Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Federal Department of the Environment, Transport, Energy and Communications DETEC

Swiss Federal Office of Energy SFOE
Energy Research and Cleantech Division

Final report from July 2, 2026

KI in der Wärmepumpenberatung

Zusatzbericht

Prediction of Household Energy Potential

Date: July 2, 2026

Location: Zurich

EKZ heat pump advisory programme

Publisher:
Swiss Federal Office of Energy SFOE
Energy Research and Cleantech
CH-3003 Bern
www.bfe.admin.ch

Subsidy recipients:

Bits to Energy Lab, ETH Zurich
Weinbergstrasse 56/58, CH-8092 Zurich
www.bitstoenergy.com

Authors:
David Schaurecker, Bits to Energy Lab, ETH Zurich, schaureckerd@ethz.ch
Prof. Dr. Thorsten Staake, Bits to Energy Lab, ETH Zurich, tstaake@ethz.ch

Other contacts:
Marina González Vayá, Elektrizitätswerke des Kantons Zürich, marina.gonzalezvaya@ekz.ch
Lorenz Deppeler, Elektrizitätswerke des Kantons Zürich, lorenz.deppeler@ekz.ch

BFE project coordinators:
Stephan Renz, info@renzconsulting.ch
Elena-Lavinia Niederhäuser, elena-lavinia.niederhaeuser@bfe.admin.ch

SFOE contract number: SI/502257-01

Summary

This report is a second addition (Work Package 2) to the overarching project "KI in der Wärmepumpenberatung" (BFE project number SI/502257). The objective of this addition was to estimate, before an on-site advisory visit takes place, which households are likely to achieve particularly high electricity savings from a heat-pump service and settings optimization. The work therefore extends the earlier retrospective savings analysis by adding a forward-looking prediction layer based on pre-visit 15-min smart-meter data. The broader project and data context are described by Brudermueller et al. [1].

The final methodology combines two components. First, achieved savings are estimated with a temperature-corrected matched difference-in-differences design. Second, a prediction model ranks households by expected savings. The treated evaluation sample contains 63 households, comprising 40 air-source and 23 ground-source heat pumps, which motivates the final stratified prediction design. Several candidate models were compared, and the best overall result was obtained with a stratified k-nearest-neighbours ensemble. In the final evaluation across repeated, randomized test sets, the model achieved a mean aggregated holdout Spearman rank correlation of 0.24, a mean absolute error of 7.46 % points relative to the household-level achieved savings labels from the matched-difference-in-differences (DiD) step, and a positive rank correlation in 82 % of holdout splits.

Average achieved savings remained positive in all repeated reruns. The mean estimated electricity saving effect was 6.92 %, which corresponds to mean achieved annual savings of about 876 kWh or CHF 176, and ranged from 4.44 % to 8.75 % across reruns. On a household-by-household basis, expected and achieved savings also remained positively associated: Spearman rank correlation reached 0.27 $(p = 0.033)$. More importantly for prioritization, when households were ordered by predicted savings, the top predicted tercile achieved 9.71 % savings on average, compared with 6.92 % across the full treated sample and 4.36 % in the bottom predicted tercile, a gap of 5.35 percentage points. For reference, the true achieved-savings bottom and top terciles averaged –3.48 % and 16.68 %, respectively, so the predicted top tercile is clearly shifted toward the genuinely high-saving households even though it does not fully recover the separation of the true top tercile. The model's ex ante predictions correspond to mean expected annual savings of about 981 kWh or CHF 195 under the time-of-use tariff assumptions used here. These are the average pre-visit savings signals the model would have assigned across treated households, whereas the ex post matched-DiD benchmark implies mean achieved savings of about 876 kWh or CHF 176. The gap is therefore about 105 kWh or CHF 19 on average, meaning that the ex ante estimate is roughly 12 % higher than the ex post benchmark.

Taken together, the results show that smart-meter data can support a practical screening tool for household prioritization. The evaluation sample remains modest, and the absolute savings and economic interpretation should therefore be viewed with appropriate caution. The configuration-change and influence analyses also remain exploratory. Nonetheless, the repeated multi-seed results provide confidence that the central ranking prediction captures a real and practically useful signal for identifying households with comparatively high savings potential before an advisory visit.

Zusammenfassung

Dieser Bericht ist ein zweiter Zusatz (Work Package 2) zum übergeordneten Projekt «KI in der Wärmepumpenberatung» (BFE-Projektnummer SI/502257). Das Ziel dieses Zusatzes war es, noch vor einem Beratungsbesuch vor Ort abzuschätzen, welche Haushalte voraussichtlich besonders hohe Stromeinsparungen durch eine Optimierung des Wärmepumpenbetriebs und der Einstellungen erzielen können. Die Arbeit erweitert damit die frühere retrospektive Einsparanalyse um eine vorausschauende Prognoseebene auf Basis von 15-Minuten-Smart-Meter-Daten vor dem Beratungstermin. Der breitere Projekt- und Datenkontext wird von Brudermüller et al. [1] beschrieben, auf welchen der Zusatz nun aufbaut.

Die finale Methodik kombiniert zwei Komponenten: Erstens werden die tatsächlich erzielten Einsparungen mit einem temperaturbereinigten, gepaarten Difference-in-Differences-Ansatz geschätzt. Zweitens ordnet ein Prognosemodell die Haushalte nach den erwarteten Einsparungen. Die behandelte Evaluationsstichprobe umfasst 63 Haushalte, davon 40 mit Luft-Wasser- und 23 mit Erdsonden-Wärmepumpen, was das finale stratifizierte Prognosedesign motiviert. Mehrere Kandidatenmodelle wurden verglichen, während das beste Gesamtergebnis mit einem stratifizierten k-Nearest-Neighbours-Ensemble erzielt wurde. In der finalen Evaluation, über wiederholte und randomisierte Testmengen, erreichte das Modell eine mittlere aggregierte Holdout-Spearman-Rangkorrelation von 0.24, einen mittleren absoluten Fehler von 7.46 % gegenüber den auf Haushaltsebene erzielten Einsparungslabels aus dem gematchten DiD-Schritt sowie eine positive Rangkorrelation in 82 % der Holdout-Splits.

Die durchschnittlich erzielten Einsparungen blieben in allen wiederholten Durchläufen positiv. Der mittlere geschätzte Stromeinspareffekt betrug 6.92 %, was mittleren tatsächlich erzielten jährlichen Einsparungen von etwa 876 kWh beziehungsweise CHF 176 entspricht, und lag über die Durchläufe hinweg zwischen 4.44 % und 8.75 %. Auch auf Haushaltsebene blieb der Zusammenhang zwischen erwarteten und erzielten Einsparungen positiv: Die Spearman-Rangkorrelation erreichte 0.27 ($p = 0.033$). Wichtiger für die Priorisierung ist, dass die nach prognostizierten Einsparungen geordneten Haushalte im obersten vorhergesagten Terzil im Durchschnitt 9.71 % Einsparungen erzielten, verglichen mit 6.92 % in der gesamten behandelten Stichprobe und 4.36 % im untersten vorhergesagten Terzil. Dies entspricht einer Differenz von 5.35 %. Zum Vergleich: Die tatsächlichen unteren und oberen Terzile der erzielten Einsparungen lagen im Mittel bei –3.48 % beziehungsweise 16.68 %. Das vorhergesagte obere Terzil ist somit klar in Richtung der tatsächlich stark einsparenden Haushalte verschoben, auch wenn es die Trennung des wahren oberen Terzils nicht vollständig reproduziert. Die Ex-ante-Prognosen des Modells entsprechen unter den hier verwendeten zeitvariablen Tarifsannahmen mittleren erwarteten jährlichen Einsparungen von etwa 981 kWh beziehungsweise CHF 195. Dies sind die durchschnittlichen Einsparsignale, die das Modell den behandelten Haushalten vor dem Beratungstermin zugewiesen hätte, während der Ex-post-Benchmark aus dem gematchten DiD-Ansatz mittlere tatsächlich erzielte Einsparungen von etwa 876 kWh beziehungsweise CHF 176 impliziert. Die Differenz beträgt damit im Mittel etwa 105 kWh beziehungsweise CHF 19. Die Ex-ante-Schätzung liegt also rund 12 % über dem Ex-post-Benchmark.

