Building Characterization through Smart Meter Data Analytics: Determination of the Most Influential Temporal and Importance-in-prediction based Features

Behzad Najafi[a,*], Monica Depalo[a], Fabio Rinaldi[a], Reza Arghandeh Joueneghani[b]

[a]*Department of Energy, Politecnico di Milano, Via Lambruschini 4, Milano 20156, Italy.*
[b]*Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen 5063, Norway.*

Abstract

The present paper aims at determining the most influential features to be extracted from smart meter data to facilitate machine learning-based classification of non-residential buildings. Smart meter-driven remote estimation of the chosen characteristics (the buildings' performance class, use type, and operation group) is significantly helpful in buildings' commissioning, benchmarking, and diagnostics applications. As the first step, state-of-the-art feature selection methods and a proposed customized approach are utilized for determining the most influential parameters in the pool of temporal features, proposed in a previous study. Next, importance-in-prediction based features, generated from an hour-ahead load prediction pipeline, that can improve the classification accuracy are proposed and added as additional input parameters. Finally, interpretations about some of the most influential features for different classification targets are provided.

The obtained results demonstrate that, while aiming at estimating the buildings' use type, through performing feature selection and adding importance-in-prediction based features, the number of utilized features is reduced from 290 (initial pool of features proposed in a previous study) to 29, while also increasing the accuracy from 71% to 74%. Similarly, number of employed features for estimating the performance class is decreased from 224 to 17 and the achieved accuracy is improved from 56% to 62%. Finally, using only 6 selected features, compared to 287 features in the initial set, the obtained accuracy for the classification of operation group is increased from 98% to 100%. It is thus demonstrated that the proposed methodology, through selecting and utilizing notably fewer features, results in a notable simplification of the feature extraction procedures, improves the achieved accuracy, and facilitates providing interpretations about the reason behind the influence of some of the most important features.

Keywords: Commercial building Characterization, Smart Meter Data Analytics, Machine Learning, Feature Extraction, Feature Selection.

*Corresponding author
Email address: behzad.najafi@polimi.it (Behzad Najafi)

Abbreviations

allDays_meanvs95thRatio_std Standard deviation of the ratio between the mean value of consumption and the corresponding 95th percentile considering all days of the week

allDays_meanvsmaxRatio_std Standard deviation of the ratio between the mean value of consumption and the maximum consumption considering all days of the week

allDays_minvs95thRatio_max Maximum of the ratio between the minimum value of consumption and the corresponding 95th percentile considering all days of the week

AreaNormalizedConsMean Mean of area normalized consumption

AreaNormalizedConsMin Minimum area normalized consumption

breakoutsNumber_i_j_k Number of breakouts where i is the minimum breakout size in days, j is the penalization level on closeness of breakout points (0 not penalized, 2 max penalization), and k represents the beta penalization threshold whose value is 0.00k

CV Cross-validation

dailyMaxVariance Maximum daily variance in consumption

dayFilterFreq_a_wh_min Minimum of DayFilter patterns where a is the alphabet size (number of equiprobable regions in which the daily consumption distribution is split) and w is the size (in hours) of the time periods in which the day is split

dayFilterFreq_a_wh_std Standard deviation of DayFilter patterns where a is the alphabet size (number of equiprobable regions in which the daily consumption distribution is split) and w is the size (in hours) of the time periods in which the day is split

eemeter_cvrmse Model fit coefficient using EEMeter

hourlyStats_maxConsHourOfDay Hour of the day with maximum consumption

hourlyStats_meanCons4hr Mean consumption at 4:00 a.m.

imp_ConsumptionX Importance-in-prediction of consumption lagged for X hours

imp_MaxUse Importance-in-prediction of the maximum consumption value in the last 24 hours

imp_meanvsmax_use24 Importance-in-prediction of the ratio between mean and maximum values of the previous 24 hours of consumption

imp_quant50vsmax_use24 Importance-in-prediction of the ratio between 50th percentile and maximum values of the previous 24 hours of consumption

$imp\_SeaLevelPressurehPaX$ Importance-in-prediction of sea level pressure lagged for X hours

$imp\_sinHour$ Importance-in-prediction of sin(hour)

imp_std_use24 Importance-in-prediction of the standard deviation of the previous 24 hours of consumption

$imp\_stdvsmax\_use24$ Importance-in-prediction of the ratio between standard deviation and maximum values of the previous 24 hours of consumption

$imp\_VisibilityKmX$ Importance-in-prediction of visibility lagged for X hours

$imp\_WindSpeedKm/hX$ Importance-in-prediction of the speed of the wind lagged for X hours

$loadshape\_mapeIntervalDaytime$ MAPE interval of prediction during day time using the loadshape model

$loadshape\_rmse\_Interval$ RMSE interval of prediction using the loadshape model

$maxDailyConsDate$ Date at which the maximum daily consumption occurs

$meta\_dateLast$ Last date of sampling

MI Mutual Information

$mostCommonHourTop10perc$ Most common hour at which the top 10% of consumption values occur

RF Random forest

RFE Recursive feature elimination

$RFECV$ Recursive feature elimination procedures that include cross validation loop

$stats\_minDailyConsDate$ Date at which the minimum daily consumption occurs

$stats\_minHourlyCons$ Minimum hourly consumption

$stats\_minHourlyConsDate$ Date at which the minimum hourly consumption occurs

$STLweeklyPatternXMean$ STL model trend mean on X weekday, X being Thu: Thursday, Fri: Friday, Sat: Saturday

$summerConsVariance$ Variance of consumption in the summer

$weekdays\_meanvs95thRatio\_min$ Minimum of the ratio between the mean value of consumption and the corresponding 95th percentile considering only weekdays

$weekdays\_meanvs95thRatio\_std$ Standard deviation of the ratio between the mean value of consumption and the corresponding 95th percentile considering only weekdays

$weekdays\_minvs95thRatio\_max$ Maximum of the ratio between the minimum value of consumption and the corresponding 95th percentile considering only weekdays

weekdays_minvs95thRatio_mean Mean of the ratio between the minimum value of consumption and the corresponding 95th percentile considering only weekdays

weekdays_minvsmaxRatio_mean Mean of the ratio between the minimum value of consumption and the maximum consumption considering only weekdays

weekdays_minvsmaxRatio_min Minimum of the ratio between the minimum value of consumption and the maximum value of consumption considering only weekdays

winterConsVariance Variance of consumption in winter

1. Introduction

The energy demand of buildings has a considerable impact on the global primary energy consumption and greenhouse gas emissions. According to a 2019 IEA report [1], buildings will play a central role in the clean energy transition as this sector is responsible for 28% of energy-related CO_2 emissions worldwide, two-thirds of which is associated with the electricity consumption. In addition, the share of electricity use in the buildings' energy demand is rapidly growing as the average incomes rise and the urban migration continues in non-OECD countries [2] and it is consequently expected to increase from 33% in 2017 to nearly 55% of total buildings' energy use in 2050 [1]. Therefore, increasing the energy efficiency and performance of buildings is a critical step towards global sustainability.

Building commissioning, and retro-commissioning in particular, is proved to have a considerable energy saving potential [3], which is also the focus of the many other energy auditing related procedures employed in this field [4, 5]. Moreover, as the majority of the existing buildings worldwide were constructed without the obligation of following mandatory energy performance related protocols and considering that these buildings will make up a considerable share of the future building stock, large-scale procedures to enhance the overall building performance should be a priority in the coming years [6].

Estimating the use type of the building, specifically while having access to the corresponding consumption profile, can help several parties including the utility companies, the grid management firm, and the public organizations (e.g. sustainability work groups) to have a proper estimation of the building's energetic performance. The latter is due to the fact that the performance metrics (e.g. yearly consumption per conditioned surface) are defined differently for various building applications (e.g. residential building, education dwellings, hospitals). As an instance, having knowledge about the performance of a large number of buildings (determining which requires knowledge about their use type) can permit public organizations to prepare the most suitable incentive programs (for improving the performance of the building) that can attract the highest number of users and to predict the impact of the corresponding energy saving interventions

in a large (e.g. national) scale.

However, performing conventional energy auditing procedures on a large number of buildings is a time-consuming and costly procedure. Furthermore, buildings types, in terms of the corresponding construction, performance, and system technologies, are often very diverse. Accordingly, a notable effort, in terms of time and economic investment, is required to perform a dedicated analysis on each specific type of building.

A promising alternative for dealing with the latter obstacle, is exploiting the enormous amount of data generated by smart electrical meters, which are already largely diffused in most of Europe (around 200 million units [7] in 2020) and USA (87 million units in 2018 [8]). It should be noted that the smart meters in different countries are installed by various organizations and, while the consumption data can be shared with other parties/organizations, it is commonly anonymized because of privacy concerns. Thus, public organizations can commonly be provided by the consumption profiles of a large number of buildings (obtained from smart meters), without having access to information about the use type of the building. Thus, facilitating the possibility of estimating the building use type while only employing the smart meter data can provide these organizations with a notable benefit.

In this context, several research works, have employed meter data analysis, for a variety of applications including load profile classification and clustering, energy disaggregation, and demand response potential estimation. Examples from the first category include studies such as the one conducted by Räsänen and Kolehmainen [9], where the use of extracted statistical features improved the clustering accuracy of electricity load curves, and the research carried out by Dasgupta et al. [10], which clustered and analyzed load curves employing elastic shape analysis, successfully discovering broad consumption patterns across different seasons and neighborhoods. In the study conducted by Najafi et al. [11], the use of non intrusive load monitoring (NILM) was investigated to classify electrical appliances with possible applications in demand prediction, mal-functioning identification and occupancy monitoring, whereas the research performed by Mathieu et al. [12] analyzed 15-min-interval electric load data for building benchmarking, demand response, peak load management, and other purposes.

An important attempt towards utilizing the smart meter data analysis for estimating the building type and performance has been conducted by Miller [13], in which the Building Data Genome Project [14] was utilized as the dataset. The latter is a large public dataset including weather and electrical meter data of several buildings along with their use categories and characteristics. In this study, an extensive investigation on extracting temporal features from the smart meter data of non-residential buildings was carried out. These features were then employed to estimate the buildings' category of use, performance index, and operation strategy. This process facilitates building characterization, which is at the basis of techniques like commissioning, benchmarking and diagnostics, while only employing the smart meter data [13]. However, making the related decisions human-interpretable is also important for the final expert's judgement. The utilization of a large number of features, despite being utilized aiming at increasing the model accuracy

and generalisability, reduces the interpretability of the results and increases the model complexity and consequently the calculation cost. Therefore, implementing a comprehensive feature selection methodology can notably reduce the complexity of the model and enhance the interpretability, while increasing the estimation accuracy.

It is noteworthy that, although a few previous studies have provided interpretations about the reason behind the possible contribution of some of the extracted temporal features to the accuracy of estimation pipelines (utilizing for building characteristics estimation), as these studies did not include the feature selection step, the extent of this influence (if any) could not be verified. Thus, the latter shortage impeded providing interpretations utilizing a set of features, the extent of contribution of which (to the achieved accuracy) is verified. The feature selection procedure that is performed in the present study permits avoiding the latter obstacle.

Previous studies have investigated the impact of feature selection methods on the models' accuracy, resulting in improved or comparable prediction performance. In particular, Zhao and Magoulès [15] analyzed the influence of feature selection on the prediction of a building energy consumption, Kapetanakis et al. [16] captured the effect of selecting input variables on thermal loads prediction of commercial buildings, while Zhang and Wen [17] proposed a feature selection procedure based on pre-processing, filtering and grouping through a wrapper method. However, in most of the cases, this process is carried out on a relatively small set of features and is commonly performed manually. Furthermore, to the authors' knowledge, no previous work has been conducted on implementing a comprehensive variable selection methodology for building characterization, for which most of the conducted studies do not employ large sets of attributes, nor extensively investigate the underlying physical behaviours. It is the case of the study conducted by Westermann et al. [18], where 27 variables are used for customer characterization, the research performed by Yang et al. [19], the result of which utilizes 6 variables for building climate zoning, and the study carried out by Piscitelli et al [20], implementing the classification of load profiles in buildings of 114 customers with 9 variables.

