallaties

Oikonomou, E., Zimmermann, N., Davies, M., & Oreszczyn, T. (2022). Behavioural Change as a Domestic Heat Pump Performance Driver: Insights on the Influence of Feedback Systems from

Multiple Case Studies in the UK. Sustainability, 14(24), 16799. https://doi.org/10.3390/su142416799

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830. Link to the exact model: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

Wolf (2023). Types of heat pumps. https://www.wolf.eu/en-de/advisor/types-of-heat-pumps

Appendices

A. Data description

An overview of the house time series data is shown in Table A1. For each variable, their associated unit, measurement frequency and aggregation method is included.

An overview of the weather time series data is shown in Table A2. This also included the unit, measurement frequency and aggregation method for each variable.

Table A1: Overview of the house time series data. The 'Time resolution' indicates the frequency at which new measurements were recorded. 'Aggregation Method' indicates the aggregation technique used for each variable.

Sensor	**Variable**	**Unit**	**Time resolution (s)**	**Aggregation Method**
Heating system	Supply temperature system	℃	5	Mean
	Return temperature system	℃	5	Mean
	Cumulative energy consumption	kWh	5	Max
	Power	kW	5	Mean
	Water flow	L/min	5	Mean
Heat Pump	Cumulative energy consumption	kWh	60	Max
	Power	kW	5	Mean
Boiler	Cumulative energy consumption	kWh	60	Max
	Power consumption	kW	5	Mean
DSMR	Power consumption	kW	10	Mean
	Power production	kW	10	Mean
	Cumulative energy consumption	kWh	60	Max
	Cumulative energy production	kWh	60	Max
	Cumulative volume gas consumption	m^3	60	Max
Thermohygrometer	Room temperature	℃	30	Mean
	Room humidity	%	30	Mean

Table A2: Overview of the weather time series *data. The 'Time resolution' indicates the frequency at which new measurements were recorded. 'Aggregation Method' indicates the aggregation technique used for each variable.*

Variable	**Unit**	**Time resolution**	**Aggregation method**
Temperature	℃	hour	Mean, Min and Max
Sun hours	hours	hour	Sum
Wind speed	m/s	hour	Mean
Rain	mm	hour	Sum
Ice form	Boolean	hour	Mean
Relative Outdoor humidity	%	hour	Mean

B. Multicollinearity

Figure B1 show the correlation between several key variables, including the saving performance score. The time series variables are associated with the 0-5 °C temperature bin. This depicts that these features are not only correlated with the saving performance score, but also with each other. Therefore, it indicates that multicollinearity is present in our data.