Insgesamt zeigen die Ergebnisse, dass Smart-Meter-Daten ein praktisches Screening-Instrument zur Priorisierung von Haushalten unterstützen können. Die Evaluationsstichprobe bleibt jedoch begrenzt, sodass die absoluten Einsparwerte und die ökonomische Interpretation mit entsprechender Vorsicht zu betrachten sind. Auch die Analysen zu Konfigurationsänderungen und Einflussparametern bleiben explorativ. Dennoch geben die wiederholten Multi-Seed-Ergebnisse Vertrauen, dass die zentrale Rangprognose ein reales und praktisch nutzbares Signal erfasst, um Haushalte mit vergleichsweise hohem Einsparpotenzial bereits vor einem Beratungstermin zu identifizieren.

Résumé

Ce rapport est un deuxième ajout, Work Package 2, au projet global «KI in der Wärmepumpenberatung», numéro de projet BFE SI/502257. L'objectif de cet ajout était d'estimer, avant qu'une visite de conseil sur place n'ait lieu, quels ménages sont susceptibles de réaliser des économies d'électricité particulièrement élevées grâce à un service de pompe à chaleur et à une optimisation des réglages. Ce travail étend donc l'analyse rétrospective précédente des économies en ajoutant une couche de prédiction prospective basée sur les données de compteurs intelligents de 15 minutes antérieures à la visite. Le projet plus large et le contexte des données sont décrits par Brudermueller et al. [1].

La méthodologie finale combine deux composantes. Premièrement, les économies réalisées sont estimées à l'aide d'un modèle de différence des différences apparié et corrigé de la température. Deuxièmement, un modèle de prédiction classe les ménages selon les économies attendues. L'échantillon d'évaluation traité contient 63 ménages, comprenant 40 pompes à chaleur à air et 23 pompes à chaleur géothermiques, ce qui motive la conception de la prédiction stratifiée finale. Plusieurs modèles candidats ont été comparés, et le meilleur résultat global a été obtenu avec un ensemble stratifié de k plus proches voisins. Dans l'évaluation finale sur des ensembles de test répétés et randomisés, le modèle a obtenu une corrélation de rang de Spearman agrégée moyenne sur les données exclues de 0,24, une erreur absolue moyenne de 7.46 % points par rapport aux étiquettes d'économies réalisées au niveau des ménages de l'étape de DiD appariée, et une corrélation de rang positive dans 82 % des divisions de données exclues.

Les économies moyennes réalisées sont restées positives dans toutes les répétitions. L'effet moyen estimé de l'économie d'électricité était de 6.92 %, ce qui correspond à des économies annuelles moyennes réalisées d'environ 876 kWh ou 176 CHF, et variait de 4.44 % à 8.75 % d'une répétition à l'autre. Sur une base individuelle par ménage, les économies attendues et réalisées sont également restées positivement associées, la corrélation de rang de Spearman a atteint 0,27 ($p = 0.033$). Plus important encore pour la hiérarchisation, lorsque les ménages ont été ordonnés selon les économies prédites, le tiers supérieur prédit a réalisé 9.71 % d'économies en moyenne, contre 6.92 % pour l'ensemble de l'échantillon traité et 4.36 % pour le tiers inférieur prédit, soit un écart de 5,35 points de pourcentage. À titre de référence, les tiers inférieur et supérieur des économies réelles réalisées étaient en moyenne de −3.48 % et 16.68 %, respectivement, de sorte que le tiers supérieur prédit est clairement orienté vers les ménages réalisant de réelles économies élevées même s'il ne retrouve pas entièrement la séparation du tiers supérieur réel. Les prédictions ex ante du modèle correspondent à des économies annuelles attendues moyennes d'environ 981 kWh ou 195 CHF selon les hypothèses de tarification basées sur le temps d'utilisation ici appliquées. Ce sont les signaux d'économies moyens avant la visite que le modèle aurait attribués aux ménages traités, tandis que la référence ex post de la DiD appariée implique des économies moyennes réalisées d'environ 876 kWh ou 176 CHF. L'écart est donc d'environ 105 kWh ou 19 CHF en moyenne, ce qui signifie que l'estimation ex ante est environ 12 % plus élevée que la référence ex post.

Dans l'ensemble, les résultats montrent que les données des compteurs intelligents peuvent servir de base à un outil de sélection pratique pour la hiérarchisation des ménages. L'échantillon d'évaluation reste modeste, et les économies absolues ainsi que l'interprétation économique doivent donc être considérées avec la prudence requise. Les analyses de changement de configuration et d'influence restent également exploratoires. Néanmoins, les résultats répétés basés sur plusieurs graines initiales confirment que la prédiction centrale de classement capture un signal réel et utile en pratique pour identifier les ménages ayant un potentiel d'économies comparativement élevé avant une visite de conseil.

Main findings

- Smart-meter data contain enough information to prioritize households with comparatively high expected savings before an advisory visit takes place.
- The selected prediction model is a stratified k-nearest-neighbours ensemble and provides the best balance between ranking quality and stability among the tested alternatives.
- The economic interpretation is promising but heterogeneous. Under the tariff assumptions used here, the median one-year break-even advisory-fee threshold is CHF 169.
- Configuration-change and household-driver analyses provide useful hypotheses, but they are not yet robust enough to justify causal conclusions for individual interventions.

Zentrale Ergebnisse

- Smart-Meter-Daten enthalten ausreichend Informationen, um Haushalte mit vergleichsweise hohen erwarteten Einsparungen bereits vor einem Beratungstermin zu priorisieren.
- Das ausgewählte Prognosemodell ist ein stratifiziertes k-Nearest-Neighbours-Ensemble und bietet unter den getesteten Alternativen den besten Kompromiss zwischen Rankingqualität und Stabilität.
- Die ökonomische Interpretation ist vielversprechend, aber heterogen. Unter den hier verwendeten Tarifsannahmen liegt der mediane Schwellenwert für eine innerhalb eines Jahres kostendeckende Beratungsgebühr bei CHF 169.
- Analysen zu Konfigurationsänderungen und haushaltsspezifischen Einflussfaktoren liefern nützliche Hypothesen, sind jedoch noch nicht robust genug, um kausale Schlussfolgerungen zu einzelnen Interventionen zu rechtfertigen.