Motivated by the above-mentioned necessity and research gap, the present work is implemented starting from the temporal features, which are proposed and extracted in [13] from the electrical meter data and the corresponding weather dataset of several commercial buildings. Next, while considering three classification objectives (building use, performance class, and operation strategy), state-of-the-art feature selection methodologies are implemented and the corresponding results, in terms of classification performance and number of selected features, for each of the considered objectives, are compared. A customized feature selection method is then proposed and implemented and the obtained results are compared with those achieved using conventional methods. It is demonstrated that the implemented feature selection methodologies, and particularly the proposed customized method, can notably reduce the number of utilized temporal features and thus the dimension of the dataset, while even improving the classification accuracy.

In the next step, importance-in-prediction based features, extracted in the framework of a short-term load prediction pipeline, are added. These features, to the author's knowledge, have never been utilized in the previous studies for non-residential building characterization purposes. It is shown that adding these features can result in an enhancement of the obtained accuracy. The influence of each feature of the final set on the overall accuracy is then demonstrated. Finally, following the methodology proposed in a previous study [21], the selected features are analysed and interpretations about the reason behind the impact of some of the most influential features are provided.

Accordingly, the contributions of the present paper are summarized as follows:

- Besides applying state-of-the-art feature selection methods, a customized approach is proposed and implemented aiming at selecting the most influential temporal features to be extracted from smart meter data, proposed in a previous study [13], aiming at building characterisation;
- The importance-in-prediction based features are proposed and extracted, in the framework of a short-term load prediction pipeline, to enhance the building characterization accuracy;
- The final proposed pipelines provide higher accuracy while utilizing notably fewer features with respect to the corresponding initial set, which significantly simplifies the feature extraction procedure and facilitates the interpretation of the obtained results;
- Interpretations about the reason behind the influence of some of the most important features on the achieved classification accuracy are provided.

In this framework, section 2 briefly introduces the dataset and the classification objectives of the study. Section 3 presents the overall methodology including a brief explanation of the extracted features. Section 4 provides a description about the utilized machine learning algorithms, the accuracy metrics, and the employed feature selection methodologies. In section 5, the obtained results of the feature selection procedure are presented and discussed and physical interpretations of selected features are provided. Finally, section 6 presents the conclusions reached based on the obtained results.

2. Case Study

In the present work, the Building Data Genome Project [14], a public, open dataset, is employed. The dataset is composed of 507 non-residential buildings, located in the USA (New York, Los Angeles, Denver, Chicago, Phoenix), London, Zurich and Singapore, and therefore representing discreetly varied climate areas (the overall temperature ranges from -30°C to +50°C). The metadata file provides information about each building's primary space use, timezone, surface area, and its corresponding weather file name. Smart meter-derived hourly electrical consumption data for a period of at least one year is also given for every building.

The dataset is finally completed by a set of hourly weather data files for each of the locations mentioned above, including temperature, humidity, pressure, visibility and wind speed. An example of the consumption and temperature profile of a university building, provided in this dataset, is demonstrated in Fig. 1.

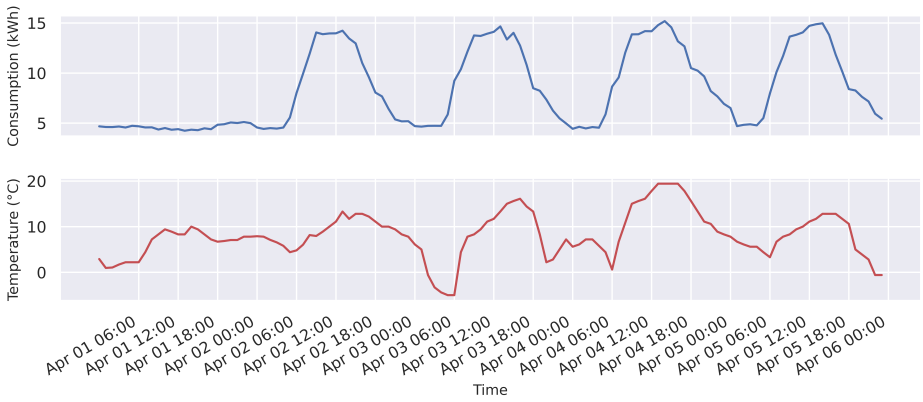


Figure 1: An example of the load and temperature profiles for one of the buildings provided in the Building Data Genome Project [14] dataset

Starting from the extraction of 315 temporal features, proposed by Miller [13], that are publicly accessible in an open source project [22], the first part of this study is focused on conducting different feature selection methods aiming at selecting the most influential features. Different categories of features, proposed in [13] that are present in the pool of features provided in [22], are summarized in Table 1. Brief descriptions about each of these categories of features are provided in sub-section 3.1. In order to ensure an effective comparison with the results obtained by Miller [13], the set of extracted features employed in the first step and the utilized classification algorithm are identical to the ones implemented in this work.

Feature Category	**Description**
Statistics-based	Application of basic statistical functions such as mean, median, maximum, minimum and standard deviation to the time series data
Regression-based	Output parameters and attributes obtained from the development and training of predicting models
Pattern-based	Frequent daily, weekly, monthly and long-term patterns extracted from the time series data

Table 1: Main categories of extracted features utilized for building characterization.

2.1. Prediction Targets

Different classification targets are set as objectives of the analysis, namely building primary use type, performance class, and operation group. The first two are identical to the ones considered in [13], while the latter is partially modified to obtain more balanced classes.

- **Principle Building Use** indicates the primary use for which a certain building was designated, in our case being either an office, a primary or secondary school classroom, a college classroom, a dormitory or a laboratory. It is possible that some spaces are utilized for multiple purposes, which complicates the classification task; thus, only the primary application of these buildings is considered as the objective [13]. Furthermore, while estimating this target, features related to in-class similarity and temporal specificity, that are generated using "jmotif" library [23] and indicate how well a certain building fits in its own class, were removed due to their dependency on the use type.

- **Performance Class** is evaluated based on the floor area-normalized consumption of buildings, which are therefore assigned to three groups of low (bottom 33% percentiles), intermediate or high (top 33% percentiles) consumption levels within each primary use category [13]. While considering this target, all features that are evidently correlated to consumption, such as area normalized consumption, "eemeter" [24] outcomes and inherent statistics, are excluded.

- **General Operation Strategy** distinguishes between different campuses and groups of buildings operated by the same authority, which can therefore have similar operation strategies. For this purpose, four distinct campuses with a comparable number of buildings were selected, excluding a few smaller campuses that were used in [13]. All variables that are indicators of weather sensitivity are removed from the set to avoid any possible relation with the location of the buildings.

Each step of the process will be applied on a dedicated set of features utilized for specific classification targets. Being able to estimate the considered targets facilitates performing techniques like commissioning, benchmarking and diagnostics. The building primary use type, as an instance, defines the benchmark used for the building's performance level assessment. Space use estimation can also be used to determine whether the building principal use type has changed over time without being recorded [13]. The performance class target is also related to the benchmarking process as it can help understanding how a building performs compared to its peers and what are the behaviours that lead to a good or poor performance. These latter evaluations can also provide useful insights for identifying critical aspects to address during the commissioning process.

In the next sections, more details about the overall implemented methodology, including the employed data pre-processing, features extraction, feature selection steps and the utilized machine learning algorithms, are provided.

3. Implemented methodology

The first step is dedicated to cleaning the dataset from invalid and missing values. Secondly, the raw temporal data, including meter and weather data, is processed utilizing multiple tools and techniques for feature extraction, aiming at achieving a comprehensive description of different phenomena. Once all the

variables are obtained and integrated together, while utilizing a chosen benchmark algorithm (Random Forest Classifier [25, 26]), different feature selection algorithms are performed. The accuracy obtained for the considered classification targets, while providing the complete set of features and different sets obtained with various feature selection methods, are then compared. Afterwards, importance-in-prediction based features, generated in the context of an hour-ahead load prediction framework, that improve the achieved accuracy, are added to the selected set of temporal features for each classification target. Finally, interpretations about the reason behind the selection of specific features for each classification target are provided.

The above mentioned steps are represented in Fig. 2.

Figure 2: Schematic representation of the implemented methodology.

3.1. Data pre-processing and feature extraction

For each building and each feature extraction technique, all invalid or missing values in the dataset are first removed. The temporal load data and the corresponding hourly weather information are next combined, depending on the variables needed for each extraction process.

In the last phase of the data processing procedure and before implementing the machine learning algorithm, each target's dataset is shuffled (to obtain an even distribution of the classes) and is then split into 50% training and 50% testing sets. In the next sub-section, brief descriptions about the main categories of temporal features, proposed in [13], are provided. Further details about the extraction of these features can be found in [13] and [22], while the distribution of the most significant variables among different classes can be observed in Fig. 9, 11 and 13.

3.1.1. Statistics-based features

The statistics-based features include temporal basic statistics such as mean, median, maximum, minimum, variance, and standard deviation, calculated on the whole time-series load vector or on shorter intervals such as the winter and summer seasons [13]. Mean and variance can be calculated using Eqs. 1 and 2 respectively, while standard deviation is defined as the square root of the variance.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{n} \tag{2}$$

Many of these features were generated through the visdom R package [27]. Variance is useful to understand how much certain values vary across a time range, whereas 97th and 3rd percentiles can be more meaningful than minimum and maximum values due to the exclusion of extreme outliers. In addition, hour and date that correspond to the consumption peak are determined and added as variables. Furthermore, a series of hour-of-day metrics, which are determined based on the time at which a particular behaviour occurs most frequently, are also extracted.

Other extracted statistics-based features include ratios of the above-mentioned statistical parameters, which can be used as a better comparison basis between different buildings, along with other normalized quantities such as the floor area-normalized consumption. An example of these is given by Fig. 3-a, which shows the daily ratio between mean consumption and maximum consumption for a selected building. Lastly, the utilized Spearman rank order correlation (ROC) coefficient indicates the correlation between the total electrical consumption of a building and the outdoor temperature in a range between -1 and +1. A highly positive correlation (+1) implies that consumption and temperature increase accordingly as in a cooling sensitive building, whereas for a heating sensitive building, consumption will tend to increase with decreasing temperatures, described by a ROC coefficient close to -1 [13]. Fig. 3-b demonstrates the determined yearly Spearman rank order correlation coefficient for a specific building.

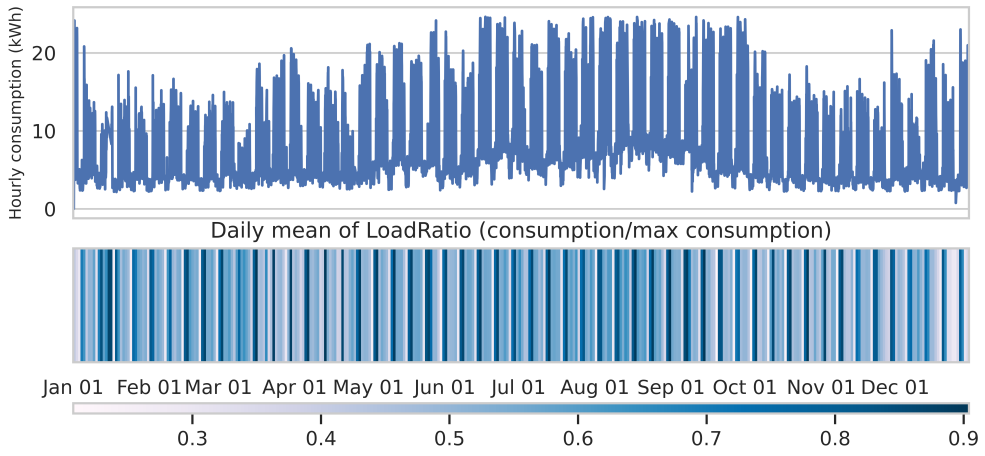
3.1.2. Pattern-based features

Pattern-based features facilitate capturing the typical (motifs) and atypical patterns (discords) in the consumption of buildings [13]. The aim of extracting these features is to understand whether a building follows some kind of daily or weekly pattern. Fig. 4-a provides a visual representation of the consumption pattern of a specific building.