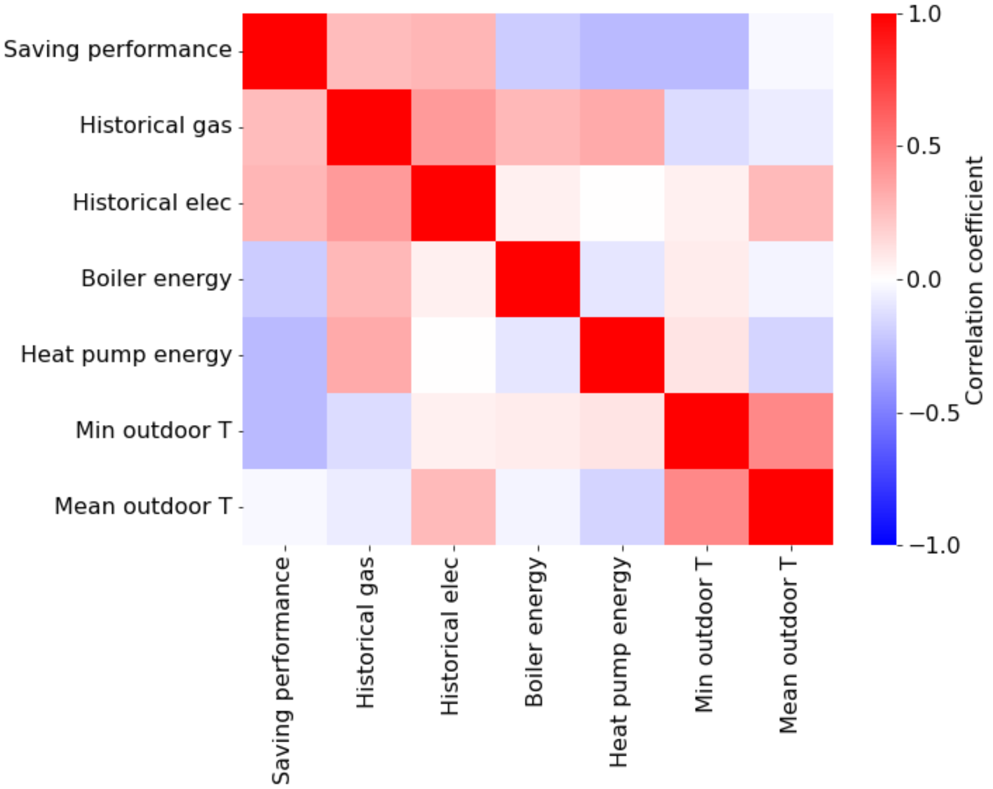


Figure B1: Heatmap of the correlation between several key variables, including the saving performance score. The time series variables are associated with temperature bin 0-5 °C.

C. Hyperparameter tuning

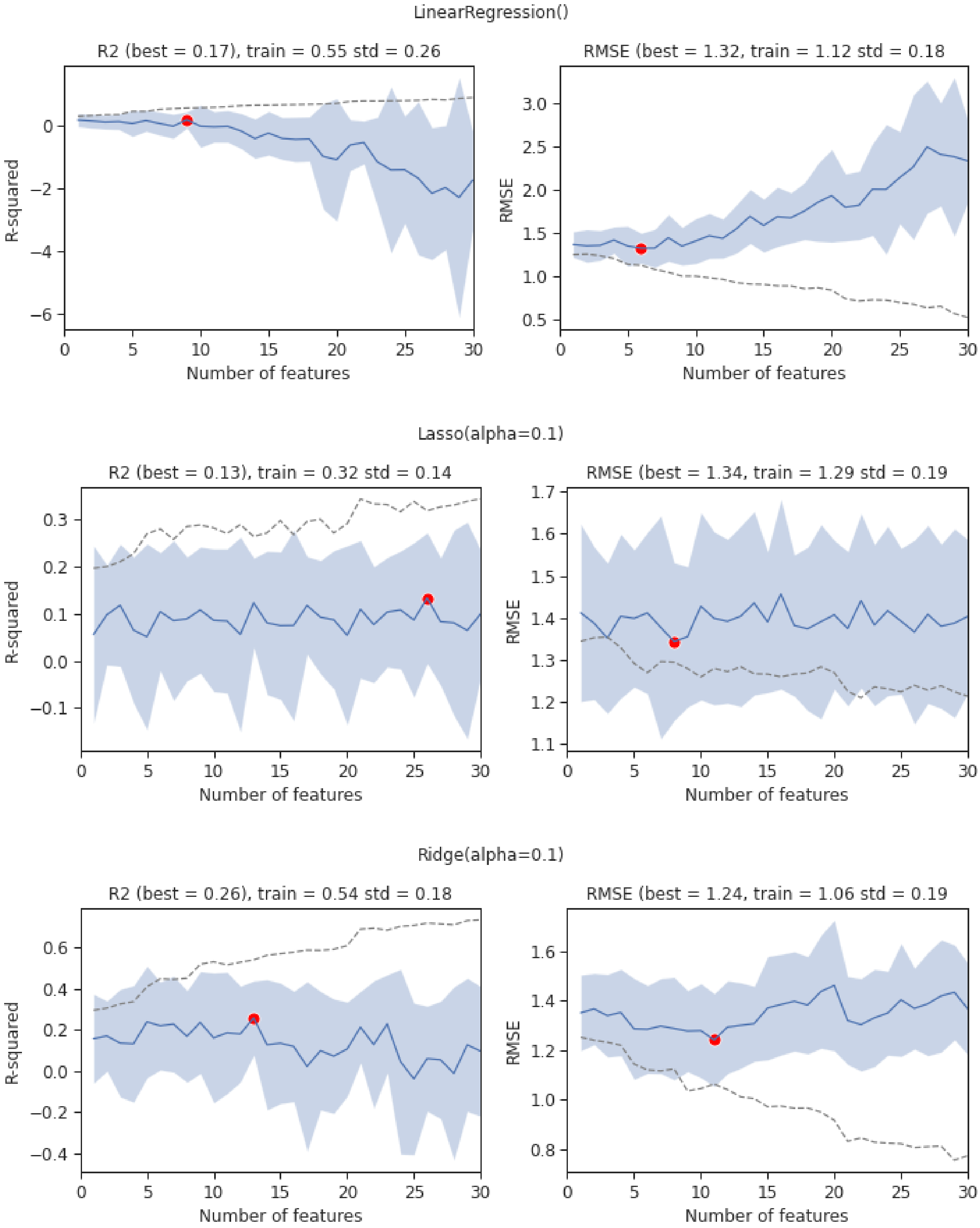
The parameters that are used in the grid search are shown in Table C1. It contains a short description of each variable and the range of values that are included in the grid search. A more detailed description of the parameters can be found in Pedregosa et al. (2011).

