Sustainable Cities and Society 39 (2018) 508–518

Contents lists available at ScienceDirect

Sustainable Cities and Society

journal homepage: www.elsevier.com/locate/scs

On the feature engineering of building energy data mining

Chuan Zhang[a,b], Liwei Cao[b,c], Alessandro Romagnoli[a,b,*]

[a] School of Mechanical and Aerospace Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore
[b] Cambridge Center for Advanced Research in Energy Efficiency in Singapore Ltd., 1 Create Way, Singapore
[c] Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge, UK

ARTICLE INFO

Keywords:
Building energy
Feature engineering
Exploratory data analysis
Principal component analysis
Random forest.

ABSTRACT

Understanding the underlying dynamics of building energy consumption is the very first step towards energy saving in building sector; as a powerful tool for knowledge discovery, data mining is being applied to this domain more and more frequently. However, most of previous researchers focus on model development during the pipeline of data mining, with feature engineering simply being overlooked. To fill this gap, three different feature engineering approaches, namely exploratory data analysis (EDA) as a feature visualization method, random forest (RF) as a feature selection method and principal component analysis (PCA) as a feature extraction method, are investigated in the paper. These feature engineering methods are tested with a building energy consumption dataset with 124 features, which describe the building physics, weather condition, and occupant behavior. The 124 features are analyzed and ranked in this paper. It is found that although feature importance depends on specific machine learning model, yet certain features will always dominate the feature space. The outcome of this study favors the usage of effective yet computationally cheap feature engineering methods such as EDA; for other building energy data mining problems, the method proposed in this study still holds important implications since it provides a starting point where efficient feature engineering and machine learning models could be further developed.

1. Introduction

Building energy consumption accounts for a considerable portion of the overall energy consumption in contemporary society. The underlying mechanism behind building energy consumption is a complex issue that has attracted research interests worldwide (Fumo, 2014; Sahakian, 2011; Shiraki, Nakamura, Ashina, & Honjo, 2016; Tian, 2013). The efforts aiming at mimicking the dynamics of building energy consumption come from two perspectives: engineering approach and statistical approach. Engineering approach, or equation-based approach (Foucquier, Robert, Suard, Stéphan, & Jay, 2013), strives to describe the interaction between building, energy, and environments, such as the heat and mass transfer process between building envelope and surrounding, in mathematical equations or equivalent modeling techniques. However, since building energy consumption is related with building physicals, meteorological parameters, and occupant behavior as shown in Fig. 1, the relationships between building energy consumption and these influence factors are usually nonlinear and non-stationary. Hence the engineering approach usually fails to capture in an efficient way such inherent nonlinearity and stochastic of building energy consumption. Even though state-of-the-art equation-based software (e.g. EnergyPlus Crawley et al., 2001, TRNSYS Trnsys, 2000) could describe all the components at a very detailed level, it is still pointed out that the engineering method produced simulation results usually present significant gap compared to the field measured data (De Wilde, 2014). The other perspective to look at building energy consumption modeling is the statistical (or data-driven) approach (Noh & Rajagopal, 2013). Such approach starts from data and ends with hypothesis; that is also the reason why in some literature it is named as *inverse method* (Zhao & Magoulès, 2012). Compared to engineering approach, the data-driven approach prefers to discover the algebraic relationships hidden behind the datasets without highlighting the description of physical processes. Actually, the paradigm shifts from equation-driven to data-driven is becoming a trend in many engineering fields (Al-Jarrah, Yoo, Muhaidat, Karagiannidis, & Taha, 2015), especially in the current era of big data (Fan, Xiao, & Yan, 2015). Moreover, the development of machine learning (ML) is pushing forward the advance of building energy analysis and forecasting; various ML techniques are adopted to different questions in this domain (Zhao et al., 2013; Zhou & Yang, 2016). Thus, a prosperous interdisciplinary research area, namely building energy data mining in this paper, is taking shape. Indeed, the capability of data mining and ML to handle

* Corresponding author at: School of Mechanical and Aerospace Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore.
E-mail address: a.romagnoli@ntu.edu.sg (A. Romagnoli).

https://doi.org/10.1016/j.scs.2018.02.016
Received 13 December 2017; Received in revised form 12 February 2018; Accepted 12 February 2018
Available online 09 March 2018