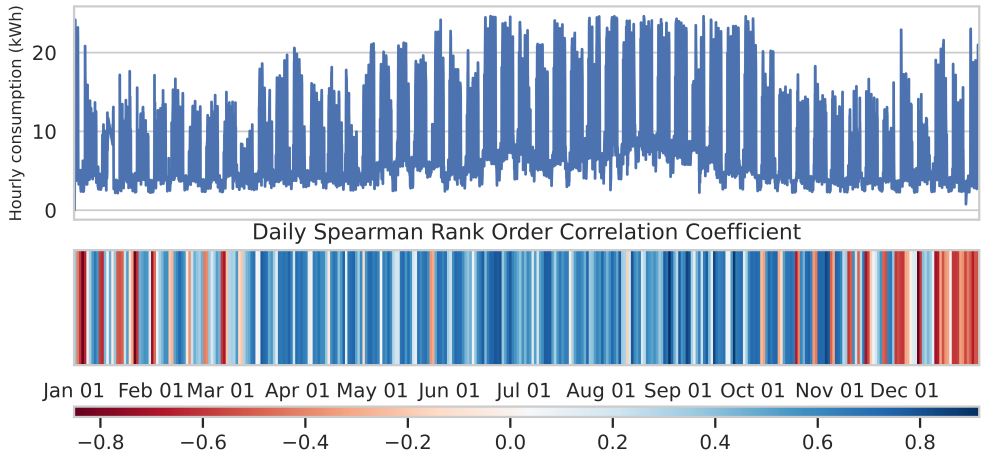
These features include categories such as diurnal patterns, long-term consistency and pattern specificity.

Employing the "Day Filter" function, which is based on Symbolic Aggregate approXimation (SAX) representation of time-series data [28], the mentioned patterns are extracted on a 24 hour period, quantifying the size and the number of motifs obtained for a particular building [13]. The volatility of a building's consumption over a long period of time (such as a year) is instead captured by the long-term pattern consistency that permits monitoring as an instance changes in the schedules between different seasons or due to particular events. These evident changes are identified as "breakouts" and are also added as features to the set, counting the number of occurrences within a chosen interval [13]. An example of the latter feature is represented in Fig. 4-b, where the cumulative number of breakouts in the chosen year is computed.

The last category of pattern-based features concerns the pattern specificity, which indicates whether a



(a) Daily ratio of mean consumption and maximum consumption



(b) Yearly Spearman rank order correlation coefficient for a specific building, blue indicates stronger cooling correlation while red demonstrates higher heating correlation

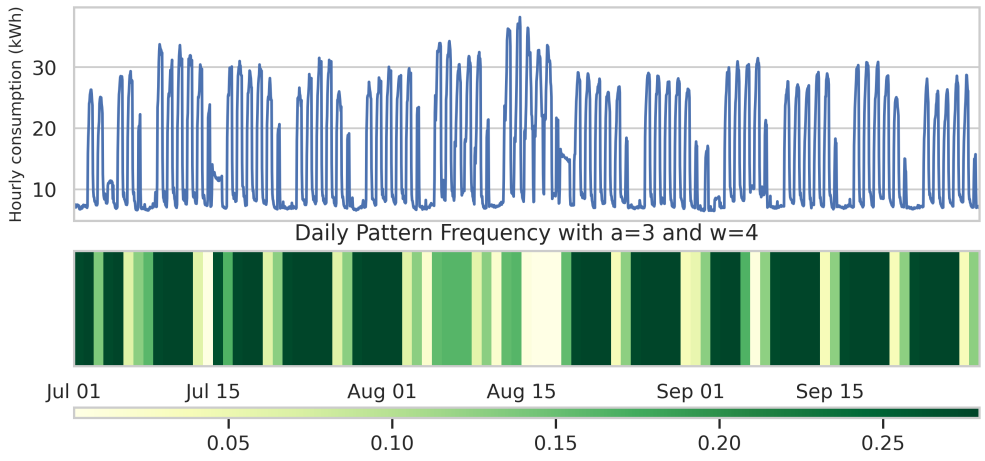
Figure 3: Visual representation of two statistics-based features

building's patterns are typical of a certain use type class; and therefore if it operates analogously to other buildings of the same group [13]. The SAX-VSM process [29] is employed for the extraction of these features.

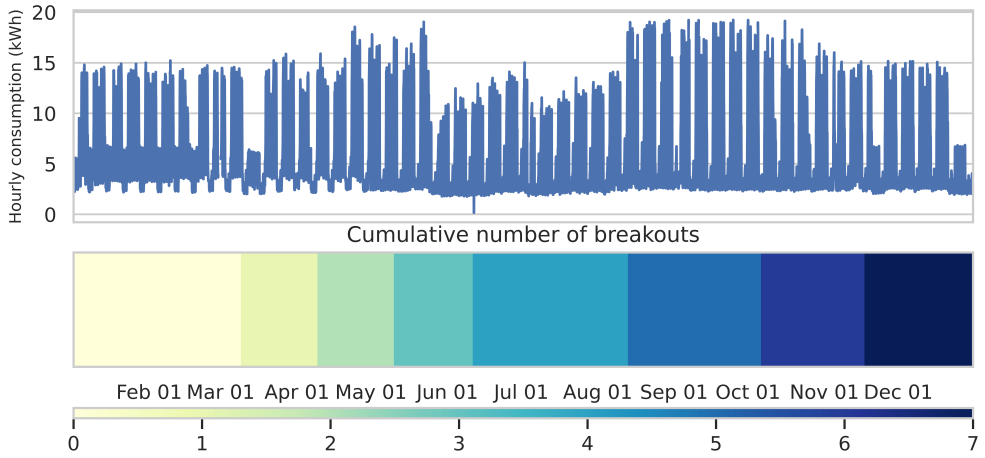
3.1.3. Regression-based features

The output parameters of performance prediction models can provide information about the physical behaviour of buildings. Several electrical consumption prediction models and libraries were employed to obtain these features, among which the Time-of-week and Temperature (TOWT) model [30] implemented in the eetd-loadshape library [31], the Seasonal Decomposition of Time Series using the STL package [32] in R, and the PRISM method [33].

The TOWT model's outcomes attempt to capture the intensity of load dependence to either scheduling or outdoor air temperature. Building load is calculated separately for occupied and unoccupied hours,



(a) Daily consumption pattern frequency



(b) Cumulative number of breakouts

Figure 4: Visual representation of two pattern-based features

which are empirically distinguished according to the different temperature dependence of the consumption during the day. For both cases, the predicted load depends on the selected one-hour time period of the week, its relative outdoor temperature interval among six equally sized intervals, and two coefficients α and β indicating respectively the base load for the selected period and the temperature dependency for that temperature interval and time. Once the prediction is performed, it is possible to obtain a series of metrics from the analysis of the fitted model, such as hourly residuals indicating the actual consumption deviation from the model, and the periods of under-prediction, which indicate whether the building is operated according to its set schedule or not [13, 30].

Another category of features is designed to capture the seasonality and trend behaviours of buildings. Seasonality typically refers to the different consumption patterns occurring between weekdays and weekends, or nights and days. Trends, instead, identify a long-term increase or decrease which usually does not follow

a pattern but are due to the external influences such as weather-related factors, change in occupancy, and loss of system efficiency. These behaviours are extracted using the seasonal trend decomposition procedure based on Loess with the STL R package. Input data is obtained from weather normalized daily consumption, then the remainder quantities R are calculated subtracting the computed trend components T and seasonal components S from the initial input data I, as shown in Eq. 3:

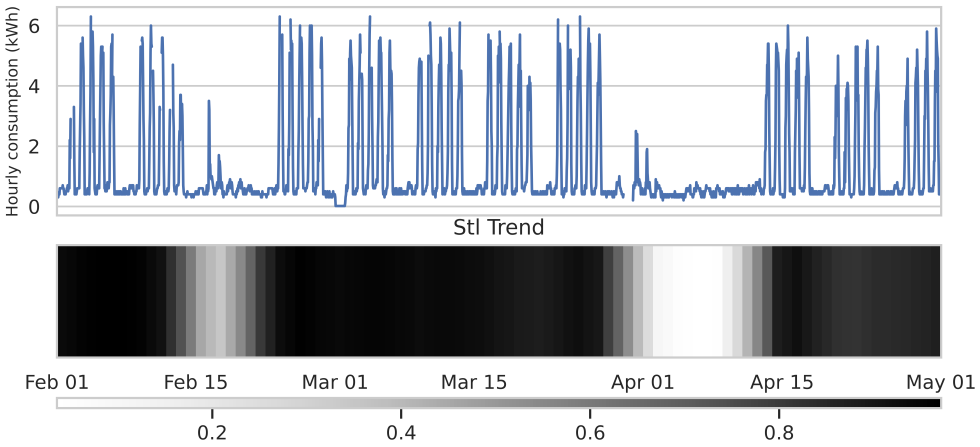
$$R = I - T - S \tag{3}$$

The remainder values indicate which decomposed loads are not well described by the STL model's prediction, as illustrated in Fig. 5-b for a specific building of the dataset, whereas Fig. 5-a provides a visual representation of the daily trend component for the same building. The resulting features consist in the weekly seasonal patterns of each building (daily trend mean), the long-term trend (monthly trend mean), and the remainders from the resulting model using the STL procedure (monthly remainder mean) [13, 32].

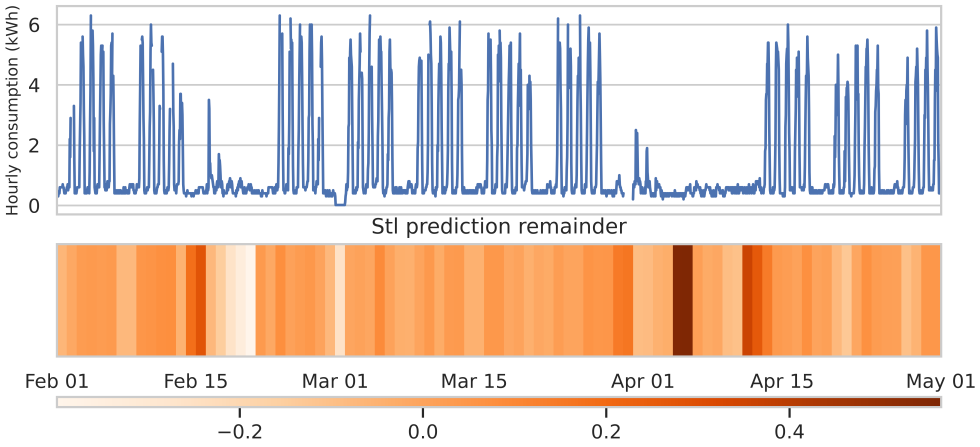
The outputs of linear change point models, based on the PRISM method [33], complete the set of regression-based features. Such models can approximate the amount of load used for each part of the HVAC system through the use of a linear regression model that identifies the point (temperature), after which the relation between consumption and outdoor temperature becomes linear. This point is called cooling balance point and the slope of the line is the cooling energy increase rate. The same procedure is conducted for the heating case, where consumption and temperature are inversely proportional [13].