Table C1: The hyperparameters of the Elastic Net model using in the grid search. A short description and the range of values are included.

Parameter	**Description**	**Range of values**
Alpha	Factor that determines the strength of the L1 and L2 regularization components	[0.01, 0.1, 0.5, 1, 5]
L1 ratio	Mixing ratio between L1 and L2 components. Ranges from 0 (only L2) to 1 (only L1)	[0.2, 0.4, 0.6, 0.8]
Max iterations	The maximum number of iterations	[1e4, 1e5]
Selection	Set the method for coordinate descent, which means the order in which coefficients are updated at each iteration	[cyclic, random]

D. Model selection

Figure D1 depicts the dependency of model performance on the number of features. Accuracy (R^2, left) and error (RMSE, right) of the 6 regression models is shown. The best mean value for each metric and the associated *k*-value is indicated by the red dot and quantified in the title of the figures.



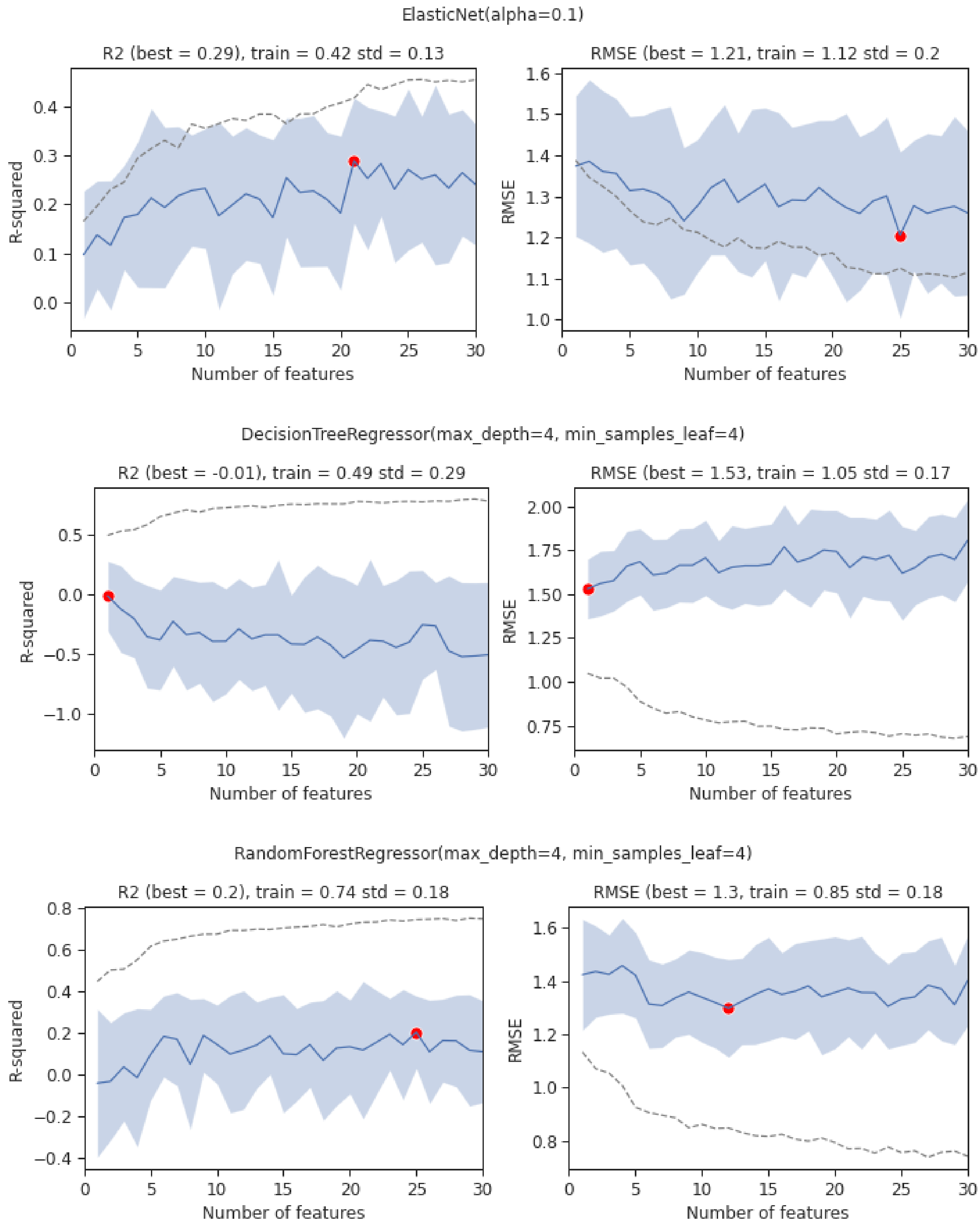