Principaux résultats

- Les données des compteurs intelligents contiennent suffisamment d'informations pour hiérarchiser les ménages ayant des économies attendues comparativement élevées avant qu'une visite de conseil n'ait lieu.
- Le modèle de prédiction sélectionné est un ensemble stratifié de k plus proches voisins, et il offre le meilleur équilibre entre la qualité du classement et la stabilité parmi les alternatives testées.
- L'interprétation économique est prometteuse mais hétérogène. Selon les hypothèses tarifaires appliquées ici, le seuil médian de rentabilité sur un an pour les frais de conseil est de 169 CHF.
- Les analyses des changements de configuration et des facteurs déterminants au niveau des ménages fournissent des hypothèses utiles, mais elles ne sont pas encore assez robustes pour justifier des conclusions causales pour des interventions individuelles.

Contents

Abbreviations

ATE average treatment effect

CI confidence interval

DiD difference-in-differences

HP heat pump

kNN k-nearest neighbours

LOO leave-one-out

MLP multilayer perceptron

MAE mean absolute error

MAPE mean absolute percentage error

RMSE root mean squared error

SFOE Swiss Federal Office of Energy

WP work package

1 Introduction

1.1 Background and motivation

Heat pumps play a central role in the decarbonization of residential heating, but their efficiency in everyday operation depends strongly on correct configuration and continued optimization. Many households can therefore benefit from advisory visits in which an expert inspects and adjusts settings such as heating curves, heating limits, or night setback. Since advisory capacity is limited and visits are costly, the practical question is which households should be addressed first.

The central objective of this addition to the project "KI in der Wärmepumpenberatung" was therefore not only to confirm that advisory visits reduce electricity consumption on average, but also to determine whether pre-visit smart-meter data can be used to identify the households that are most likely to benefit. If such a signal is sufficiently reliable, it can improve the allocation of advisory resources and provide more informative communication with households before a visit takes place.

1.2 Objective of the project

This Work Package 2 aimed to develop and evaluate a method for predicting the savings potential of individual households before an advisory visit. The intended practical benefits were the prioritization of high savings-potential households, the economic interpretation of expected savings, the comparison of expected and achieved savings, and an exploratory analysis of configuration changes and household characteristics associated with higher or lower savings.

2 Approach and Methodology

This section describes the data basis, the construction of the predictive features, and the estimation and evaluation procedures used in the work package.

2.1 Data basis

The analysis relies on the heat pump (HP) advisory dataset published by Brudermueller et al. [1], which originates from "Additional Work: Work Package 1" of the final BFE report for the project "KI in der Wärmepumpenberatung" [2] (see section 3.5 on future work). For the present work package, four data components were combined at household level, namely smart-meter electricity measurements at 15-minute resolution, hourly outdoor-temperature data that can be aligned to the same timestamps, household metadata from the survey and project administration, and advisory protocols documenting visit dates and technical observations from the on-site consultation.

Before the final quality filtering, the assembled data set contained 31,834,560 quarter-hour electricity records from 307 households, namely 68 treatment households and 239 control households. Treatment households are those with a documented advisory visit. Control households do not receive such a visit and are used to approximate the background evolution of electricity demand in the absence of treatment. Because the target of the prediction model is the achieved saving effect of a treated household, only treatment households enter the supervised learning stage, whereas the control households are essential for the upstream causal-estimation step and for placebo checks.

The final evaluation sample was obtained by applying strict completeness filters. Treatment households had to have exactly one advisory visit, a known visit date, valid 15-minute smart-meter data, and at least one full year of observations both before and after the visit. Control households had to provide at least two full years of valid data so that an analogous split into a pre and post period was possible. Electricity consumption was normalized by living area, yielding comparable quarter-hour consumption values in kWh per 100 m^2. After the final household-level filtering, the matched-DiD evaluation used 63 treated households and 233 control households. Within the treated sample, 40 households use air-source and 23 use ground-source heat pumps.

For treatment households, the split between "before" and "after" is given by the true advisory date recorded in the protocol. For control households, a virtual visit date was assigned. This date was sampled from the empirical distribution of treatment-visit dates, but only among dates that leave a full 365-day observation window on both sides within the control household's available data. If no such date was feasible, the midpoint of the available time series was used instead. This design keeps the control group seasonally comparable to the treated group while preserving symmetric yearly windows before and after the split date.

In addition to the smart-meter and weather series, the metadata and protocol tables provide contextual descriptors of the households and systems. These include living area, building type, number of housing units, construction-year interval, renovation status, number of residents, heat-pump installation type and heating capacity, distribution-system indicators such as radiators or floor heating, and domestic-hot-water-related indicators such as production system, circulation, and sterilization settings. These context variables were mainly used for descriptive and exploratory analyses and are therefore not included in the final predictive feature vector. In particular, several protocol-derived variables are only observed during the advisory visit, whereas the final prediction model was intentionally restricted to pre-visit smart-meter signatures. Living area was used upstream for normalization, and heat-pump type was used only for subgroup stratification.

2.2 Construction of prediction features

To ensure that the task is genuinely predictive, all model inputs were derived exclusively from the pre-visit year. Let $c_{i,d,t}$ denote the quarter-hour electricity consumption of household i on day d and quarter-hour slot $t \in \{1, \dots, 96\}$, after normalization by living area. The first feature family is the normalized mean daily load profile,

$$\bar{c}_{i,t} = \frac{1}{D_i} \sum_{d=1}^{D_i} c_{i,d,t}, \qquad \tilde{c}_{i,t} = \frac{\bar{c}_{i,t}}{\frac{1}{96} \sum_{s=1}^{96} \bar{c}_{i,s}}, \tag{1}$$

where D_i is the number of available pre-visit days which is fixed to 365 days in our analysis. The vector $(\tilde{c}_{i,1}, \dots, \tilde{c}_{i,96})$ captures the typical intraday load shape of a household independently of its absolute consumption level.

The second feature family is a compact scalar summary representation derived from nine interpretable smart-meter descriptors:

- `daily_cv`: coefficient of variation of the daily total consumption over the pre-visit year,
- `gini_load`: Gini coefficient of the daily totals, measuring how unevenly the annual load is distributed across days,
- `morning_day_ratio`: mean normalized load between 06:00 and 10:00 divided by the mean normalized load between 10:00 and 18:00,
- `morning_peak_above_base`: maximum normalized load between 06:00 and 10:00 minus the mean normalized load between 00:00 and 02:00,
- `cold_warm_ratio`: ratio of average daily consumption on cold days (mean outdoor temperature below 0 ℃) to average daily consumption on mild days (between 5 ℃ and 15 ℃),
- `night_day_ratio`: mean normalized load between 00:00 and 06:00 divided by the mean normalized load between 10:00 and 18:00.
- `temp_r2`: coefficient of determination R^2 from a household-level linear regression of quarter-hour consumption on outdoor temperature,
- `temp_sensitivity`: slope coefficient from the same regression, quantifying how strongly load changes with temperature,
- `weekday_weekend_shape_diff`: mean absolute difference between the normalized average weekday and weekend profiles,

The final stratified predictor uses these feature families differently for the two technical subgroups. For air-source heat pumps, the scalar summary vector gave the more stable representation and was therefore used on its own. For ground-source heat pumps, the final model used the concatenation of the 96-dimensional normalized load profile and the scalar summary vector. This reflects the empirical finding that the detailed shape of the daily profile carries additional signal for ground-source systems, whereas a more compact summary was sufficient for air-source systems. Apart from living-area normalization and heat-pump-type stratification, the additional survey and protocol variables listed above were not direct inputs to the final selected predictor.