3.2. Feature Selection

The above-mentioned feature extraction procedure leads to the generation of a large number of features, including which results in an elevated computational cost. Furthermore, employing features that do not contribute to the achieved accuracy (do not provide any benefit to the estimation procedure) can result in over-fitting (to the training set), as the model attempts to include uninfluential parameters as inputs, which in turn reduces the accuracy obtained on the test set. Therefore, implementing a feature selection procedure, which facilitates choosing only the influential parameters [34], permits reducing the calculation cost and models' complexity while increasing (even if marginally) the achieved accuracy (through evading the mentioned over-fitting issue). Different feature selection methods, including the state-of-the-art methodologies and a customized method, are thus employed. The results are determined using a 3-fold cross validation (CV) procedure and the achieved accuracy and number of features, obtained using different feature selection methodologies for different classification targets, are then compared. Detailed descriptions about the employed feature selection procedures are provided in section 4.5.



(a) Daily trend component of consumption obtained through the STL decomposition



(b) Remainder values of the the STL model's prediction

Figure 5: Visual representation of two regression-based features

3.3. Importance-in-prediction based features

In the next step, importance-in-prediction based features, which are the coefficients generated based on the features' influence and relevance in an hour-ahead load prediction pipeline, are extracted. These features, can help the model to correctly classify buildings based on the importance of different features in the corresponding load prediction process. Accordingly, a load prediction pipeline is first developed, in which the target to be predicted is the consumption of the building in the next hour. Next, starting from the raw consumption and weather data, several parameters including statistics-based, seasonality related (calendar based), and lagged features (which are explained in details in sub-section 3.3.1), are extracted and provided to the pipeline as input features. In the next step, different coefficients are generated, using both correlation and importance-related indexes. The employed correlation coefficients are Mutual Information for Regression [26, 35] (with $n\_neighbors = 3$), Pearson Correlation [36, 37], Spearman's Rank Correlation [38, 39] (described

in details in section 4.3) that represent the correlation and mutual dependence between each of the extracted features and the target (consumption in the next hour).

In order to evaluate the contribution of each feature to the prediction process, the pipeline is first implemented, trained and validated using different algorithms. For this purpose, a subset of 15 buildings, randomly selected with equal numbers from different use type categories, are employed, while Random Forest Regressor [26, 40], Multilayer Perceptron (a back propagation-based neural network) [26, 41] and Support Vector Regression [26, 42] are utilized and tuned [43] as prediction algorithms. In the implemented pipelines, the algorithms were trained on 80% of data of each building and tested on the remaining 20%. The algorithm with the highest accuracy (measured using coefficient of determination R^2) for the chosen buildings was determined to be the Random Forest Regressor (with n_estimators=50 and max_depth=10). Accordingly, the feature importance coefficients, generated by Random Forest Regressor through the $feature\_importances\_$ attribute, constitute the last type of generated coefficients that are utilized to represent the influence of each feature in the prediction process.

All of the generated coefficients are then sorted based on the corresponding mutual information values and are progressively added, as additional features, to the previously obtained selected features of each target. The impact (in terms of achieved score improvement and the number of added features) for each index type is then compared in order to select the most suitable one for each classification target. For each case, only the coefficients adding which results in an increment in the achieved average cross validation scores is kept.

3.3.1. Feature Extraction for the load prediction pipeline

As was previously pointed out, different types of features including, statistics-based, seasonality related, and lagged features are extracted from the from the raw consumption and weather data in order to be provided as input features in the load prediction pipeline. For the extraction of statistics-based variables, total load consumption and temperature temporal statistics have been obtained on a daily basis and have been included in the dataset. The latter include basic metrics such as mean, maximum, minimum, variance and different quantiles in the previous 24 hours (with respect to the target's timestamp) [30]. In addition, the ratios and differences of some of the mentioned metrics have also been extracted and added. Lastly, Spearman correlation coefficients between temperature and consumption are generated for intervals of the previous 4, 6, 12 and 24 hours.

Seasonality related features instead include: hour of the day (along with $cos(hour)$ and $sin(hour)$), month, day of the week, week of the year, the weekend flag and the night flag. Lagged features, which are the values of parameters in the previous timestamps, are the last category of provided inputs. These input features are particularly important to take into account the effect of features such as temperature and other ambient conditions that do not have an immediate influence on the load variations (e.g. owing

to the influence of heating or cooling systems' consumption). In the present study, lagged values of total consumption and temperature up to 26 hours (to capture a possible day-ahead consumption correlation), and other features like sin(hour), cos(hour), sea level pressure, humidity, visibility, and wind speed up to 12 hours are extracted and added to the dataset.

3.4. Interpretation of the selected final set of features

Once the final set of selected features is obtained, the relative importance of features along with their contribution to the overall classification performance is demonstrated through graphical illustration. Furthermore, it is attempted to provide interpretations about the reason behind the importance of the most influential features. In this context, distribution plots of different features in various classes have been employed to understand the feature's classification effectiveness and spot differences among classes.

4. Machine learning based pipeline implementation and improvement concepts: utilized classifier, accuracy metrics, correlation indexes and feature selection methods

This present section is focused on providing further theoretical explanations about the machine learning algorithm, accuracy metrics, and correlation indexes along with the feature selection methods which have been utilized in this study.

4.1. Random forests classifier

Random forests, or random decision forests, which is utilized in the present study as the classification algorithm, is an ensemble learning method that is based on building several decision trees in the training process while minimizing a given error metric and providing the average of their predictions as output [40, 44]. The resulting model predicts the value of a target variable by learning simple decision rules from the data features. Considering $T_i(x)$ to be a single regression tree built based on a subset of input features and the bootstrapped samples [40], the tree can be expressed as:

$$\hat{f}_{RF}^{C}(\mathbf{x}) = \frac{1}{C}\sum_{i=1}^{T} T_i(\mathbf{x}) \tag{4}$$

in which C represents number of trees and x is the vectored input variable [40].

A random forest classifier is employed for the classification task, the performance of which is assessed both using average 3-fold cross validation scores and directly employing a 50% training set to generate confusion matrices and other visual indicators. Cross validation scores are mostly utilized during the feature selection process to ensure that the whole dataset is considered when evaluating the impact of features addition or removal.

4.2. Accuracy metrics

In the present sub-section, the key metrics that have been employed to evaluate the model's performance are presented. It is worth noting that, as the pipelines implemented in the present work are defined as multi-class classification problems, per-class scores have to be combined to obtain a single averaged value. This value is called micro-averaged if the average is determined on the total number of elements, macro-averaged if it is computed for each class and then divided by the number of classes, and weighted-averaged if the number of elements belonging to each class is considered in the averaging procedure [26, 45, 46].

4.2.1. Accuracy

While considering false positive (FP) (number of elements wrongly labeled as positive) and false negative (FN) (number of elements wrongly labeled as negative) as the two types of possible errors in a classification process, accuracy can be defined as the ratio between correct predictions (true positives (TP) and true negatives (TN)) and the total number of cases [26, 46]:

$$Accuracy = \frac{\sum TP + TN}{\sum (TP + TN + FP + FN)} \tag{5}$$

4.2.2. Precision,recall, and F1-score

Precision, also called positive predicted value, indicates the fraction of correctly selected elements among the relevant ones in a classification context [46, 47].

$$Precision = \frac{\sum TP}{\sum (TP + FP)} \tag{6}$$

Recall, or true positive rate (sensitivity), is the ratio of the true positives to the total elements classified as positive, defined as [26, 46]:

$$Recall = \frac{\sum TP}{\sum (TP + FN)} \tag{7}$$

The F1-score is a measure that combines precision and recall in the following way [46]:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{8}$$

4.2.3. Coefficient of determination

The coefficient of determination (R^2 score) is the proportion of the variance in the dependent variable that can be estimated by the independent variable(s). It measures the extent that the model replicates the observed outcomes, based on the proportion of total variation of outcomes explained by the model [44, 48, 49].

Considering SS_{res} as the sum of squares of residuals, and SS_{tot} to be the total sum of squares, then R^2

is expressed as [44, 48, 49]:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{9}$$

357

358 *4.3. Coefficients utilized to represent importance-in-prediction*

359 In the preset sub-section, the theoretical description about the indexes that have been employed for gener-
360 ation of importance-in-prediction based features are provided.

361 *4.3.1. Mutual information*

Mutual Information (MI) quantifies the mutual dependence between two random variables that are sampled simultaneously. It measures the amount of information acquired about a random variable through observing the other variable. The mutual information, while considering two random variables X and Y, is determined employing the following equation [50, 51]:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{10}$$

where $p(x, y)$ represents the joint probability mass function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability mass functions of X and Y [50, 51] that can be expressed as:

$$P_X(x) = \sum_{y \in Y} P_{XY}(x,y) \tag{11}$$

362 Mutual information is employed to determine the importance of each feature with respect to the estimation
363 target, which can be either discrete or continuous.

364 *4.3.2. Pearson correlation*

365 The Pearson correlation coefficient is a measure of the linear correlation between two variables and is
366 expressed as the ratio between the corresponding covariance and the product of their standard deviations
367 [36, 37], as shown in equation (12).

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{12}$$

368 The range of this index is between -1 and +1 [36, 37].

369 *4.4. Spearman correlation*

Spearman's rank correlation coefficient is a measure of the rank correlation, which is the statistical dependence between the two variables' rankings. While Pearson's correlation evaluates linear relationships,

Spearman's correlation assesses the extent that two variables' relationship can be described using a monotonic function (whether linear or not). In other words, the Spearman correlation coefficient can be defined as the Pearson correlation coefficient [36, 39] between the rank variables. Ideal Spearman correlation of +1 or -1 occurs when each of the variables is an idea monotone function of the other one [38, 39]. Considering n ranks as distinct integers, it can be computed using Eq. (13) [38]:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{13}$$

4.4.1. Random forest's feature importance

Once a random forest model is fit using the training data, it is possible to access the corresponding feature importance coefficients (employing the $feature\_importances\_$ attribute) [40], each of which is an index of how well a certain variable can help predicting the target. As expressed in Eq. (14), the importance of input feature X_i for predicting Y is found by summing the importances of the j-th nodes ni_j on which X_i is split, divided by all nodes' importances, and finally averaged over all T trees in the forest [52].

$$Imp(X_i) = \frac{1}{T} \sum_{t \in allTrees} \frac{\sum_{j \in nodeSplitOnX_i} ni_j}{\sum_{k \in allNodes} ni_k} \tag{14}$$

4.5. Feature Selection

As was previously pointed out, selecting features is an effective method for reducing the computational cost and model's complexity, while maintaining an acceptable accuracy and even marginally improving it [53] by evading the over-fitting issue. Accordingly, different state-of-the-art feature selection methods (which are implemented in SciKit-Learn [26, 54] library) along with a customized approach are employed in the present work, a brief description of which is provided in the present sub-section.

4.5.1. Univariate selection

This method uses univariate statistical tests to select the features of a data set which have the strongest relationship with the output variable [26, 54] , it includes different approaches:

- *SelectKBest*, which only keeps the k highest scoring features;
- *SelectPercentile* only holds a user-specified percentage of the highest scoring features;
- Common univariate statistical tests like *SelectFpr*, *SelectFdr* and *SelectFwe* respectively for false positive rate, false discovery rate and family wise error [26, 54];
- *GenericUnivariateSelect* allows to select the best univariate selection strategy among the previous methods and possible scoring functions with hyper-parameter search [26, 54] .

These objects take as inputs a scoring function which vary according to the nature of the machine learning problem and include: ANOVA F-value between label/feature for classification tasks ($f\_classif$), Chi-squared stats of non-negative features for classification tasks ($chi2$), F-value between label/feature for regression tasks ($f\_regression$), mutual information between features and the target ($mutual\_info\_regression$) [26, 54].

4.5.2. Select from model

Select from model method can be used with any estimator which has a $coef\_$ or $feature\_importances\_$ attribute and it consequently removes insignificant features according to the given threshold parameter. Accordingly, an option is using tree-based feature selection within estimators such as random forest regressor [40] or extra trees regressor [55] that are able to compute features' importance. The parameters of such estimators include $n\_estimators$ (number of trees in the forest), $criterion$ (the function according to which the quality of a split is measured) and $max\_features$ (number of the best features to keep).