Figure D1: Dependency of model performance on the number of features. Accuracy (R^2, left) and error (RMSE, right) of the 6 regression models is shown. The blue line indicates the mean performance of the 50 iterations, while the shading depicts one standard deviation from this mean. The red dot shows the highest mean value for both performance metric. The best mean and standard deviation of the metric are shown in the title of the figures. Additionally, the dashed grey line indicates the mean performance on the training set.

E. Measurement days in bin ≥15

Figure E1 shows that the measurement time affects the relationship between the mean daily cumulative rainfall and the saving performance score for bin ≥15 °C.

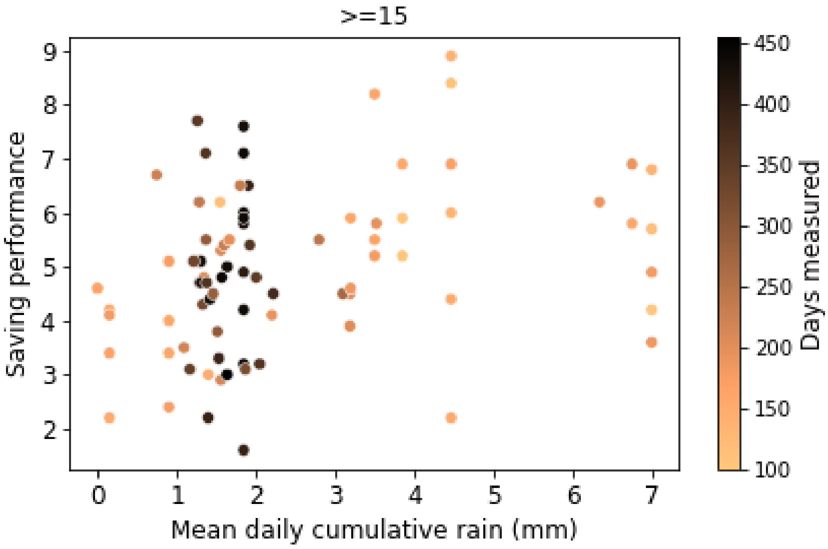


Figure E1: The relationship between mean daily cumulative rainfall and the saving performance score for bin ≥15 °C. The colours indicate the measurement time in days for each house.

Figure E2 shows that the measurement time affects the relationship between the mean outdoor temperature and the saving performance score for bin ≥15 °C.