Defining the above, let $T_{i,d,t}$ denote the household-aligned outdoor temperature, let $E_{i,d} = \sum_{t=1}^{96} c_{i,d,t}$ denote the daily total consumption, let $\bar{T}_{i,d} = \frac{1}{96}\sum_{t=1}^{96} T_{i,d,t}$ denote the daily mean temperature, let $E_{i,(d)}$ denote the sorted daily totals, and let $\tilde{c}_{i,t}^{\mathrm{wd}}$ and $\tilde{c}_{i,t}^{\mathrm{we}}$ denote the normalized mean weekday and weekend profiles defined analogously to $\tilde{c}_{i,t}$. Furthermore, let $(\hat{\alpha}_i, \hat{\beta}_i)$ denote the household-specific ordinary-least-squares coefficients from

$$(\hat{\alpha}_i, \hat{\beta}_i) = \arg\min_{\alpha,\beta} \sum_{d,t} \left(c_{i,d,t} - \alpha - \beta T_{i,d,t}\right)^2 . \tag{2}$$

The nine scalar inputs retained in the final predictor are then:

- `morning_day_ratio`: $\dfrac{\frac{1}{16}\sum_{t=25}^{40} \tilde{c}_{i,t}}{\max\left(\frac{1}{32}\sum_{t=41}^{72} \tilde{c}_{i,t}, 10^{-3}\right)}$
- `daily_cv`: $\dfrac{\mathrm{sd}(E_{i,d})}{\max(\mathrm{mean}(E_{i,d}), 10^{-3})}$
- `gini_load`: $\dfrac{2\sum_{d=1}^{D_i} d\, E_{i,(d)}}{D_i \sum_{d=1}^{D_i} E_{i,(d)}} - \dfrac{D_i + 1}{D_i}$
- `morning_peak_above_base`: $\max\limits_{25 \le t \le 40} \tilde{c}_{i,t} - \dfrac{1}{8}\sum_{t=1}^{8} \tilde{c}_{i,t}$
- `cold_warm_ratio`: $\dfrac{\mathrm{mean}_{d:\bar{T}_{i,d}<0} E_{i,d}}{\max\left(\mathrm{mean}_{d:5\le\bar{T}_{i,d}<15} E_{i,d}, 10^{-3}\right)}$
- `temp_r2`: $1 - \dfrac{\sum_{d,t}\left(c_{i,d,t} - \hat{\alpha}_i - \hat{\beta}_i T_{i,d,t}\right)^2}{\sum_{d,t}\left(c_{i,d,t} - \mathrm{mean}_{d,t}(c_{i,d,t})\right)^2}$
- `temp_sensitivity`: $\hat{\beta}_i$, i.e. the slope of the pre-visit OLS regression above
- `weekday_weekend_shape_diff`: $\dfrac{1}{96}\sum_{t=1}^{96} \left|\tilde{c}_{i,t}^{\mathrm{wd}} - \tilde{c}_{i,t}^{\mathrm{we}}\right|$
- `night_day_ratio`: $\dfrac{\frac{1}{24}\sum_{t=1}^{24} \tilde{c}_{i,t}}{\max\left(\frac{1}{32}\sum_{t=41}^{72} \tilde{c}_{i,t}, 10^{-3}\right)}$

In the implementation, `cold_warm_ratio` defaults to 1.5 if fewer than five cold and five mild days are available.

3 Results

This section reports the main quantitative findings of the work package together with the additional exploratory and robustness-analyses that help interpret the final approach.

3.1 Estimation of achieved savings

Achieved savings were estimated with a temperature-corrected matched DiD design, following standard DiD logic [3]. The idea of DiD is to measure not just whether treated households' consumption changed after the advisory visit, but whether they changed more than a comparable control group over the same period. This is important because electricity demand also moves for reasons unrelated to the advisory visit, such as weather, price, seasonality, or broader behavioural trends.

Using the household-specific pre-visit OLS coefficients $(\hat{\alpha}_i, \hat{\beta}_i)$ introduced above, the pre-visit and post-visit quarter-hour series are temperature-adjusted as

$$c^{\mathrm{adj}}_{i,d,t} = c_{i,d,t} - (\hat{\alpha}_i + \hat{\beta}_i T_{i,d,t}) + \bar{c}^{\mathrm{pre}}_i, \tag{3}$$

where $\bar{c}^{\mathrm{pre}}_i$ denotes the mean pre-visit consumption of household i. The before-versus-after change is then computed on the average adjusted consumption and compared with the corresponding change in the matched control group:

$$\mathrm{DiD}_i = \Delta y^{\mathrm{treat,adj}}_i - \overline{\Delta y^{\mathrm{ctrl,adj}}}_i. \tag{4}$$

To make the two components explicit, define the adjusted pre-visit and post-visit means as

$$\bar{c}^{\mathrm{pre,adj}}_i = \frac{1}{|\mathcal{P}^{\mathrm{pre}}_i|} \sum_{(d,t)\in\mathcal{P}^{\mathrm{pre}}_i} c^{\mathrm{adj}}_{i,d,t}, \qquad \bar{c}^{\mathrm{post,adj}}_i = \frac{1}{|\mathcal{P}^{\mathrm{post}}_i|} \sum_{(d,t)\in\mathcal{P}^{\mathrm{post}}_i} c^{\mathrm{adj}}_{i,d,t}, \tag{5}$$

where $\mathcal{P}^{\mathrm{pre}}_i$ and $\mathcal{P}^{\mathrm{post}}_i$ denote the pre-visit and post-visit quarter-hour observations of household i. The treated-household reduction is then

$$\Delta y^{\mathrm{treat,adj}}_i = 100\,\frac{\bar{c}^{\mathrm{pre,adj}}_i - \bar{c}^{\mathrm{post,adj}}_i}{\bar{c}^{\mathrm{pre,adj}}_i}, \tag{6}$$

and the matched control benchmark is

$$\overline{\Delta y^{\mathrm{ctrl,adj}}}_i = \frac{1}{|\mathcal{M}_i|} \sum_{j\in\mathcal{M}_i} 100\,\frac{\bar{c}^{\mathrm{pre,adj}}_j - \bar{c}^{\mathrm{post,adj}}_j}{\bar{c}^{\mathrm{pre,adj}}_j}, \tag{7}$$

where $\mathcal{M}_i$ denotes the set of control households whose assigned visit month lies within ± 2 months of the visit month of treated household i. In the implementation, if fewer than three such controls are available, this matched control mean is replaced by the overall mean reduction across all control households. In words, the method asks: how much more did treated households reduce their energy consumption than one would have expected from the background trend alone? The credibility of this comparison depends on a parallel-trends type assumption: after temperature correction and calendar matching, the treated and control groups should evolve similarly in the absence of the advisory visit.
The average treatment effect is then given by

$$\widehat{\mathrm{ATE}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{DiD}_i. \tag{8}$$

After filtering, the final evaluation covered 63 treated households and 233 control households. The analysis was repeated across eight complete reruns, each with a fresh random assignment of virtual visit dates to control households, to assess robustness. Across these eight reruns, the seed-level average treatment effect was 6.92 %, with a standard deviation of 1.42 percentage points and a range from 4.44 % to 8.75 %. This confirms a positive average effect, while also showing that the estimate remains sensitive to the assignment of virtual visit dates on the control side.
Figure 1 provides a complementary diagnostic view at the treated-household level. It compares the distribution of simple pre/post percentage changes before and after the household-level temperature correction, that is, before the matched control subtraction is applied. The figure therefore does not show the final DiD estimate

itself, but it makes transparent how the weather adjustment reshapes the raw treated-household change distribution that underlies the upstream savings labels.