4.5.3. Recursive feature elimination

Recursive Feature Elimination (RFE) recursively considers smaller sets of features and eventually selects the best set with the best scores that are given by an external estimator which assigns weights to features [56, 57]. The estimator is first trained on the initial set obtaining the importance of each feature through a $coef\_$ or $feature\_importances\_$ attribute, secondly the least important features are dismissed from the current set. These two steps are recursively repeated until the desired number of features is reached. RFECV, is an extended version of RFE that includes a cross-validation loop to find the optimal number of features [56].

4.5.4. Customized Feature Selection

A customized feature selection method, in which the mutual information and above-mentioned accuracy metrics are employed, was also proposed and implemented. In this approach, the features are first sorted based on their mutual information coefficient. Next, starting with the most correlated feature, the loop adds a new element to the set only if it leads to an improvement in the averaged cross validation scores (either weighted F1-score or accuracy). The use of cross validation allows to consider whether one feature is significant on average and not only for the portion of data on which the model is tested.

It is noteworthy that, while implementing this approach, besides mutual information, RF feature importance, Pearson correlation and permutation importance were also tested as sorting criterion, though mutual information turned out to give better overall results. Different steps of this method are represented in Fig. 6.

5. Results and discussions

In the present section, results of the feature selection procedures, in terms of accuracy and the number of selected features, are first presented and discussed. The improvement in the classification performance

Figure 6: Schematic representation of the implemented customized feature selection methodology.

through adding the importance-in-prediction based features (extracted from prediction pipelines) are then determined and demonstrated. Lastly, interpretations about the selected features and the obtained results are provided.

5.1. Feature selection results

As was previously pointed out, different feature selection procedures were conducted for all three classification targets, starting from the corresponding initial feature set. Fig. 7 represents the resulting weighted F1 scores, accuracy, and the number of selected features. It is worth mentioning that the classification outcomes (classification performance and number of selected features) were found to be fairly sensitive to each method's tuning parameters and the most promising ones, obtained after performing a thorough test of different parameters, are provided in this figure.

It can be observed that all of the implemented feature selection methods can substantially reduce the number of utilized features, while, in most of the cases, leading to even an improvement in the obtained classification performance. Furthermore, it can be noticed that the proposed customized method results in the highest performance and lowest number of features for use type and performance class targets. Therefore, for these two targets, the feature sets obtained by the customized method are chosen as the final ones. Accordingly, the 13 selected features for the performance class target result in an accuracy score of 0.609 (and a weighted F1 score of 0.608), while the 23 chosen features for the use type target lead to an accuracy score of 0.724 (and a weighted F1 score of 0.711).

For the case of operations group target, the features selected using the RFECV and "Select from Model" methods provide a higher classification performance compared to the one obtained using the proposed customized method (weighted F1 scores of 0.997 and 0.994 respectively compared to 0.985 obtained by the customized method). Although the features selected by RFECV method provide a slightly higher classification performance than the ones obtained using "Select from Model", 16 features are chosen using the former method, while only 8 features are selected using the latter one. Since, owing to the reasons provided in section 4.5, having a small number of features is preferred, the feature set obtained employing "Select From Model", which leads to the second highest score and the fewest features, is chosen as the final set for the operation group target. This latter set is then further reduced using a similar loop to that of the customized feature selection, which allows obtaining an even smaller set (only four features) with the same accuracy.

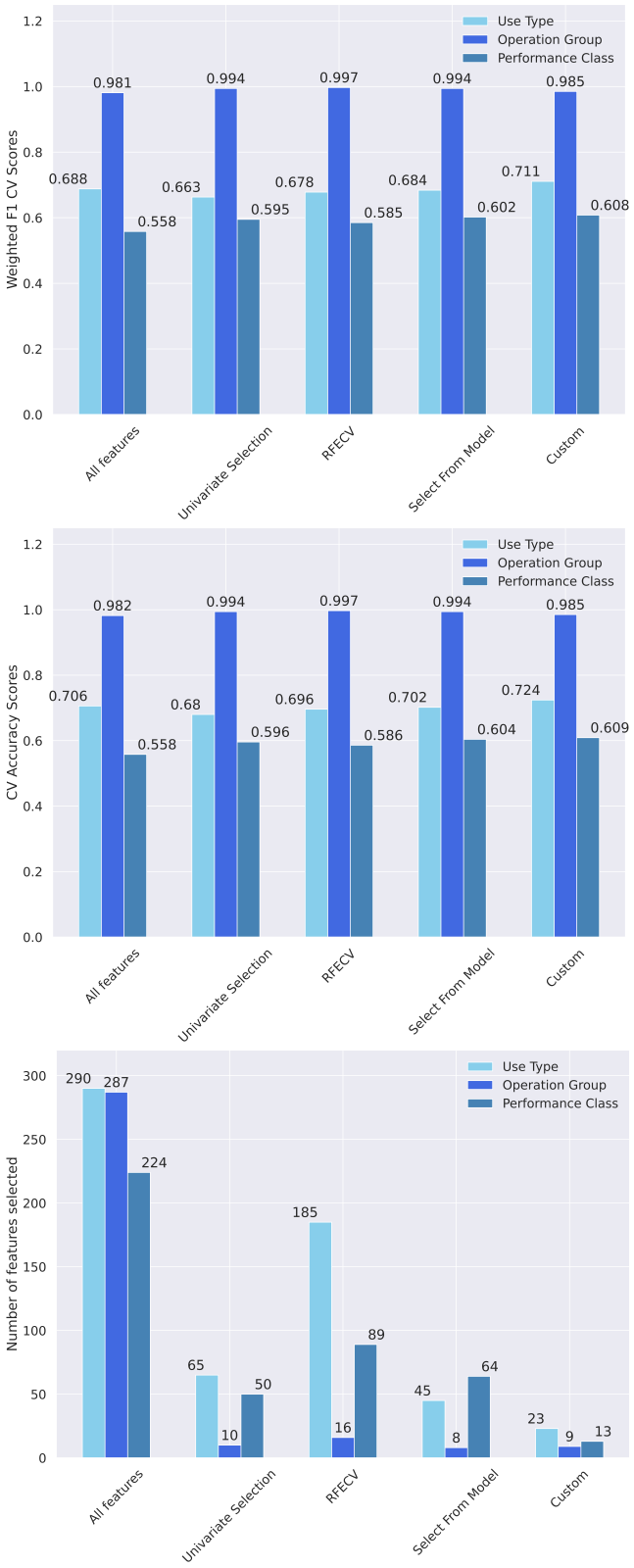


Figure 7: Comparison of the results of performing different feature selection methods for the considered classification targets

5.2. Addition of Importance-in-prediction based Features

Starting from the features sets and the resulting scores obtained from the feature selection step, the importance-in-prediction based features (extracted from an hour ahead prediction pipeline) are progressively added and are only kept in case the addition results in an improvement in the obtained score. For all of the considered targets, the four different indexes that measure the influence of features in the hour ahead pipelines (which were described in section 4.3) are extracted and the resulting impact of adding them to the set of features is assessed. It was demonstrated that for the use type and operation group targets, the scaled mutual information based features lead to the highest improvement in the achieved scores. For the case of performance class target, the Spearman's Rank Correlation coefficients were determined to result in a slightly higher improvement. Pearson correlation and random forest importance coefficients were also demonstrated to improve the scores but were not as effective as the latter indexes.

Accordingly, six new features are added to the selected set for the use type target, while four features are added to that of the performance class and two features are added to the operation group target's selected set. Table 2 reports the improvements obtained in the classifications score for different targets by adding the latter features. It can specifically be noticed that, through adding the mentioned additional features, the accuracy and F1 score achieved for the performance group target is increased to 1.

	Building Use Type			**Performance Class**			**Operation Group**		
Pipeline	A	B	C	A	B	C	A	B	C
Mean CV Accuracy	0.706	0.724	0.742	0.558	0.609	0.621	0.982	0.994	1.000
Mean CV F1-score	0.688	0.711	0.731	0.558	0.608	0.621	0.981	0.994	1.000
Number of Features	290	23	29	224	13	17	287	4	6

Table 2: Performance of model employing pipeline (A): initial set of features , pipeline (B): selected features, and pipeline (C): selected features plus the addition of importance-in-prediction based features.

5.3. Selected features and interpretation of the results

In the present sub-section, the selected features and the added importance-in-prediction based features, for each of the considered targets, along with the corresponding effect on the obtained classification score are presented. Next, an interpretation about the reason behind the impact of some of the most influential parameters is provided. Having a deeper understanding of how, why, and which variables affect the building classification performance is a meaningful step towards improving the models, that can favour a more focused approach in this area. In this context, the selected features and the corresponding interpretation is provided for each target individually and the results of the analysis is compared with those presented in the foregoing works [58] and [21].

5.3.1. *Use type*

Fig. 8 represents all of the selected features and the added importance-in-prediction based features (denoted by the initial term "imp") for the building use type target along with the contribution of adding each feature to improving the achieved classification scores. As can be observed in this figure, the minimum area normalized consumption (*"AreaNormalizedConsMin"*), coherently with what was reported in [58], is still among the first selected features and results in a notable score improvement as it is an intuitive index of the energy intensity of each space, which is higher for labs and more similar for the other use types. Similarly, the average of this normalized variable (*"AreaNormalizedConsMean"*) is among the selected features.

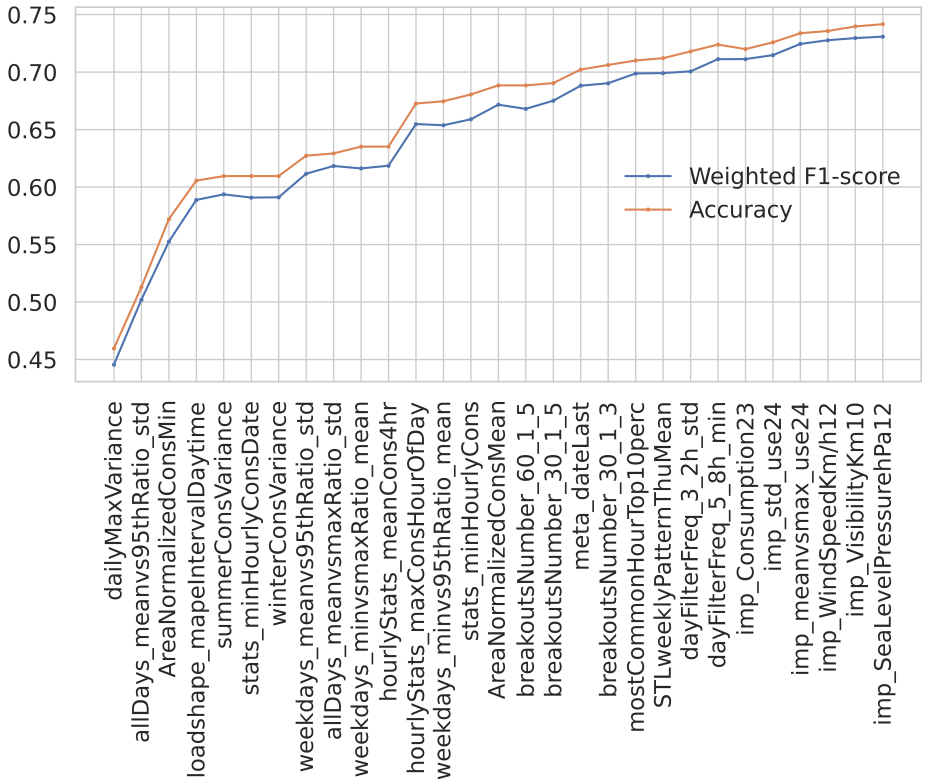


Figure 8: Improvements in the achieved scores after each feature is added to the selected set for building use type target

Considerable improvement in accuracy is also observed to be caused by the addition of maximum daily variance in consumption (*"dailyMaxVariance"*) and the standard deviation of average consumption with respect to the corresponding 95th percentile value both for all days (*"allDays_meanvs95thRatio_std"*) and weekdays (*"weekdays_meanvs95thRatio_std"*). These features are promising describers of the higher variance of both daily maximum consumption and of the daily ratio mean/95th percentile consumption of Primary/Secondary Classrooms compared to the other classes. The latter difference can be clearly observed in the box plot provided in Fig. 9, which represents the distributions of significant features for different classes for the Building Use Type target. The latter difference can be attributed to the fact that primary and

secondary classrooms (PrimClass) are among the buildings with the lowest area-normalized consumption and have a less regular use of electric appliances, the impact of which is easily detected owing to their lower base-load.