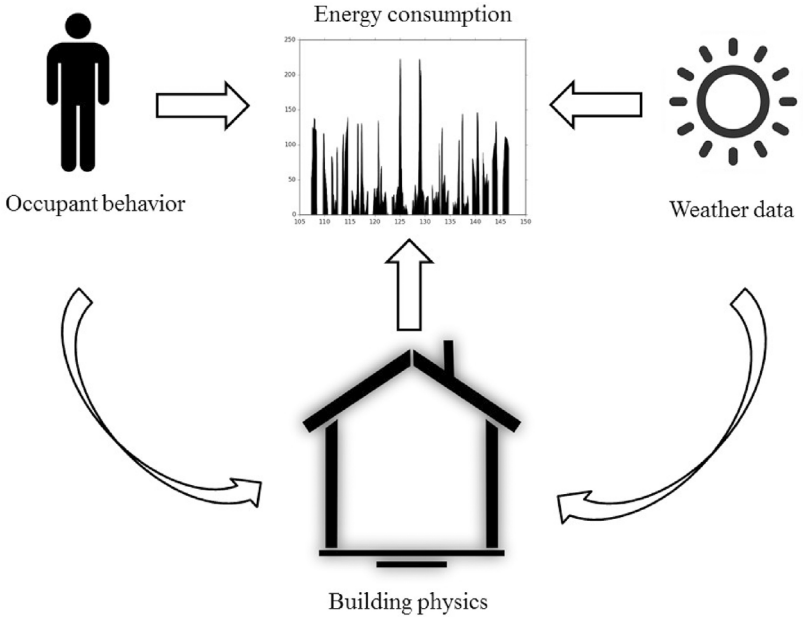


Fig. 1. Interaction relationship between building physics, weather condition, and occupant behavior.

high complexity make it a powerful tool for the building energy analysis and prediction problem (Yu, 2012).

Data mining, also known as knowledge discovery in databases (KDD), is a technique to extract patterns or knowledge from huge amount of data. A typical data mining procedure is shown in Fig. 2 (Yu, Fung, & Haghighat, 2013). From Fig. 2 it can be seen that data mining usually begins with raw data, goes through the procedure of data wrangling (e.g. data selection, cleaning, and preprocessing), feature engineering, machine learning, and it ends with knowledge evaluation. In the context of building energy data mining, most of the current studies focus on machine learning models development and evaluation; for instance, Li et al. applied support vector machine (SVM) to predict hourly building electricity load in their work (Li, Meng, Cai, Yoshino, & Mochida, 2009); Rodger et al. proposed a K-nearest neighbor (KNN) model to predict natural gas demand in public buildings (Rodger, 2014); Ekici et al. also explored the possibility of using artificial neural network (ANN) to predict building energy consumption (Ekici & Aksoy, 2009). However, with the development of data mining techniques, more and more people realize that feature engineering is equally, if not more, important than ML model development during the pipeline of data mining shown in Fig. 2. Feature engineering, defined as "process of using domain knowledge of the data to create features that make machine learning algorithms work more efficiently" (Domingos, 2012), is mainly addressing the question on which factors (referred as features in ML) have the largest effect on the effectiveness and accuracy of ML algorithms. Particularly, as discussed in the last paragraph, building energy consumption is jointly influenced by various features originating from building physics, weather condition, and occupant behavior. So the questions are: which features have more impacts on the ML model effectiveness, which have less? How the feature importances change with ML models? Which features should be used as inputs for different ML models? Are there any efficient and computationally cheap methods to do feature engineering for building energy data mining? These are the questions which will be addressed in this paper.

The rest of this paper is organized as follows: the principles and methods of several different feature engineering methods are discussed in Section 2; these different feature engineering methods are applied to an illustrative building energy dataset example in Section 3, the results are discussed in this section as well; inspired by the results obtained in Section 3, some future perspectives are summarized in Section 4; finally

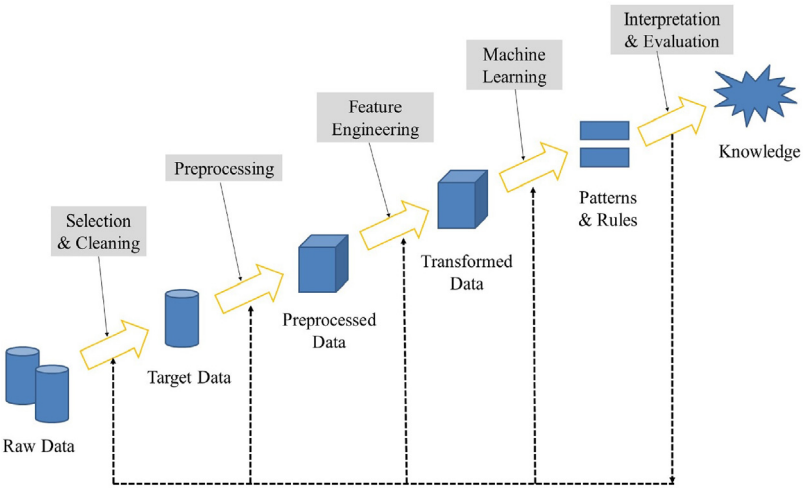


Fig. 2. The typical data mining procedure.