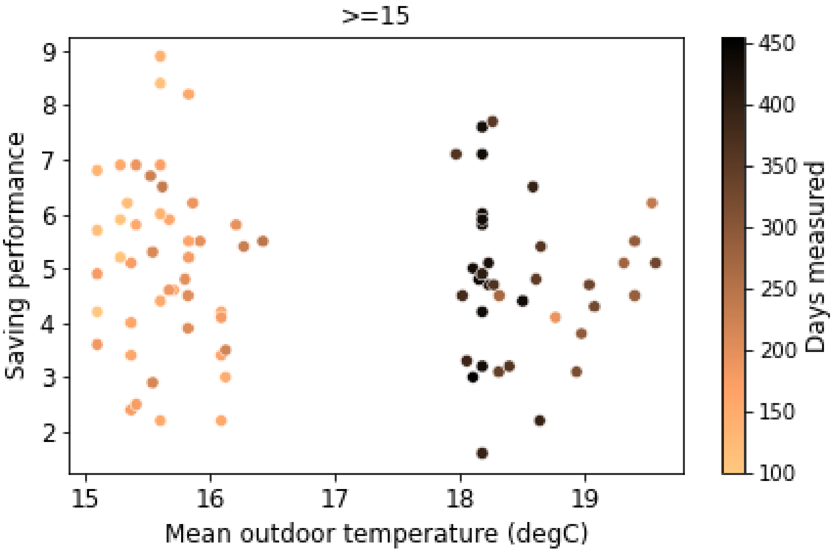


Figure E2: The relationship between mean outdoor temperature and the saving performance score for bin ≥15 °C. The colours indicate the measurement time in days for each house.

F. Correlation outdoor temperature with saving performance score

Figure F1 shows the correlation between outdoor temperature and the saving performance score for every temperature bin. The minimum and maximum outdoor temperature and the difference between the two is shown. It can be observed that the difference between the minimum and maximum outdoor temperature is more strongly correlated with the saving performance score for bin 0-5 °C and 10-15 °C.

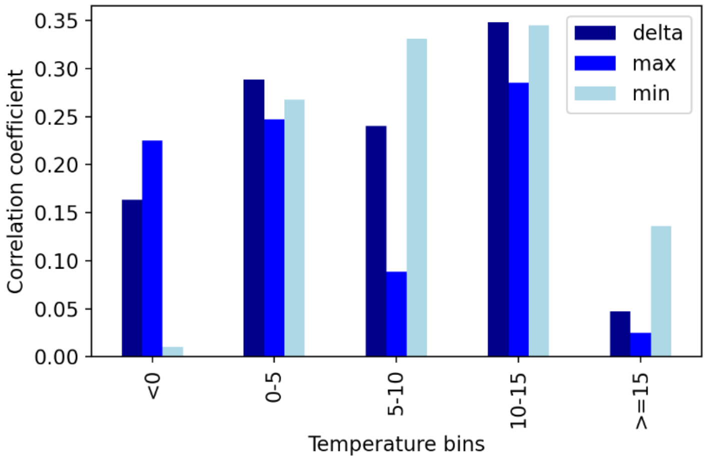


Figure F1: Correlation between outdoor temperature and the saving performance score for each temperature bin. The minimum and maximum outdoor temperature, and difference between the two are shown. Note that the correlation coefficients of minimum outdoor temperature are in fact negative, but their absolute value is plotted to enable a comparison between the variables.

G. Relative energy supply by heat pump

Figure G1 shows the energy supply by the boiler and the heat pump for each temperature bin. The energy supply by the boiler is estimated by converting the gas consumption to kWh, and subtracting the gas consumption in bin ≥15. The motivation for this is that we assume that at these temperatures, the gas consumption is merely related to heating up water for showering.

The energy supply by the heat pump is estimated from the energy demand of the heat pump, and using an SCOP of 3.8 to convert this to energy supply (Installatiemonitor, 2022). Note that this is a rough estimation of the real energy supply by the heat pump.

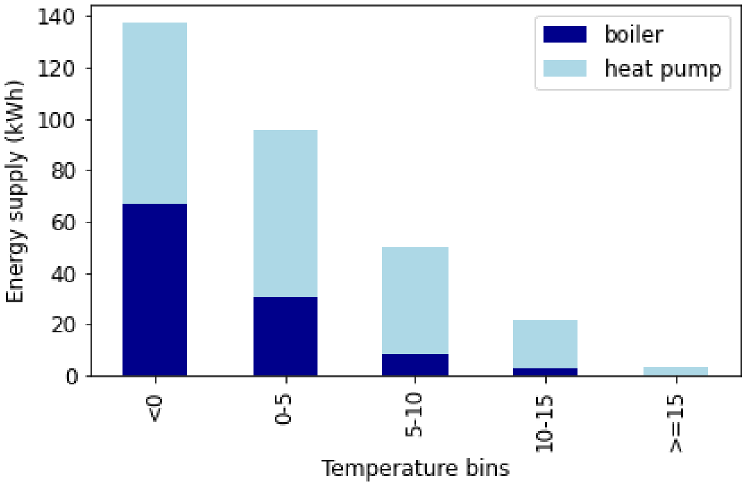


Figure G1: The energy supply by the boiler and the heat pump for each temperature bin. Note that these values are estimated from the gas consumption of the boiler and electricity usage by the heat pump.