Other features which notably enhance the achieved accuracy are the hour at which the maximum consumption takes place (*"hourlyStats_maxConsHourOfDay"*) and the most common hour at which the top 10% of consumption takes place *"mostCommonHourTop10perc"*, which are indicators of the hour at which the maximum consumption occurs. These features help distinguishing between dormitories and other use types as their maximum consumption is typically taking place later at night when most people are back from office or university. The latter difference can be easily noticed in Fig. 9.

Another variable that is useful to spot primary and secondary classes is variance in the winter consumption *"winterConsVariance"*, a seasonal consumption statistics, which once again underlines the significantly higher consumption variance of such class. In addition, number of breakouts are also confirmed to be suitable indexes to detect primary and secondary classrooms, that are more likely to have frequent changes of schedule for holiday breaks and similar events, while other categories like Offices often follow only few days of national holidays and have more regular schedules. Finally, the selected STL normalized weekly pattern, is useful to spot human-behavior influenced patterns of dormitories.

Regarding the importance-in-prediction based features, as was pointed out in section 5.2, for the case of the use type target, the scaled mutual information based coefficients were determined to be the ones adding which results in the highest improvement in the achieved accuracy. Therefore, the importance-in-prediction based features (denoted by the initial term "imp" in Fig. 8) for this target, refer to the scaled mutual information of each indicated feature with the target (consumption in the next hour). Among these features, the ones corresponding to 23-hour-lagged consumption (*"imp_Consumption23"*) and deviation of daily consumption (*"imp_std_use24"*) seem to be promising indicators of different use types. As an instance, for the case of dormitories, the 23 hours lagged consumption (thus, the consumption of the building 24 hours before the time-stamp to be predicted) has a higher correlation with the predicted consumption. This observation can be attributed to the fact that the consumption of dormitories in a specific day is pretty similar to the previous one, thus implying that day-ahead load can be a good predictor for the next-hour forecast. Standard deviation of the previous 24 hours of consumption also has an elevated correlation with the target for dormitories and primary/secondary classes (as can be observed in Fig. 9) that could be the consequence of the fact that specific consumption schedules are not followed in these spaces, which in turn results in a stronger importance of the consumption deviation as a useful parameter for load prediction. The latter interpretation can also be extended to the next added feature, importance-in-prediction of the ratio between mean and maximum consumption values of the last 24 hours, which is higher for primary/secondary classes and that can be attributed to their higher consumption variance. The remaining importance-in-prediction features are mainly related to weather parameters, which could be identifying different weather-related

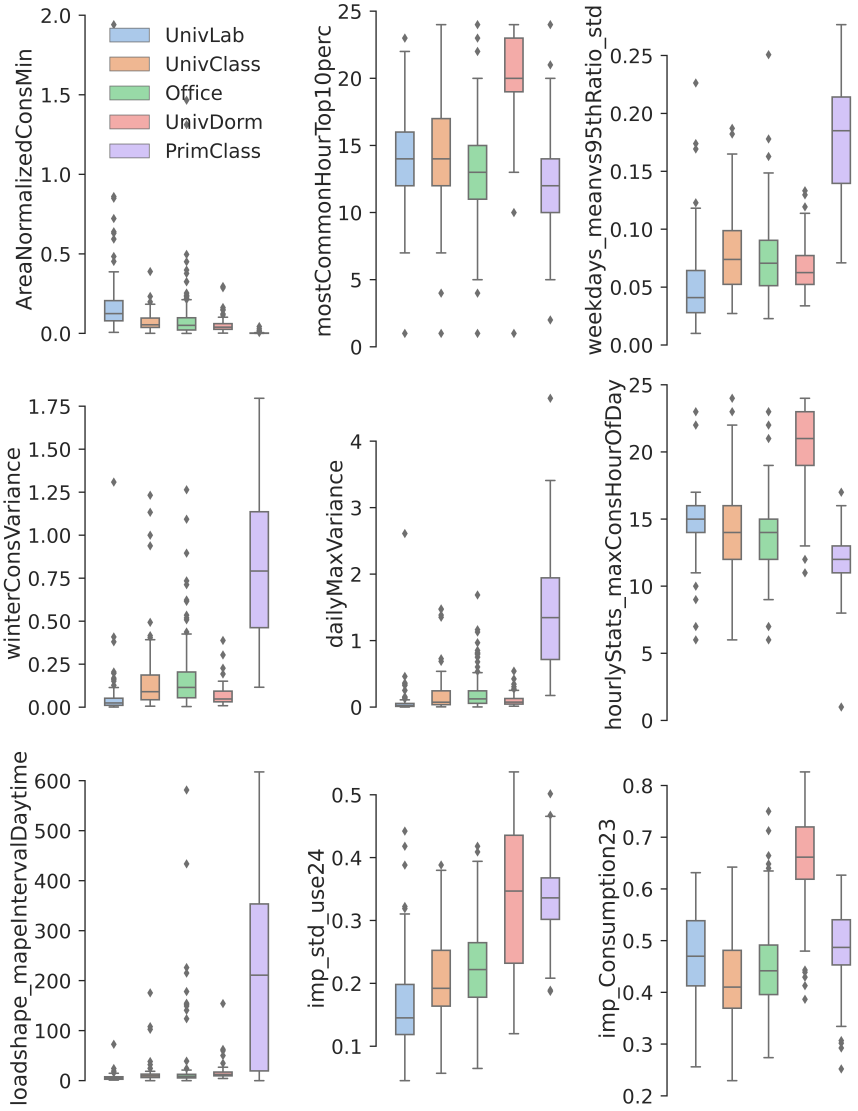


Figure 9: Boxplots of distributions of significant features for building use type target

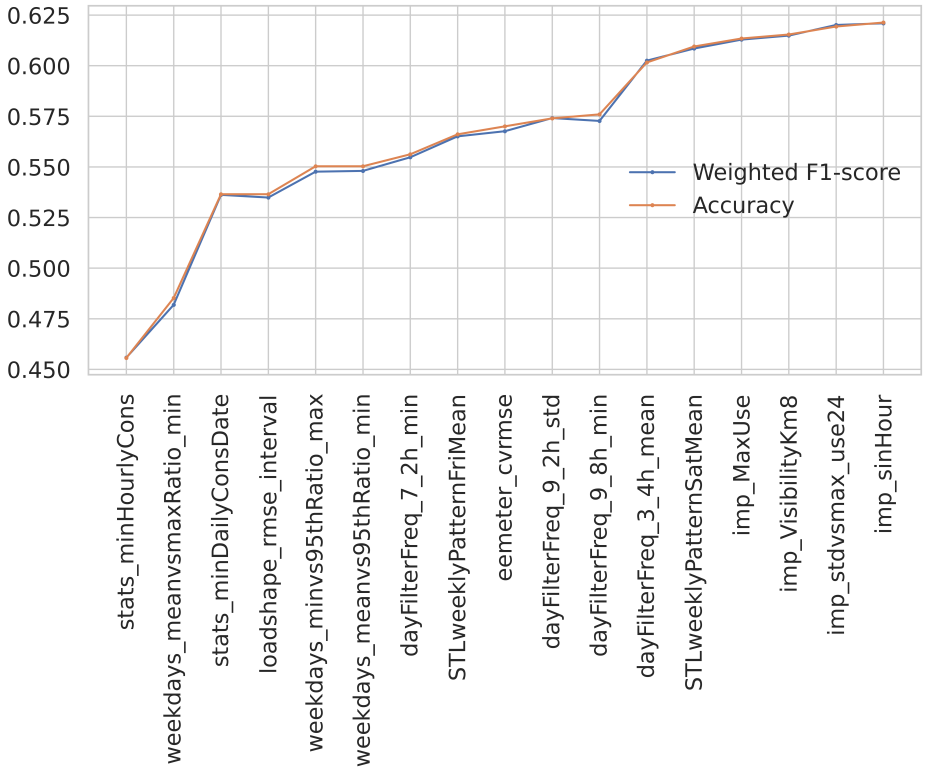


Figure 10: Improvements in the achieved scores after each feature is added to the selected set for performance class target

behaviours for different building types, even though values distribution for these features are very similar among classes.

Differently from what was reported in the previous investigations ([21, 58]), consumption statistics and other statistical features, in particular indicating variance, are the most common category in the characterization of building use type. It is worth noting that, based on the classification outcomes, the offices are often mistaken for university classrooms or laboratories. This might be due to the resemblance of these spaces which can be used with similar purposes, or in part to outdated/inaccurate labeling.

It is also worth mentioning that, although Fig. 8 illustrates the improvements obtained in the achieved accuracy owing to the addition of each feature, the observed improvement does not necessarily correspond to the exact individual importance of each specific feature, as it can also be attributed to the joint influence of the added features and the existing (previously added) ones. A list of the exact accuracy values obtained after each selected feature is given in Tab. 3.

5.3.2. Performance class

Fig. 10 represents the selected feature and the added importance-in-prediction features for the performance class target along with the corresponding resulting influence on the achieved accuracy scores. It can be no-

ticed that the most notable improvements in accuracy is observed while adding features such as the date on which the minimum daily consumption takes place (*"stats_minDailyConsDate"*), *"dayFilterFreq_3_4h_mean"* and *"weekdays_meanvsmaxRatio_min"*. The date on which the minimum daily consumption occurs appears to be more variable for low consuming buildings, that can be attributed to the corresponding lower dependence on the weather condition and, therefore, on the time of the year. The latter difference in the distribution of this variable can also be observed in Fig. 11.

Coherently with the observation reported in [58], the two most relevant groups of features, for the classification of this target, include the load diversity (represented by load ratios) and consumption patterns. The first group, which includes the above mentioned minimum mean/max ratio (*"weekdays_meanvsmaxRatio_min"* and *"weekdays_meanvs95thRatio_min"*), indicates the magnitude of the mean consumption compared to the peak and tends to be lower for low consuming buildings, implying that a building is more likely to consume less overall if there are only limited moments of high peak consumption and a low mean load. The pattern-based features manage to indicate some pattern differences among classes, which confirms the link between variety of patterns (possibly due to the implemented energy-saving or demand-response policies) and lower consumption. *"loadshape_rmse_interval"* is particularly effective to distinguish low consumers, for which the regression fitting error is sensibly lower than that of the other classes. This implies that in the buildings with a regular and predictable schedule, reasoned control strategies might have been implemented and are thus the ones that are supposed to consume less.

Lastly, four importance-in-prediction based features (denoted by the initial term "imp") were also added. These, as pointed out in section 5.2 for the performance class target, are the Spearman's Rank Correlation coefficients between the indicated parameter and the consumption in the next hour. One of these features is the one that represents the correlation between the visibility and the consumption to be predicted (*"imp_VisibilityKm8"*), which can be attributed to the differences in the weather dependence among different performance classes. As can be observed in Fig. 11, for the low consumers, visibility has a slightly lower correlation (with the predicted target) compared to the other classes, that can be linked to the fact that low consumption is also related to higher energy efficiency, better envelope insulation, and consequently lower dependence on weather related parameters.