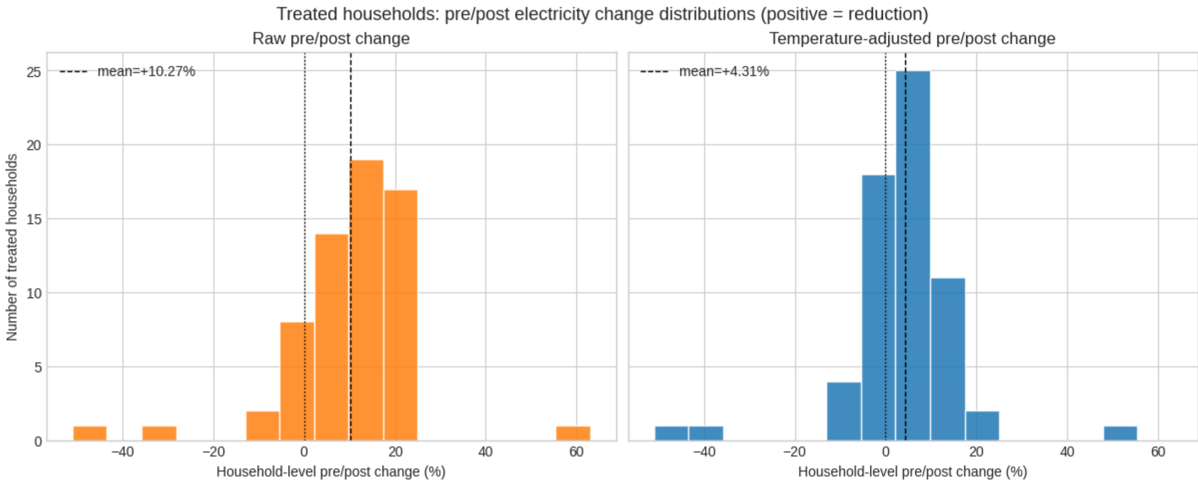

Figure 1: Distribution of household-level pre/post electricity changes within the treated sample, shown before and after the household-level temperature correction. Positive values indicate lower post-visit consumption. This is a diagnostic view of the treated-household before/after change only; the matched control subtraction used in the final DiD estimate is not yet applied.

- Across eight complete reruns, the advisory visit reduced electricity consumption by about 7 % on average.
- The savings effect remained positive in every rerun, which supports the existence of a real average saving effect.
- The exact effect size remains sensitive to the virtual control-date assignment and should therefore be interpreted with appropriate caution.

3.2 Prediction of savings potential

3.2.1 Prediction models

The prediction task used pre-visit smart-meter features only. Several candidate models were compared, including a single-neighbourhood k-nearest-neighbours baseline, a multi-neighbourhood k-nearest-neighbours ensemble, ridge regression, and a simple linear blend. The final model was a stratified k-nearest-neighbours ensemble.
Each household is represented by a feature vector x_i derived from its pre-visit load profile. To predict savings for a target household, the algorithm searches for the k most similar households in feature space and averages their observed savings. Similarity is measured by Euclidean (L_2) distance after feature standardization so that variables with large numerical ranges do not dominate the distance calculation. In the present implementation, the prediction for a household with feature vector x can be written as

$$\hat{y}^{(k)}(x) = \frac{\sum_{j \in \mathcal{N}_k(x)} w_j y_j}{\sum_{j \in \mathcal{N}_k(x)} w_j}, \qquad w_j = \frac{1}{d(x, x_j) + \varepsilon}, \tag{9}$$

where $\mathcal{N}_k(x)$ is the set of the k nearest neighbours, $d(\cdot, \cdot)$ is the Euclidean (L_2) distance between standardized feature vectors, y_j is the observed saving of neighbour j, and ε is a small constant to avoid division by zero. Neighbours that are more similar therefore receive more weight than neighbours that are further away.

The final model treats air-source and ground-source heat pumps separately and averages predictions across neighbourhood sizes 3, 5, 7, and 9:

$$\hat{y}^{\text{ens}} = \frac{1}{4} \sum_{k \in \{3,5,7,9\}} \hat{y}^{(k)}. \tag{10}$$

This predictor can be viewed as a simple model-averaging variant of distance-weighted k-nearest neighbours (kNN): instead of committing to one neighbourhood size, it averages the predictions for $k \in \{3, 5, 7, 9\}$, which reduces sensitivity to a single arbitrary choice of k [4]. Because these neighbourhoods are nested, the closest neighbours contribute to several component predictors and therefore receive more total influence than more distant neighbours, although this is not a simple four-versus-one multiplicity because the distance weights are renormalized within each k. Equivalently, the ensemble can be interpreted as a distance-weighted kNN with an additional rank-dependent weighting over neighbour order. The stratification by heat-pump type reflects the observation that air-source and ground-source systems differ in both operating patterns and achievable savings, so that "similar" households should first be searched within the same technical subgroup. More concretely, the air-source model uses only the nine retained scalar descriptors from the summary vector introduced above. By contrast, the ground-source model uses the same nine scalar descriptors together with the full 96-dimensional normalized mean daily load profile $(\tilde{c}_{i,1}, \ldots, \tilde{c}_{i,96})$. Thus, both models rely exclusively on pre-visit smart-meter-derived inputs, with outdoor temperature entering only through the weather-related summary features rather than as a separate predictor block.

Ridge regression was included as a linear benchmark, again using standardized feature vectors [4]. In ridge regression, the coefficient vector β is estimated by minimizing

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - x_i^{\top} \beta \right)^2 + \lambda \|\beta\|_2^2 \right\}, \tag{11}$$

where $\lambda \geq 0$ controls the strength of the penalty. The additional L_2 penalty shrinks large coefficients toward zero and stabilizes estimation when predictors are correlated or when the sample is small relative to the feature dimension. In the present analysis, λ was selected within each training split by `RidgeCV` over a logarithmically spaced grid from 10^{-3} to 10^{3}; held-out test households were used only for the final evaluation. Ridge therefore serves as a useful smooth and interpretable baseline, but unlike kNN it imposes a global linear relationship between the features and the savings outcome.

3.2.2 Evaluation metrics

Prediction quality was assessed mainly with Spearman rank correlation [5], because the practical objective is to rank households by expected savings rather than to predict exact point values. Spearman's ρ is the correlation between the ranks of predicted savings and the ranks of achieved savings:

$$\rho_S = \operatorname{corr}(\operatorname{rank}(\hat{y}), \operatorname{rank}(y)). \tag{12}$$

If there are no ties, it can also be written as

$$\rho_S = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}, \tag{13}$$

where d_i is the difference between the predicted and achieved rank of household i. A value of $\rho_S = 1$ means perfect ordering, $\rho_S = 0$ means no monotonic association, and negative values imply that the model tends to rank households in the wrong direction. In the result tables below, ρ refers to this Spearman rank-correlation quantity. This metric is more appropriate than R^2 for the present application because the practical goal is targeting and prioritization, not exact household-level point prediction.
For completeness, point-prediction error could also be summarized by mean absolute percentage error (MAPE). The root mean squared error (RMSE) and mean absolute error (MAE) are computed on the household-level matched-DiD savings labels and are therefore reported in percentage points. By contrast, MAPE expresses the absolute prediction error relative to the magnitude of the achieved household-level matched-DiD saving and would therefore be reported in percent. However, because some achieved DiD savings in our sample are close to zero or negative, MAPE becomes unstable and mechanically large. It is therefore not reported here, whereas MAE is the more interpretable absolute-error summary.