H. Heat pump energy and gas usage

Figure H1 shows the relationship between heat pump energy usage and gas consumption for temperature bins <0 °C (left) and 10-15 °C (right). The colour indicates the historical annual gas consumption for each household. It can be observed that the negative relationship between both variables in bin <0 °C breaks down for bin 10-15 °C.

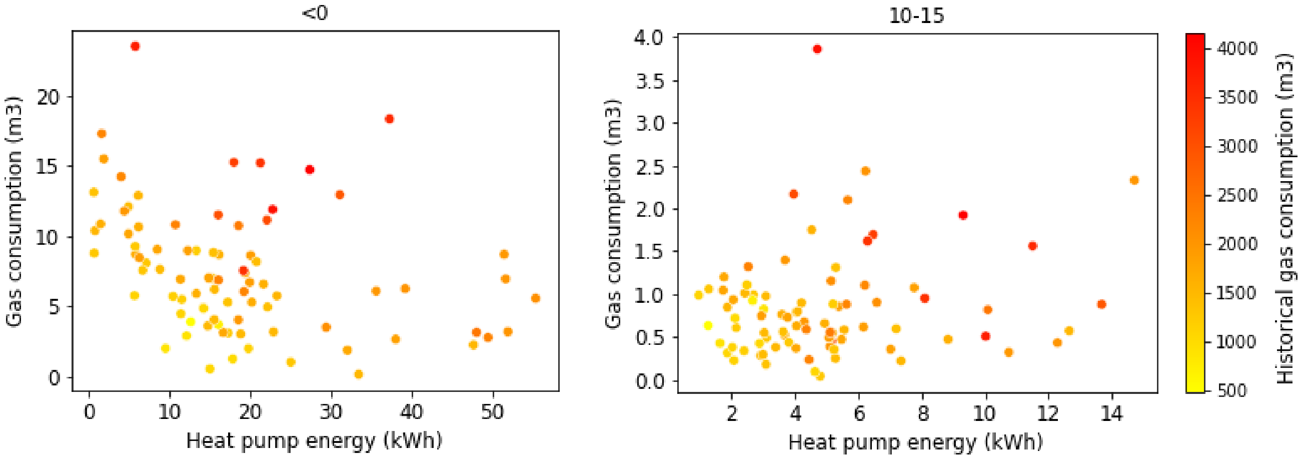


Figure H1: Relationship between heat pump energy and gas consumption for bin <0 °C (left) and 10-15 °C (right). The historical gas consumption of each house is indicated by the colours.

I. Comparison with classification and clustering

Table I1 shows the most important features found by the best performing regression, classification and clustering models. A dataset excluding cross-terms and weather data is used for the regression and classification model. It should be noted that manual feature removal is not performed on the dataset used by the classification and clustering models. The clustering model does not include temperature bins in addition to this.

Table I1: Most important features found by the best performing regression, classification and clustering model. Respectively, these include Elastic Net regression, Random Forest classifier and K-means clustering. Note that the dataset used for the analysis are not identical.

Regression	**T. bin**	**Classification**	**T. bin**	**Clustering**
Supply-return temperature	10-15	Supply temperature	5-10	Supply-return temperature
Historical electricity		Max indoor temperature	5-10	Indoor-outdoor temperature
Insulation count		Flow rate heating system	10-15	Heat pump power
Supply-return temperature	≥15	Return temperature	0-5	Rain
Number of inhabitants		Indoor humidity	<0	Sun hours
Flow rate heating system	0-5	Boiler energy	≥15	Indoor humidity
Indoor humidity	5-10	Boiler power	≥15	Ice formation
Boiler energy	<0	Boiler energy	10-15	
Electricity consumption	0-5	Mean indoor temperature	0-5	