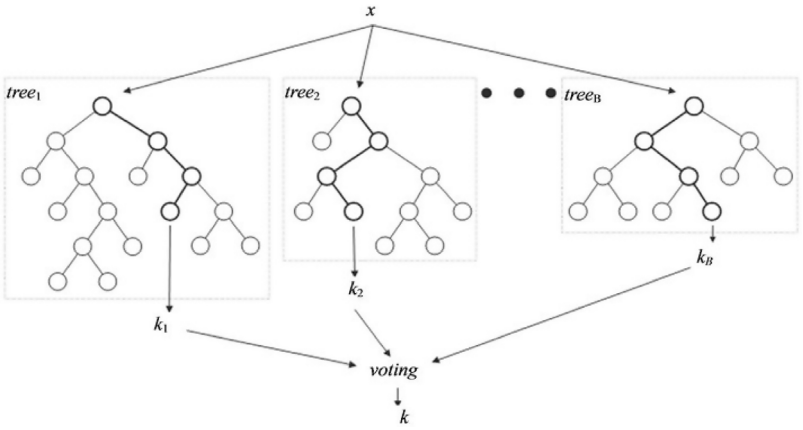


Fig. 3. Schematic of random forest algorithm.

conclusions are given in Section 5.

2. Feature engineering: principles and methods

In data mining, there are two main types of problems: supervised and unsupervised learning. In supervised learning problems, the data are labeled; whereas in unsupervised learning problems, the data are not. In this paper, we are only concerned with the supervised learning problems (either classification or regression); that is to say, for each data record, we already have information about the correct model output, noted as $\boldsymbol{y}$ in the paper. Furthermore, the vector of input features $\boldsymbol{x} = [x^1, x^2, \ldots, x^D]$ compose a feature space $\boldsymbol{\Omega}$. Under such definitions, the objective of feature engineering can be described in the following mathematical expression:

$$\min_{\hat{x},f} \ \|\boldsymbol{f}(\hat{\boldsymbol{x}}) - \boldsymbol{y}\| \tag{1}$$

where $\hat{\boldsymbol{x}} = [x^1, x^2, \ldots, x^K]$ compose a feature subspace $\hat{\Omega} \subseteq \boldsymbol{\Omega}$, given that $K \leqslant D$. $\boldsymbol{f}$ represents the ML models that can transform the input features $\hat{\boldsymbol{x}}$ into outputs so that the distance between predicted output and correct output is minimized. Such distances can be measured by different metrics, Euclidean distance for instance (Davies & Bouldin, 1979). In other words, the overall objective of feature engineering is no more than selecting the optimal feature subspace that gives the prescribed ML model best performance. So it is not difficult to understand that feature engineering is actually dependent on ML models, meaning that the optimal feature subspace could be different for different ML models. As a result, investigation of all feature engineering possibilities under all ML models is a non-trivial work beyond the scope of this paper; in this paper, only three typical feature engineering methods are discussed, namely feature visualization, feature selection, and feature extraction. The detailed procedures for other feature engineering methods with different ML models might be different, yet the general principles remain the same.

2.1. Feature visualization: exploratory data analysis

Feature visualization is a helpful technique that can provide a clear and comprehensive understanding of the feature space; however, feature visualization is not an easy task because usually, the feature space is very high dimensional. So, instead of visualizing the feature space at one time, it is recommended to analyze the pair-wise correlations through exploratory data analytics (EDA). One of the most common EDA techniques is correlation matrix. The correlation matrix is a square matrix based on Pearson product-moment correlation coefficients (Pearson's r), which is a metric for linear dependence between features and outputs. Pearson's r is calculated with the following formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{2}$$

Pearson's r provides a quantitative index for measuring the linear correlation between feature (x) and output (y). If y is perfectly positive linear related with x, r equals to 1; if y is perfectly negative linear related with x, r equals to -1; if y is not linear related with x at all, r equals to 0. Based on such interpretation, Pearson's r based EDA can help to get some basic insights about the linear correlations between outputs and features; hence, the features that are relatively high related to outputs can be chosen as "exploratory feature" for further ML model construction.

2.2. Feature selection: random forest

In cases where the given datasets are high dimensional, it is always suggested to conduct dimensionality reduction through feature selection. The basic idea of feature selection is to remove the features that have less influence on the performance of ML models while only keep the features that are most influential on the ML models. Again it has to be underlined that when ML models are different, the selected features are usually different as well. In this paper, only random forest algorithm is used as an example to show how feature selection works. The schematic of random forest algorithm is shown in Fig. 3; it can be seen from Fig. 3 that random forest is essentially an ensemble model of decision tree classifiers or predictors. The detailed explanations of random forest can be found in literature (Liaw & Wiener, 2002), detailed discussion will not be provided here since the emphasis of this paper is on feature engineering rather than model development.