3.2.3 Evaluation strategy

Because the achieved-savings labels themselves depend on the upstream DiD estimation, and that step varies across random control-date assignments, the full prediction workflow was rerun eight times. In each rerun, the household-level achieved-savings outcomes were first recomputed, the pre-visit feature vectors were then reconstructed, and only afterward were the candidate prediction models trained and evaluated on the treated households. Two evaluation views were used. In the leave-one-out (LOO) setting, each household was predicted once using all remaining households as the reference set. In the aggregated-holdout setting, 50 random train-test splits were drawn, using a test share of 20% within each heat-pump stratum. More specifically, the air-source households were randomly partitioned into an air-source training and test subset, and the ground-source households were partitioned analogously. The final training set and test set for that split were then formed by combining the two stratum-specific subsets. Thus the subgroup composition is preserved, but the specific households assigned to train and test vary from split to split. For each split, the model was fitted on the training households only and used to predict the held-out households. Because a household can appear in the test set in several splits, its final aggregated-holdout prediction is the mean of all out-of-sample predictions obtained for that household. Spearman ρ, MAE, and RMSE are then computed from these aggregated household-level predictions, while the share of positive holdout splits records how often the split-specific Spearman correlation was above zero. This repeated random repartitioning of the train and test sets should not be confused with the later permutation-based p-values, which are computed after prediction by reshuffling the final prediction-outcome pairing under the null hypothesis of no rank association. This aggregated hold-out setting is more conservative and closer to the practical screening use case, because it combines repeated out-of-sample testing with smaller training sets and averages predictions over many random partitions instead of relying on a single near-full-sample fit per household.

3.2.4 Selected Model Results in the Aggregated-Holdout Evaluation

Under this evaluation scheme, the selected stratified k-nearest-neighbours ensemble achieved the best overall compromise between predictive signal and robustness across repeated reruns. In the aggregated-holdout evaluation, it reached a mean Spearman rank correlation of **0.24**, which indicates a modest but positive ability to rank households by savings potential. The mean MAE was **7.46 percentage points** and the mean RMSE was **12.03 percentage points**. MAPE is not reported, because households whose achieved DiD savings are close to zero or negative would make this percentage error unstable and mechanically hard to interpret. Household-level point predictions thus remain fairly noisy even though the ranking signal is good. The rank correlation was positive in **82% of holdout splits**, which shows that the ordering signal persisted in most repeated train-test partitions. Across the eight full reruns, the analytic Spearman p-values, which test the null hypothesis of no rank association between predicted and achieved savings, ranged from 0.025 to 0.141. The corresponding permutation-based p-values, obtained by repeatedly reshuffling the prediction-outcome pairing and recomputing the rank correlation, ranged from 0.014 to 0.071 and were below 0.05 in **six reruns**. It is important to keep in mind here, that a p-value is the probability of observing a result at least this extreme if the null hypothesis were true, it is not the probability that the null hypothesis itself is true. Figure 2 summarizes the comparison across the main candidate prediction methods, and Table 1 reports the final repeated-evaluation metrics for the selected model. Taken together, these results are consistent with a real but modest prioritization signal: the model is useful for ranking households, but not accurate enough for household-level guarantees.

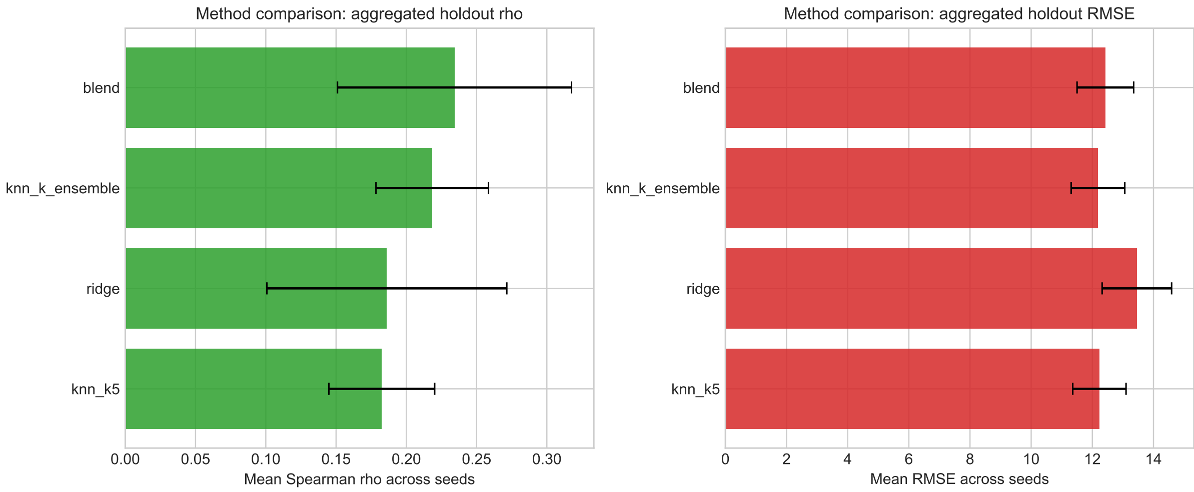

Figure 2: Comparison of the main prediction methods considered in the final model-selection stage. In the plot, `knn_k5` denotes a single k-nearest-neighbours model with $k = 5$, `knn_k_ensemble` denotes the average of k-nearest-neighbours predictions for $k \in \{3, 5, 7, 9\}$, `ridge` denotes ridge regression on the same standardized feature vectors, and `blend` denotes a simple 50:50 average of the `knn_k_ensemble` and ridge predictions. The selected stratified k-nearest-neighbours ensemble offered the best overall trade-off between ranking quality and stability.

Table 1: Final repeated evaluation of the selected prediction approach.

Quantity	Value
Repeated full reruns	8
Treated households	63
Mean achieved saving effect	6.92 %
Saving effect range	[4.44 %, 8.75 %]
Mean LOO Spearman ρ	0.27
LOO Spearman p-value range	[0.008, 0.080]
LOO permutation p-value range	[0.006, 0.044]
Mean aggregated-holdout Spearman ρ	0.24
Aggregated-holdout Spearman p-value range	[0.025, 0.141]
Aggregated-holdout permutation p-value range	[0.014, 0.071]
Std aggregated-holdout Spearman ρ	0.03
Mean aggregated-holdout MAE (percentage points)	7.46
Mean aggregated-holdout RMSE (percentage points)	12.03
Mean share of positive holdout splits	82 %
Reruns with agg. permutation $p < 0.05$	6 / 8

- The final model is useful as a ranking tool, not as a household-level guarantee.
- In the conservative aggregated-holdout evaluation, the ranking signal remained positive across all eight reruns.
- The permutation-based prediction test was below 0.05 in six of the eight reruns, which is consistent with a real but modest prioritization signal.