Other features are related to the importance of different maximum load statistics-related variables (*"imp_MaxUse"*), which are on average slightly more important for the load prediction of high consuming buildings. The last feature (*"imp_sinHour"*) concerns the importance of the "sin(hour)" variable, that is slightly higher for low consuming buildings, implying that, as was previously pointed out, a higher regularity and respect of schedules during the day can be linked to a lower overall consumption.

Lastly, based on the classification results, it can also be observed that the accuracy is promising for high and low consuming buildings, whereas the the intermediate class are often mis-classified. This phenomenon can clearly be attributed to the similar distribution of values of the features for this target.

5.3.3. Operation group

For the last target, the selected features are only six, while the resulting scores are nonetheless notably high as an accuracy and F1 score of 1 can be achieved. The contribution of adding each feature to the obtained accuracy is shown in Fig. 12. As can be noticed in this figure, the most influential feature to differentiate between groups is a pattern-based one (*"dayFilterFreq_9_6h_min"*). Other important features are statistics-based, including *"allDays_minvs95thRatio_max"*, *"stats_minDailyConsDate"* and *"maxDailyConsDate"*. As can also be noticed in Fig. 13, the first mentioned feature describes Group 2 class, while the other two variables help distinguishing Group 4 from the other groups. Overall, it can be concluded that the features that better underline different operation strategies and schedules are mainly consumption statistics/ratios and pattern-based ones.

To conclude, two importance-in-prediction based features are added; similarly to the case of use type target, these are referring to the scaled mutual information of each indicated parameter with the consumption in the next hour. These features include the importance of a load-ratio feature (*"imp_quant50vsmax_use24"*) and the importance of a 5-hour lagged consumption (*"imp_Consumption5"*), both of which help separating Group 2 from the other groups. A summary of the scores improvement during the feature selection process for all classification targets can be found in Tab. 3.

6. Conclusion

In the present work, the most influential temporal and importance-in-prediction based features, which can be extracted from smart meter data, aiming at remote characterisation of non-residential buildings, were determined. Remote estimation of the defined targets (use type, performance class, and operation group) can be notably helpful in the large-scale building commissioning, benchmarking, and diagnostics processes. In this context, reducing the number of features utilized in the procedure can notably simplify the corresponding implementation, significantly reduce the calculation cost in large-scale deployment, help evading the model over-fitting, enhance the interpretability, and even improve the achieved accuracy.

Accordingly, state-of-the-art feature selection methods and a proposed customized approach were first employed for determining the influential parameters in a pool of temporal features proposed in a previous study [13]. It was demonstrated that employing the latter procedures can notably reduce the number of utilized features; while, even if marginally, improving the obtained accuracy. Furthermore, a set of importance-in-prediction based features, which are coefficients that represent the correlation of various parameters with the consumption in the next hour (to be predicted), were added to the previously obtained selected set of features. It was shown that adding these features can improve the obtained accuracy for all of the considered classification targets.

It was demonstrated that, through performing the latter steps, number of the utilized features for

	Use type	
	Accuracy	*F1-score*
dailyMaxVariance	0.46	0.445
allDays_meanvs95thRatio_std	0.513	0.502
AreaNormalizedConsMin	0.572	0.552
loadshape_mapeIntervalDaytime	0.606	0.589
summerConsVariance	0.609	0.594
stats_minHourlyConsDate	0.609	0.591
winterConsVariance	0.609	0.591
weekdays_meanvs95thRatio_std	0.627	0.612
allDays_meanvsmaxRatio_std	0.629	0.618
weekdays_minvsmaxRatio_mean	0.635	0.616
hourlyStats_meanCons4hr	0.635	0.619
hourlyStats_maxConsHourOfDay	0.673	0.655
weekdays_minvs95thRatio_mean	0.675	0.654
stats_minHourlyCons	0.68	0.659
AreaNormalizedConsMean	0.688	0.672
breakoutsNumber_60_1_5	0.688	0.668
breakoutsNumber_30_1_5	0.69	0.675
meta_dateLast	0.702	0.688
breakoutsNumber_30_1_3	0.706	0.69
mostCommonHourTop10perc	0.71	0.699
STLweeklyPatternThuMean	0.712	0.699
dayFilterFreq_3_2h_std	0.718	0.701
dayFilterFreq_5_8h_min	0.724	0.711
imp_Consumption23	0.72	0.711
imp_std_use24	0.726	0.715
imp_meanvsmax_use24	0.734	0.724
imp_WindSpeedKm/h12	0.736	0.728
imp_VisibilityKm10	0.74	0.73
imp_SeaLevelPressurehPa12	0.742	0.731

	Performance class	
	Accuracy	*F1-score*
stats_minHourlyCons	0.456	0.456
weekdays_meanvsmaxRatio_min	0.485	0.482
stats_minDailyConsDate	0.536	0.536
loadshape_rmse_interval	0.536	0.535
weekdays_minvs95thRatio_max	0.55	0.548
weekdays_meanvs95thRatio_min	0.55	0.548
dayFilterFreq_7_2h_min	0.556	0.555
STLweeklyPatternFriMean	0.566	0.565
eemeter_cvrmse	0.57	0.568
dayFilterFreq_9_2h_std	0.574	0.574
dayFilterFreq_9_8h_min	0.576	0.573
dayFilterFreq_3_4h_mean	0.602	0.602
STLweeklyPatternSatMean	0.609	0.608
imp_MaxUse	0.613	0.613
imp_VisibilityKm8	0.615	0.615
imp_stdvsmax_use24	0.619	0.62
imp_sinHour	0.621	0.621

	Operation group	
	Accuracy	*F1-score*
stats_minDailyConsDate	0.751	0.749
maxDailyConsDate	0.79	0.791
dayFilterFreq_9_6h_min	0.976	0.976
allDays_minvs95thRatio_max	0.994	0.994
imp_quant50vsmax_use24	0.997	0.997
imp_Consumption5	1	1

Table 3: Accuracy scores improvement during the feature selection process for all classification targets

estimating the buildings' use type is reduced from 290 to 29 while augmenting the accuracy from 71% to 74%. The classification accuracy for the performance class was instead improved from 56% to 62% while employing 17 features compared to 224 features available in the initial pool of temporal features. While aiming at estimating the buildings' operation groups, employing only 6 selected features, an accuracy of 100% was achieved. In the last step, multi-class box-plots were utilized to demonstrate the distributions of various features in buildings belonging to different classes, which were then employed to provide interpretations about the capability of some features in distinguishing specific classes.

It is noteworthy that, in order to enhance the generalisability of the implemented method, the authors have provided, the processed dataset, the obtained optimal pipelines (selected feature sets), and the implemented feature selection procedures in an online repository (link provided in Appendix A). The latter scripts permit the researchers to perform the same procedures for other datasets (including more buildings, additional building use types, or classification targets) and obtain similar feature distribution and feature selection plots along with the corresponding optimal feature sets.

Appendix A. Online repository of the implemented procedures

The utilized processed dataset, the obtained optimal sets of features (for each considered target), and the implemented feature selection procedures are provided in an online repository (Link).

References

[1] International Energy Agency, . The critical role of buildings: Perspectives for the clean energy transition. 2019. URL: www.iea.org/reports/the-critical-role-of-buildings.

[2] Energy Information Administration, . International energy outlook 2019. 2019. URL: www.eia.gov/outlooks/ieo/.

[3] Mills, E.. Building commissioning: a golden opportunity for reducing energy costs and greenhouse gas emissions in the united states. Energy Efficiency 2011;4:145—-173. doi:10.1007/s12053-011-9116-8.

[4] Department of Energy, U.S.. Building energy use benchmarking. 2020. URL: www.energy.gov/eere/slsc/building-energy-use-benchmarking.

[5] Department of Energy, U.S.. Retrofit existing buildings. 2020. URL: www.energy.gov/eere/buildings/retrofit-existing-buildings.

[6] International Energy Agency, . Tracking buildings 2020. 2020. URL: www.iea.org/reports/tracking-buildings-2020.

[7] European Commission, . Benchmarking smart metering deployment in the eu-27 with a focus on electricity. 2014. URL: ses.jrc.ec.europa.eu/publications/reports/benchmarking-smart-metering-deployment-eu-27-focus-electricity.

[8] Energy Information Administration, U.S.. How many smart meters are installed in the united states, and who has them? 2019. URL: www.eia.gov/tools/faqs/faq.php?id=108&t=3.

[9] Räsänen, T., Kolehmainen, M.. Feature-based clustering for electricity use time series data. Conference Paper: Lecture Notes in Computer Science 2009;doi:10.1007/978-3-642-04921-7_41.

[10] Dasgupta, S., Srivastava, A., Cordova, J., Arghandeh, R.. Clustering household electrical load profiles using elastic shape analysis. 2019 IEEE Milan PowerTech 2019;doi:10.1109/PTC.2019.8810883.

[11] Najafi, B., Moaveninejad, S., Rinaldi, F.. Data analytics for energy disaggregation: Methods and applications. Big Data Application in Power Systems 2018;:377–408doi:10.1016/B978-0-12-811968-6.00017-6.

[12] Mathieu, J., Price, P., Kiliccote, S., Piette, M.. Quantifying changes in building electricity use, with application to demand response. IEEE Trans Smart Grid 2011;2:507–518. doi:10.1109/TSG.2011.2145010.

[13] Miller, C.. Screening Meter Data: Characterization of Temporal Energy Data from Large Groups of Non-Residential Buildings. 2016. doi:10.3929/ethz-a-010811999.

[14] Miller, C., Meggers, F.. The building data genome project: An open, public data set from non-residential building electrical meters. Energy Procedia 2017;122:439–444. doi:10.1016/j.egypro.2017.07.400.

[15] Zhao, H.X., Magoulès, F.. Feature selection for predicting building energy consumption based on statistical learning method. Journal of Algorithms & Computational Technology 2012;6(1):59–77. doi:10.1260/1748-3018.6.1.59.

[16] Kapetanakis, D.S., Mangina, E., Finn, D.P.. Input variable selection for thermal load predictive models of commercial buildings. Energy and Buildings 2017;137:13 – 26. doi:10.1016/j.enbuild.2016.12.016.

[17] Zhang, L., Wen, J.. A systematic feature selection procedure for short-term data-driven building energy forecasting model development. Energy and Buildings 2019;183:428 – 442. doi:10.1016/j.enbuild.2018.11.010.

[18] Westermann, P., Deb, C., Schlueter, A., Evins, R.. Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data. Applied Energy 2020;264:114715. doi:10.1016/j.apenergy.2020.114715.

[19] Yang, L., Lyu, K., Li, H., Liu, Y.. Building climate zoning in china using supervised classification-based machine learning. Building and Environment 2020;171:106663. doi:doi.org/10.1016/j.buildenv.2020.106663.

[20] Piscitelli, M.S., Brandi, S., Capozzoli, A.. Recognition and classification of typical load profiles in buildings with non-intrusive learning approach. Applied Energy 2019;255:113727. doi:doi.org/10.1016/j.apenergy.2019.113727.

[21] Miller, C.. What's in the box?! towards explainable machine learning applied to non-residential building smart meter classification. Energy and Buildings 2019;doi:10.1016/j.enbuild.2019.07.019.

[22] Miller, C.. Temporal features for non res. buildings library. 2018. URL: github.com/buds-lab/temporal-features-for-nonres-buildings-library.

[23] Senin, P., Lin, J., Wang, X., Oates, T., Gandhi, S., Boedihardjo, A.P., et al. Grammarviz 3.0: Interactive discovery of variable-length time series patterns. ACM Trans Knowl Discov Data 2018;12(1):10:1–10:28. doi:10.1145/3051126.

[24] openeemeter, . Eemeter. 2018. URL: www.github.com/openeemeter/eemeter.