Feature selection is usually conducted by sequential backward selection (SBS) algorithm. In SBS algorithm, features are sequentially removed from the initial space until the reduced space only contains the desired feature number. The steps of SBS can be noted as:

1. Initialize the algorithm with original feature space dimension D and desired feature subspace K.
2. Remove feature $x^1, x^2, \ldots, x^D$ one by one, use "one versus all" method to get the feature x^- that has the least influence on the model performance.
3. Remove feature $x^- \in [x^1, x^2, \ldots, x^D]$ from original feature space and repeat step 2.
4. Terminate if feature subspace dimension equals to K.

By applying such SBS algorithms, the most important K features in random forest algorithm can be picked up, thus improving the following ML model efficiency.

2.3. Feature extraction: principal component analysis

Besides feature selection, feature extraction is another quite useful aspect for dimensionality reduction in feature engineering. Compared to the former two methods, feature extraction aims to create a new feature subspace by projecting the original feature space with certain rules. principal component analysis (PCA) is perhaps the best-known technique, thus it is chosen as an illustrative example. Implementation of other feature extraction techniques, such as linear discriminant analysis and kernel principal component analysis, can be done in a similar manner (Raschka, 2015).

As its name implies, PCA aims to find the principal component of the features in the sense that the covariance between such component and outputs are largest. In PCA, a $D \times K$ dimensional transformation matrix $\boldsymbol{W}$ is constructed to convert the original feature space $\boldsymbol{x} = [x^1, x^2, \ldots, x^D]$ into a new feature space $\hat{\boldsymbol{x}} = [x^1, x^2, \ldots, x^K]$ to facilitate further analysis. Usually, the transformation matrix is constructed based on the covariance matrix between different features. The covariance between feature x^i and x^j can be calculated as:

$$\sigma_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_k^i - \overline{x^i})(x_k^j - \overline{x^j}) \tag{3}$$

Based on such covariance definition, a $D \times D$ dimensional covariance matrix for feature space $\boldsymbol{x} = [x^1, x^2, \ldots, x^D]$ can be obtained; then by choosing the K largest eigenvalues and the corresponding eigenvectors, the transformation matrix could be constructed. In such a framework, the feature importance is just defined as the ratio between its corresponding eigenvalue and the overall sum of all eigenvalues:

$$\frac{\lambda_i}{\sum_{i=1}^{D}\lambda_i} \tag{4}$$

3. Feature engineering: results

In this section, the aforementioned feature engineering methods are applied to an illustrative dataset. Descriptions of this dataset are given in Section 3.1; the feature importance information with different feature engineering methods are discussed in Section 3.2; finally visualization of feature space is provided in Section 3.3 to give a better understanding about how the proposed feature engineering methods work.

3.1. Description of dataset

The dataset used in this study comes from the Pecan Street Project (Street, 2010). Pecan Street Project is an Energy Internet demonstration project located in Austin, Texas, initialized by U.S. Department of Energy; it monitors the home energy consumption of 1,000 residences of the community in a real-time manner. It also records information about weather data and occupant behavior. It is treated as one of the most comprehensive databases as the testbed for building energy data mining (Glasgo, Azevedo, & Hendrickson, 2016). All the gathered data can be retrieved from a cloud storage named *DATAPORT* that can be freely accessed by academia. In this study, information from the following four tables in the database are used: *electricity-egauge-hours*, *survey-2013-all-participants*, *audits-2013-main*, *weather*. *Electricity-egauge-hours* table stores the electricity consumption information of different buildings collected by Pecan Street's smart meters; *survey-2013-all-participants* table and *audits-2013-main* table store information gathered from the survey and audits conducted in 2013 respectively; *weather* table stores the meteorological parameters. More in detail, variables from *survey-2013-all-participants*, *audits-2013-main* and *weather* are used as input features (shown in Table 1), variables from *electricity-egauge-hours* are used as model output.

From Table 1, it can be seen that the feature space investigated in this paper is 124 dimensional. There are two types of data in the feature space: numerical and categorical. Numerical data are those with quantitative values, such as house volume and temperature; categorical data are the data that are only described qualitatively without numerical values, such as front door orientation and house foundation type. All candidate options for the categorical data could be further referred to the supplemental materials of this paper. Category data are further classified into nominal data and ordinal data; nominal category data cannot be sorted; whereas ordinal category data can. For instance, resident age is ordinal category data whereas the ethical group is nominal. One important data wrangling step is mapping these categorical data into integers through dictionary-mapping approach. Another important issue during data wrangling is handling the missing data. It is common that there are missing values in the dataset due to various reasons, so it is important to come up with a solution that can fill in such null values before further using them for modeling. In the case study, the popular statistics methods are used; more in particular the mode number of the corresponding features are used as placeholders for the missing value. It is assumed that other missing data handling strategies can be implemented with moderate effort, so no detailed discussion would be provided here.