3.3 Comparison of expected and achieved savings

Predicted savings were also compared directly with achieved savings on a household-by-household basis after averaging predictions across the repeated reruns. This yields a positive rank association between expected and achieved savings: Spearman $\rho = 0.27$ with $p = 0.033$, an RMSE of 11.63 percentage points, and an MAE of 7.08 percentage points. Importantly, when households are ordered by predicted savings, the top predicted tercile exceeds the bottom predicted tercile by 5.35 percentage points in achieved savings. Given that the overall mean achieved saving is 6.92 %, this is substantial: the top predicted tercile reaches 9.71 % achieved savings on average, compared with only 4.36 % in the bottom predicted tercile. It also captures 8 of the 21 actual top-tercile savers (38%; equivalently precision = recall = F1 = 0.38 for top-tercile identification). This result supports the practical use of the model for prioritization, even though the error level remains too high for household-level guarantees.

3.4 Economic interpretation

To address the economic objective of the work package, predicted and achieved percentage savings were translated into annual kWh and annual CHF values using household-specific annual consumption and time-of-use electricity tariffs. This yields a first-order economic interpretation that can be used to assess whether an advisory visit is likely to pay back within one year.

Mean predicted annual savings correspond to about 981 kWh or CHF 195, while mean achieved annual savings correspond to about 876 kWh or CHF 176. The expected values are obtained by converting the model's pre-visit percentage predictions for each treated household into annual kWh and CHF terms and then averaging across households; the achieved values are computed analogously from the matched-DiD savings estimates. The median one-year break-even advisory-fee threshold is CHF 169. Under the tariff assumptions used here, 88.9% of households would be predicted to break even within one year at a CHF 100 advisory fee, while 36.5% would do so at CHF 200. These values should be interpreted as an energy-cost view rather than a complete bill simulation, since additional cost components and a fixed advisory fee were not directly part of the underlying data.

> Mean expected savings correspond to roughly CHF 195 per year, mean achieved savings correspond to roughly CHF 176 per year, and the median one-year break-even threshold is CHF 169 under the tariff assumptions used in the analysis.

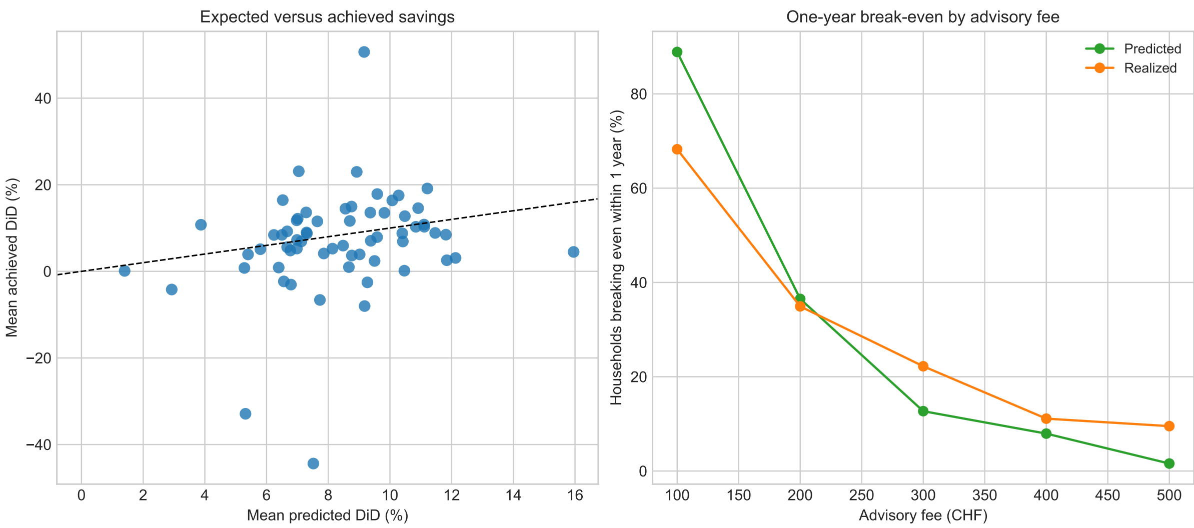

Figure 3: Comparison of expected and achieved savings, and the corresponding one-year break-even shares for different advisory-fee assumptions.

3.5 Configuration changes and influence parameters

Two further exploratory analyses were carried out to address the work-package goals on configuration effects and influence parameters. First, households with and without key intervention patterns were compared, in particular heating-curve reduction, heating-limit reduction, and changes in night setback. The reported values in this comparison denote percentage-point differences in mean achieved household savings, where achieved savings are measured by the household-level DiD outcome averaged across reruns. They are therefore subgroup contrasts within the treated sample, not separate causal treatment-effect estimates for the individual configuration changes themselves. None of these subgroup differences were statistically robust in the present sample, these results should therefore be interpreted as descriptive only. The largest positive subgroup contrast was observed for night-setback deactivation (+2.63 percentage points), meaning that households with this change had on average 2.63 percentage points higher achieved savings than households without it, while heating-curve reduction was associated with 3.76 percentage points lower achieved savings. Within the night-setback deactivation subgroup, there is no evidence that achieved savings differ systematically between air-source and ground-source heat pumps.

Second, smart-meter and protocol/context features were screened for association with savings using Spearman rank correlation. The clearest recurring signal came from load-shape features derived directly from smart-meter data. The strongest realized-savings correlate was `morning_day_ratio` with $\rho = -0.36$ and $p = 0.004$, meaning that households with a more pronounced morning load relative to daytime load tended to realize lower achieved savings. The negative sign of ρ indicates this inverse monotonic association, while the small p-value means that such an association would be unlikely to appear this strongly under the null hypothesis of no association. Larger heated floor area also showed a weak negative tendency, meaning that larger homes tended on average to realize somewhat lower achieved savings, whereas most protocol/context variables were weak. These analyses are therefore best understood as hypothesis generation for future work rather than as validated decision rules.

3.6 Other methods explored during the project

Before the final methodology was fixed, the project went through a broad exploratory phase. This was relevant because the work package required both a credible savings benchmark and a practically useful prioritization signal, so several alternative approaches were tested for comparison. Early branches included naive before-versus-after comparisons, propensity-style scoring, and panel-style DiD regressions with additional controls. These experiments helped narrow the design space and clarify which ingredients were needed for the final reported methodology.

Given the limited number of treated households, the project placed particular weight on methods that remain stable, interpretable, and hard to overfit in small samples. This made it plausible from the outset that relatively simple approaches would be difficult to outperform, but richer prediction branches were nevertheless explored to test that expectation against the data. These included alternative k-nearest-neighbours representations based on summary statistics, monthly profiles, full daily profiles, and mixed feature sets, as well as ridge and gradient-boosted direct regressors. We also tested more explicitly causal formulations, including treatment-only neural models such as direct-effect multilayer perceptron (MLP) and TARNet-style variants, counterfactual control-model approaches based on gradient boosting, uplift T-learners with Fourier features and inverse-probability weighting, and more ambitious causal-machine-learning estimators such as T-learners, X-learners, and causal-forest-style methods. In the end, these richer variants did not improve enough to justify their additional complexity: under the small treated sample, they were typically less stable, less interpretable, weaker on placebo and robustness checks, or more prone to overfitting. In that setting, the final matched-DiD plus stratified k-nearest-neighbours approach offered the most convincing balance between robustness, transparency, and operational usefulness.