[25] Scikit-learn, . Random forest classifier. 2020. URL: www.scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.

[26] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 2011;12:2825–2830.

[27] The Board of Trustees of the Leland Stanford Junior University, . Visdom: Visualization and insight system for demand operations and management 2016;URL: github.com/ConvergenceDA/visdom.

[28] Lin, J., Keogh, E., Lonardi, S., Chiu, B.. A symbolic representation of time series, with implications for streaming algorithms. 2003, p. 2–11. doi:10.1145/882082.882086.

[29] Senin, P., Malinchik, S.. Sax-vsm: Interpretable time series classification using sax and vector space model. 2013,doi:10.1109/ICDM.2013.52.

[30] Price, P.. Methods for analyzing electric load shape and its variability. Lawrence Berkeley National Laboratory 2010;URL: escholarship.org/uc/item/8gf1w6q4.

[31] Berkeley Lab, . eetd loadshape library. 2017. URL: www.bitbucket.org/berkeleylab/eetd-loadshape.

[32] Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I.. Stl: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics 1990;6:3–73.

[33] Kissock, J.K., Eger, C.. Measuring industrial energy savings. Applied Energy 2008;85:347–361. doi:10.1016/j.apenergy.

2007.06.020.

[34] Najafi, B., Bonomi, P., Casalegno, A., Rinaldi, F., Baricci, A.. Rapid fault diagnosis of pem fuel cells through optimal electrochemical impedance spectroscopy tests. Energies 2020;13(14):3643. URL: dx.doi.org/10.3390/en13143643. doi:10.3390/en13143643.

[35] Scikit-learn, . Mutual info regression. 2020. URL: www.scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.

[36] SciPy, . Pearson correlation. 2020. URL: www.docs.scipy.org.

[37] Pearson, K.. Notes on regression and inheritance in the case of two parents proceedings of the royal society of london, 58, 240-242. 1895.

[38] Wikipedia, . Spearman's rank correlation coefficient. 2020. URL: www.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient.

[39] Daniel, W.W.. The spearman rank correlation coefficient. Biostatistics: A Foundation for Analysis in the Health Sciences 1987;.

[40] Breiman, L.. Random forests. Machine Learning 2001;45(1):5–32.

[41] Glorot, X., Bengio, Y.. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. 2010, p. 249–256.

[42] Crammer, K., Singer, Y.. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of machine learning research 2001;2(Dec):265–292.

[43] Zhao, Y., Zhang, C., Zhang, Y., Wang, Z., Li, J.. A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. Energy and Built Environment 2020;1:149–164. doi:10.1016/j.enbenv.2019.11.003.

[44] Manivannan, M., Najafi, B., Rinaldi, F.. Machine learning-based short-term prediction of air-conditioning load through smart meter analytics. Energies 2017;10(11):1905.

[45] Mosley, L.. A balanced approach to the multi-class imbalance problem. 2013.

[46] scikit learn, . Scikit-learn: Model evaluation. 2020. URL: scikit-learn.org/stable/modules/model_evaluation.html.

[47] Davis, J., Goadrich, M.. The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning. 2006, p. 233–240.

[48] Wikipedia, . Coefficient of determination. 2020. URL: www.wikipedia.org/wiki/Coefficient_of_determination.

[49] Glantz, S.A., Slinker, B.K., Neilands, T.B.. Primer of Applied Regression & Analysis of Variance, ed. McGraw-Hill, Inc., New York; 2001.

[50] Latham, P.E., Roudi, Y.. Mutual information. Scholarpedia 2009;4(1):1658.

[51] Learned-Miller, E.G.. Entropy and mutual information. Department of Computer Science, University of Massachusetts, Amherst 2013;.

[52] Ronaghan, S.. Towards data science - the mathematics of decision trees, random forest and feature importance in scikit-learn and spark. 2018. URL: towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3.

[53] Najafi, B., Di Narzo, L., Rinaldi, F., Arghandeh, R.. Machine learning based disaggregation of air-conditioning loads using smart meter data. IET Generation, Transmission & Distribution 2020;14(21):4755–4762.

[54] Scikit-learn, . Feature selection. 2020. URL: www.scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_selection.

[55] Geurts, P., Ernst, D., Wehenkel, L.. Extremely randomized trees. Machine Learning 2006;63(1):3–42.

[56] Scikit-learn, . Recursive feature elimination. 2020. URL: scikit-learn.org/stable/modules/feature_selection.html#recursive-feature-elimination.

[57] Granitto, P.M., Furlanello, C., Biasioli, F., Gasperi, F.. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. Chemometrics and Intelligent Laboratory Systems 2006;83(2):83–90.

[58] Miller, C., Meggers, F.. Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings. Energy and Buildings 2017;156:360–373. doi:10.1016/j.enbuild.2017.09.056.

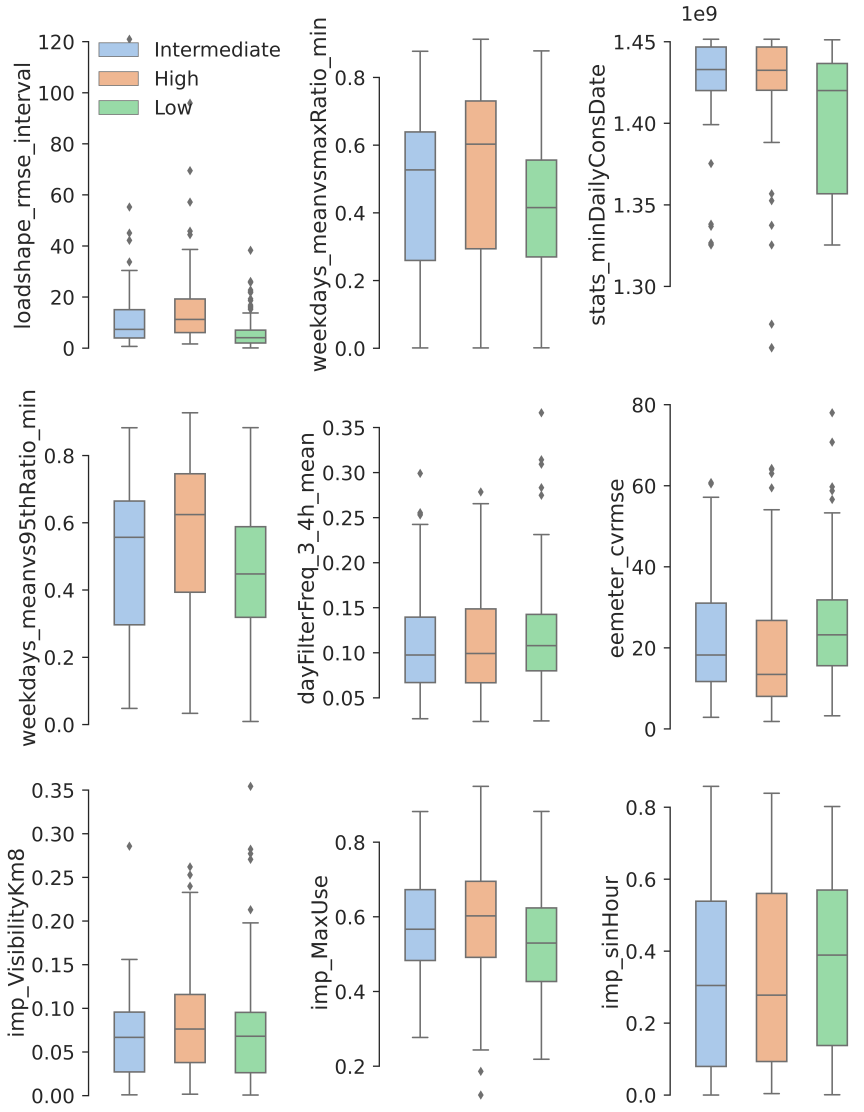


Figure 11: Boxplots of distributions of significant features for performance class target

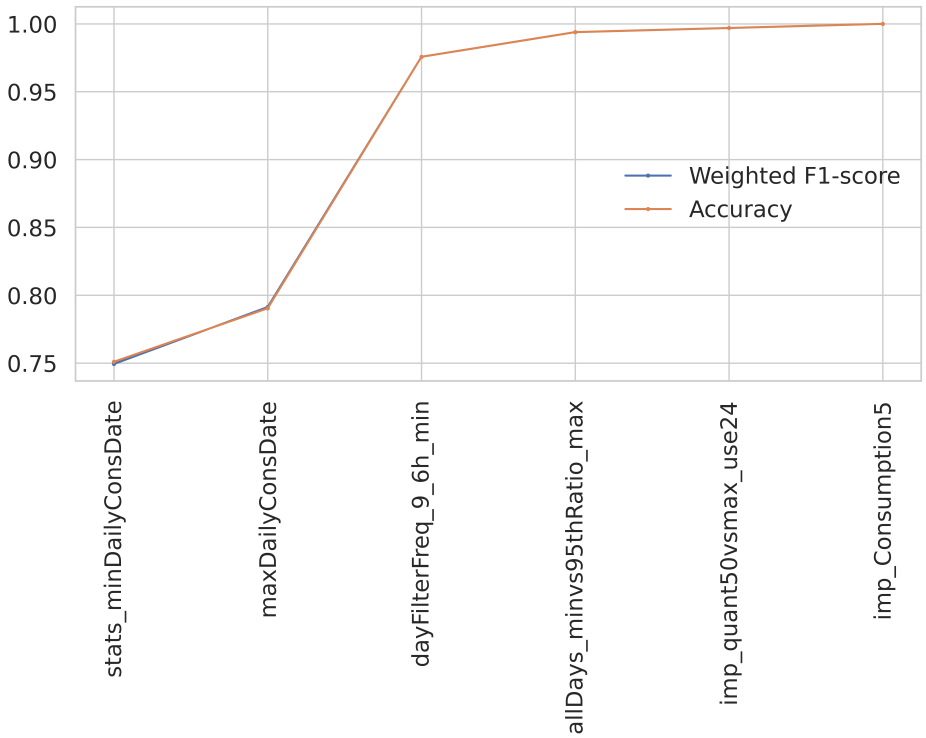


Figure 12: Improvements in the achieved scores after each feature is added to the selected set for the operation group target

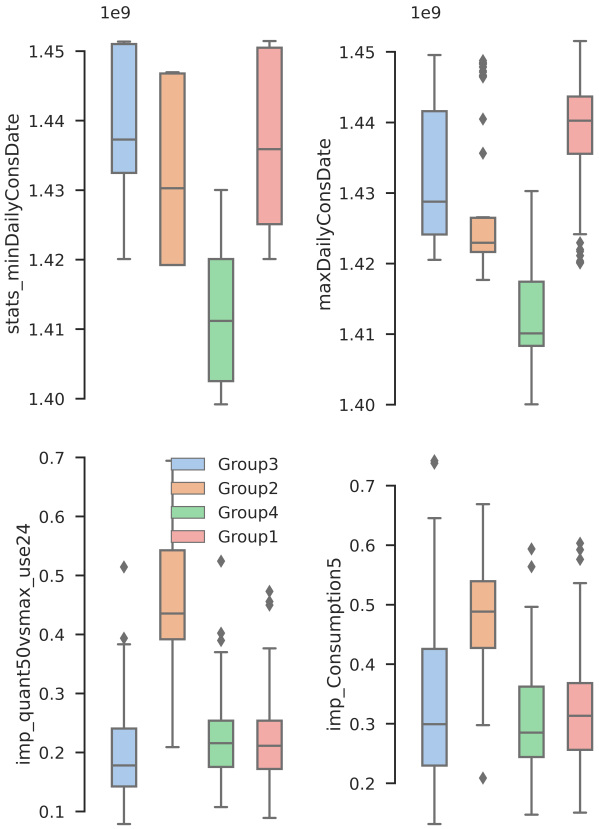


Figure 13: Boxplots of distributions of significant features for operation group target