The scale of the targeted dataset in this study is another topic which merits discussion. In the *electricity-egauge-hours* table, *survey-2013-all-participants* table, and *audits-2013-main* table, information about 826, 301, and 67 different buildings are recorded respectively; however there are only 38 buildings in common between them. So these 38 buildings are targeted as research objects in this paper because all the feature information described in Table 1 about them is available. Since the survey and audits were conducted in 2013, the electricity consumption information in 2014 is used for analytics in the study. It is assumed that all information gathered through the survey and audits in 2013 are still up-to-date in 2014. Combining all these tables together, ideally, would lead to a table with $38 \times 24 \times 365$ rows and 124 columns (e.g. features). In reality, a 314121×124 dimensional table can be obtained after querying into the tables above (referred to the supplemental materials of this paper). Although some data entries are missing in our table, the scale of cell numbers is still at 10^8 level. Indeed, it is very hard to apply traditional data analytic techniques to such high volume data without feature engineering, so in the following sections, a systematic feature engineering analysis upon such data is conducted. It is expected that some basic insights about the dataset could be gathered from the feature engineering results so that better understanding about the building energy consumption natures can be achieved.

3.2. Feature importances

By applying the feature engineering methods introduced in Section 2 to the dataset described in Section 3.1, the direct outcome would be rankings of feature importances. The feature importances rankings under different feature engineering methods are shown in Figs. 4–6 respectively. Due to the large amount of features investigated, the feature names are not listed in the figures; the whole list of such feature importance rankings can be referred to the supporting online materials of this paper. In Fig. 4, the features are ranked according to Pearson's r values, which correspond with the linear correlations between features and outputs; similarly, in Figs. 5 and 6, the features are also ranked according to the corresponding importances. From these figures, firstly it can be seen that there are 19 features which have exact zero impact on the output in all three methods, which means there is no relationship between these features and the building overall electricity consumption. Surely these 19 features should be zeroed out during further ML modeling. Secondly, it can be seen from these figures that there are no simple dominant features for building energy prediction problem. Even the most important features in EDA are only 0.3 positively and 0.2 negatively linear correlated with the output, which actually implies a

Table 1
Building energy related features investigated in this study.

Table	Data type	Feature
Audits-2013-main	Numerical (48 in total)	Bedroom number, construction year, conditioned area, house volume, central heat pump number, central AC system number, window AC number, central gas heating number, wall furnace number, gas space heater number, heat recovery system number, electric space heater number, hydroponic heater number, manual thermostats number, digital thermostats number, north window area, northwest window area, west window area, southwest window area, south window area, southeast window area, east window area, northeast window area, north solar screen film area, northwest solar screen film area, west solar screen film area, southwest solar screen film area, south solar screen film area, southeast solar screen film area, east solar screen film area, northeast solar screen film area, skylight number, exterior door number, weather stripped exterior door number, sealed plumbing penetration number, fireplace number, fireplace vented to outside number, fireplace with damper number, fireplace with external combustion number, hourly air change, attic floor square footage, attic R value, attic average insulation depth, radiant barrier, window number, window shading, distance from neighbors.
	Categorical nominal (3 in total)	Front door orientation, foundation type, home type
Survey-2013-all	Numerical (33 in total)	Number of rooms, total square feet, male sex number, female sex number, vehicle number, ceiling fans number, compressor number, summer temperature weekday hours, summer temperature weekend hours, summer temperature sleep hours, winter temperature weekday hours, winter temperature weekend hours, winter temperature sleep hours, thermostat number, indoor thermal comfort, dishwasher number, refrigerator number, cloth-washer number, cloth dryer number, water heater number, oven range number, micro-oven number, toaster-oven number, TV number, ceiling fans number, power tool number, EV number, sprinkler number, swimming pool number, electric cable box number, electric dryer number, electric router number
	Categorical nominal (19 in total)	Weekly schedule, ethnicity group, education level, total annual income, smart phone own, tablet own, pv own, building retrofits, appliance own, irrigation system, care about energy cost, willing to reduce energy cost, HVAC type, heating type, pets own, programmable thermostat, cooling and heating even, AC filter change frequency, light bulbs type
	Categorical Ordinal (7 in total)	Resident age, weekday cooking timetable, weekend cooking timetable, summer blind usage, winter blind usage, thermostat setting, TV hours
Weather	Numerical (14 in total)	Latitude, longitude, ozone, temperature, dew point, humidity, visibility, apparent temperature, pressure, wind speed, cloud cover, wind bearing, precipitation intensity, precipitation probability.

weak relationship between such features and the output. This further supports the argument that building energy consumption is jointly, to some extent evenly, influenced by various features; which proofs the complexity of building energy consumption.