> The project tested many more alternatives than this report can describe in detail, including classical regressions, several nearest-neighbour variants, boosted-tree counterfactual models, and neural causal models. This breadth strengthens confidence that the final selected method was chosen after genuine comparison rather than by default.

4 Conclusions and outlook

4.1 Key findings and limitations

An important limitation is that the achieved-savings estimate is not completely invariant across repeated reruns. The repeated evaluation showed a range from 4.44 % to 8.75 % for the average achieved saving effect estimated in the upstream DiD step. Here, the rerun seed mainly affects the random assignment of virtual visit dates on the control side in the upstream DiD step. This changes the estimated household-level achieved-savings labels and therefore also propagates into the downstream prediction metrics, although more weakly. Figure 4 shows this pattern directly: the Step 1 DiD estimate varies more strongly across seeds than the aggregated-holdout rank correlation and the aggregated-holdout RMSE (in percentage points). The prediction layer itself is more stable than this upstream savings estimate, but the remaining error level means that the model should be used as a prioritization tool rather than as a guarantee of household-level savings.

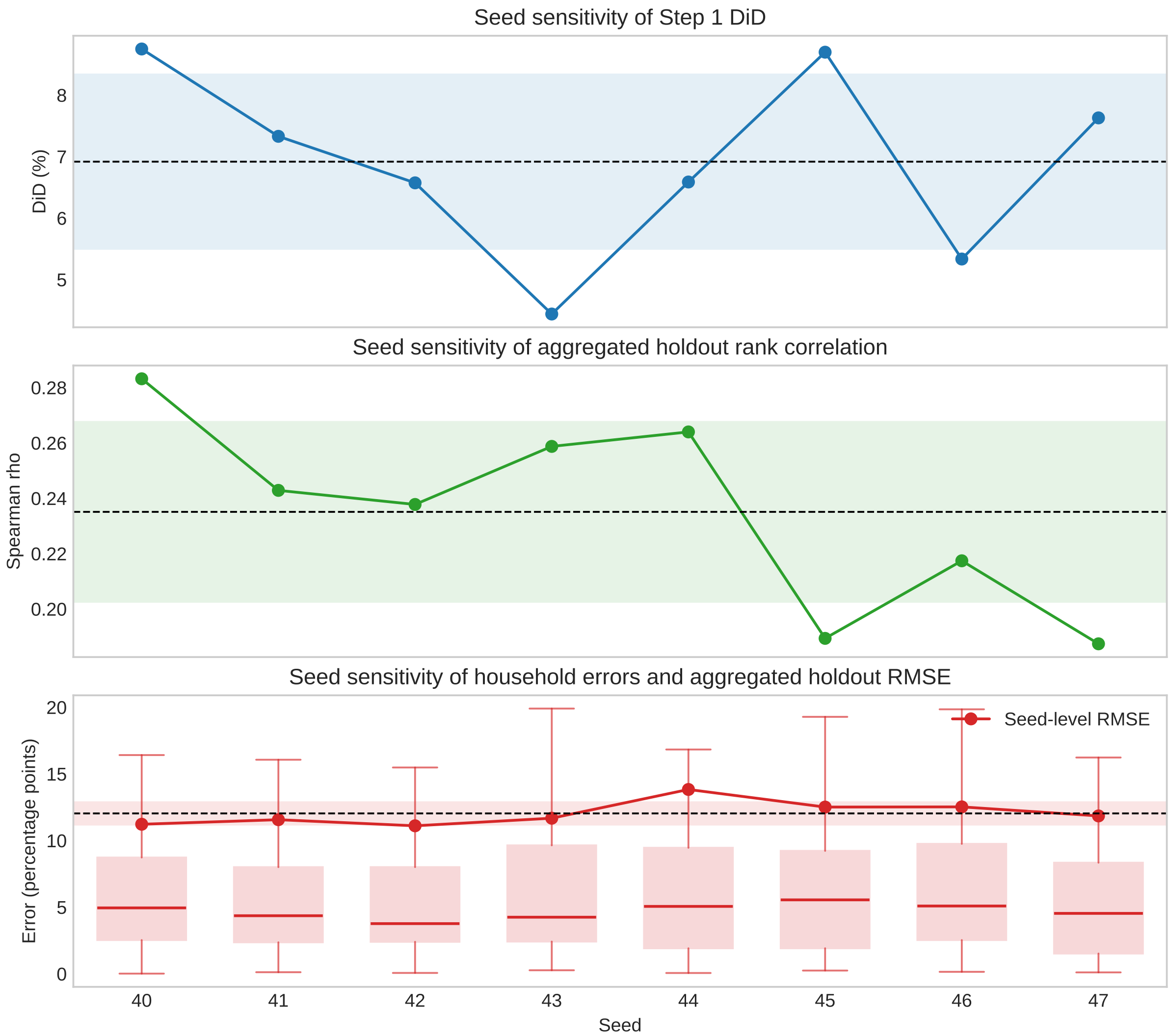

Figure 4: Robustness of the final evaluation across repeated full reruns. The achieved-savings estimate varies more strongly than the prediction metrics, which is an important limitation for interpretation. In the lower panel, the points show the seed-level aggregated-holdout RMSE, while the boxplots show the corresponding distribution of household-level absolute prediction errors within each seed.

The economic layer provides an informative estimate of annual savings and break-even thresholds, but does not include a fully specified advisory-service fee, all bill components, or a formal amortization model. Similarly, the configuration-change and influence-parameter analyses remain exploratory and should be validated on an independent future cohort before being used for operational intervention design.

4.2 Contribution to the project goals

1. **Prioritization of households with high savings potential.** Achieved. The final model provides a stable positive ranking signal and can therefore be used to identify households with comparatively high expected benefit from advisory visits.
2. **Feedback on the economic rentability of an advisory visit.** Achieved. The analysis provides annual kWh and CHF savings as well as one-year break-even thresholds. A complete household-level amortization statement could be easily adapted for real world use cases based on specific assumptions.
3. **Comparison of expected and achieved savings.** Achieved. Expected and achieved savings are compared directly on a household basis and remain positively associated.
4. **Investigation of the effect of specific configuration changes.** Achieved. The project includes exploratory subgroup analyses for key intervention types. No robust causal pattern could be established.
5. **Investigation of influence parameters of savings.** Achieved. The analysis identifies several promising smart-meter-derived correlates, but these findings remain descriptive and require a larger dataset source.

4.3 Outlook

The main conclusion of this work package is summarized as follows: household-level savings potential can be estimated sufficiently well to support prioritization before an advisory visit. The next step is to adapt the approach to specific economic assumptions and to extend the data basis. More data would strengthen statistical significance, improve representativeness, and provide a firmer basis for assessing the more exploratory findings.

> The practical implication for EKZ is not that household savings can already be predicted perfectly. Rather, a transparent model can already support better prioritization of advisory visits, while the economics and the exploratory driver analyses are refined as more households become available.

References

[1] T. Brudermueller, E. Fleisch, M. González Vayá, and T. Staake, "Heapo – an open dataset for heat pump optimization with smart electricity meter data and on-site inspection protocols," 2025.

[2] P. Gerner, J. Marckhoff, M. González Vayá, and T. Staake, "KI-WP – KI in der Wärmepumpenberatung," Schlussbericht SI/502257, Bundesamt für Energie (BFE), Bern, Switzerland, Dec. 2024.

[3] J. D. Angrist and J.-S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press, 2009.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2 ed., 2009.

[5] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.