To assess the most important features more precisely, the top 20 most important features in all three feature engineering methods are shown in Fig. 7–9 respectively. The features are further classified into top 10 positively important features in Fig. 7a and top 10 negatively important features in Fig. 7b since EDA provides not only the magnitude of correlations but also the sign of correlations. Through comparison of the top 20 important features for three different methods, it is found that there are 10 common features among them, namely *house volume, temperature, construction year, house square feet, air conditioned area, ACH50 calculation (index of house air tightness), bedroom number, attic insulation, irrigation system*, and *13-18 years old residents number*. Moreover, there are nine features appearing more than twice in these three methods, namely *TV number, PV system, EV number, total window area in the south direction, sprinkler system number, total window area in the west direction, humidity*, and *distance from neighbors*. Compared to features listed in Table 1, these features can be treated as the so-called major features that will mostly affect the building energy consumption; intuitively, by applying the feature engineering methods, the feature space dimension can be reduced from 124 to 19.

3.3. Feature space visualization

To better illustrate how feature engineering help ML modeling, some feature space visualization effort is shown in this section.

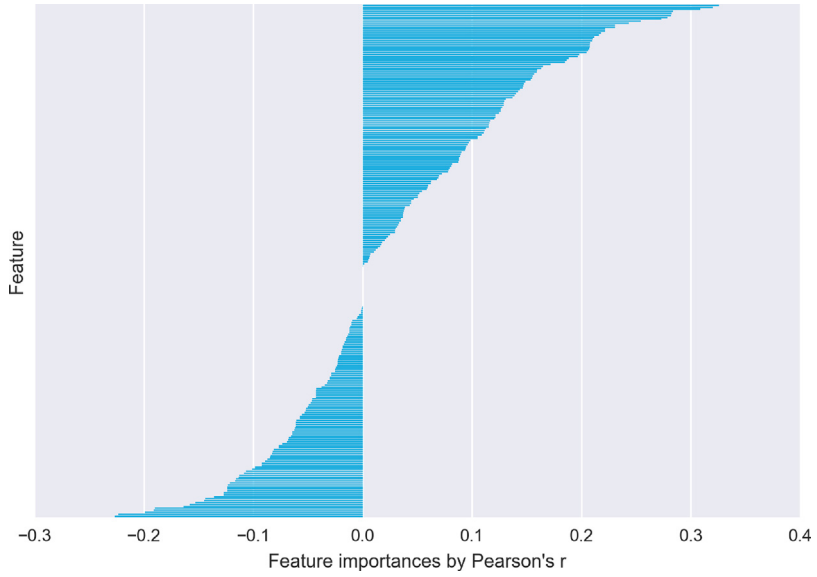


Fig. 4. Feature importances in exploratory data analysis.

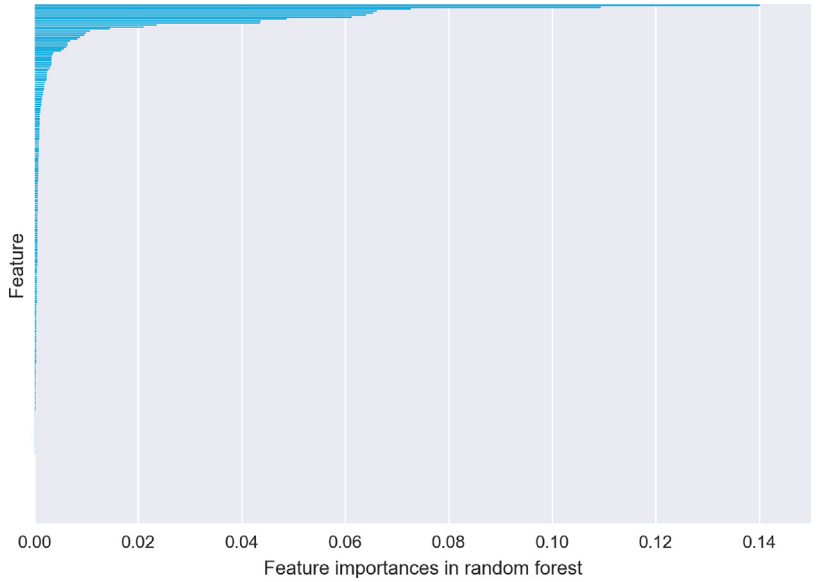


Fig. 5. Feature importances in random forest.

Although it is quite difficult to visualize the feature distribution in high dimensional space, there are still several useful techniques that could give us some fundamental insights. Pairwise scatter plot is one of them. In pairwise scatter plot, the features are plotted against output such that the one-dimensional distribution of output in the feature space can be obtained. In Fig. 10, pairwise scatter plot in the original feature space is shown; only the five most important features, namely apparent temperature, house volume, 13–18 residents number, ACH50 air tightness index, and house square feet data, together with the hourly electricity consumption (in the unit of kWh) are analyzed in this figure as proof of conception. It has to be noted that all the features shown in Fig. 10 are normalized non-dimensional data to achieve higher PCA accuracy. Similarly, in Fig. 11, pairwise scatter plot in the transformed feature space through PCA is shown, again only the five most important features are explored in the plot. Principal Component in Fig. 11 are non-dimensional variables in the new feature space without any real physical meaning. In Fig. 10, it can be seen that the outputs are largely clustered, especially among the feature dimension where feature data type is categorical; for instance, the "electricity consumption versus 13–18 residents" sub-figure in Fig. 10. Since the reported number of 13–18 years old residents are categorized, as a result the data distribution in this dimension becomes striped. However, through the feature transformation of PCA, the distribution of data becomes much more sparse. For instance, the "electricity consumption versus PC2" sub-figure in Fig. 11, although there is still no obvious patterns in this figure, yet compared to the original feature space, the distribution has been widely flattened, which of course would make it easier to develop more efficient predictors. Similar patterns can be observed from the visualization at other dimensions in the PCA-transformed feature space. This means that, feature engineering not only helps to significantly reduce the feature space dimension but also contributes to construct a new feature space where data are more sparsely distributed, therefore ML models are easier to develop.

4. Future perspectives

In the last section, the results of applying feature engineering methods to the Pecan Street database are analyzed. It is shown that feature engineering can not only help to reduce the feature space dimension, but also can reshape the feature space to make ML modeling

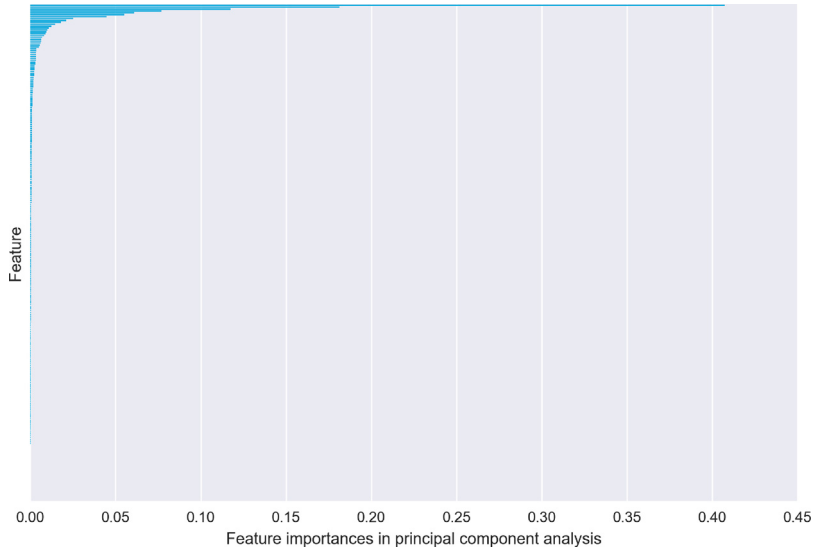
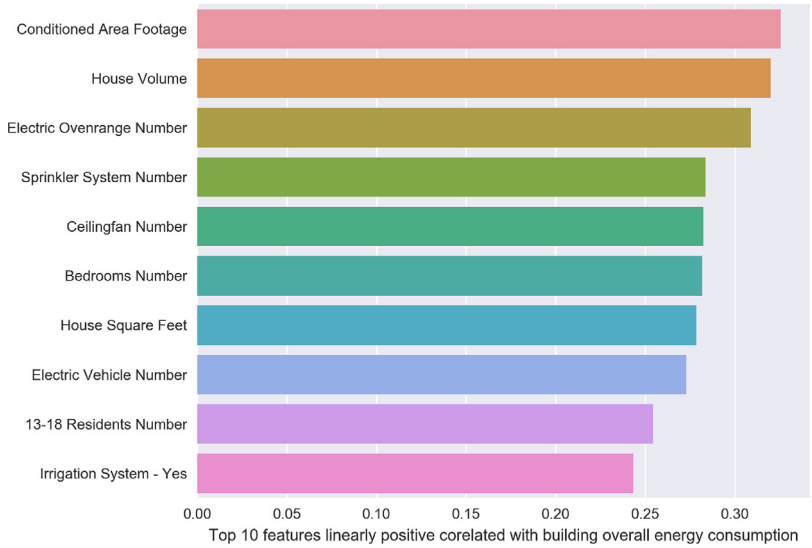
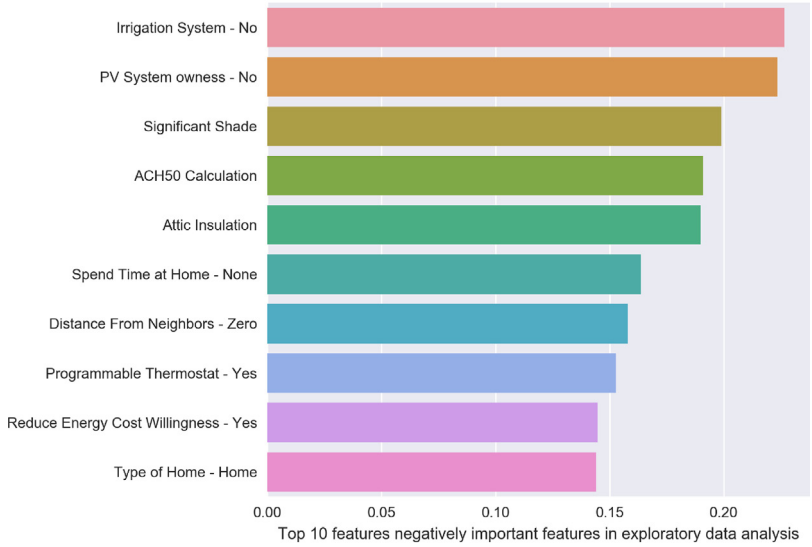


Fig. 6. Feature importances in principal component analysis.



(a) Top 10 positively important features in EDA



(b) Top 10 negatively important features in EDA

Fig. 7. Top 20 important features in EDA.

easier. However, merely based on results from one dataset, it is not enough to cover all perspectives in building energy data mining problems; so in this section, we would like to inspire several new perspectives to further look at the feature engineering problem in building energy data mining.

4.1. Use of domain knowledge

Just as discussed in Section 1, ideally, feature engineering is a process that should combine domain knowledge and machine intelligence. However, in Section 3 no domain knowledge has been used in the feature engineering; this is designed like so on purpose, because we just feed all the data available to the feature engineering and let itself pick up whatever is useful. On one hand, this is good because this is exactly where the advantages of machine intelligence come from; on the other hand, this is not a good practice because handling too many features might bring additional computational cost. Moreover, given the feature engineering results in Section 3, it is found that there are three kinds of rules discovered through feature engineering analytics: the first type aligns with our domain knowledge well. For instance, building overall energy consumption is positively correlated with the house volume, negatively correlated with the attic insulation. Such relationships are something that can even be projected before conducting feature engineering, the results of feature engineering further confirm our hypothesis; the second type of rules might not align with human domain knowledge, yet do make sense upon further reflection. For instance, building overall energy consumption is positively correlated with the 13–18 years old resident number in the building. Such covariance might be difficult to project before feature engineering, yet once being discovered through feature engineering, is supported by the domain knowledge as well. The third kind of rules is actually quite difficult to understand even after feature engineering has pointed them out. For instance, it is very hard to realize that hairdryer use pattern would be an important feature in random forest ML models even though the big data analytics insights clearly specifies it. It leads to a question: should we try to understand the machine intelligence based on human domain knowledge, or is that viable? Another problem is if the dataset described in Section 3.1 was given to an expert in building energy

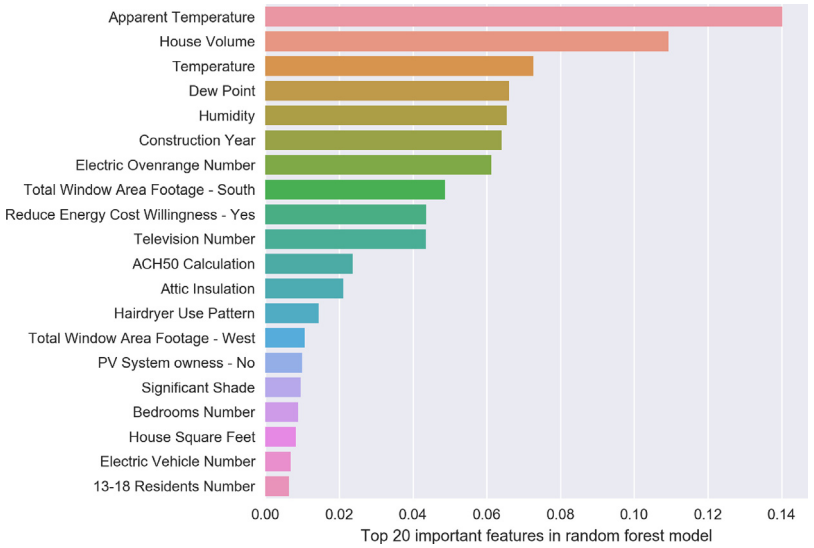


Fig. 8. Top 20 important features in random forest.

systems, and if he or she was asked to pick up the top 20 most important features by his domain knowledge, what is the possibility that he or she would pick up the same features as machine intelligence does? Furthermore, if we feed the ML models a dataset tailored by domain knowledge beforehand, will the performance of ML models increase or decrease? These are all open questions that merit further discussion.

4.2. Curse of dimensionality

In Section 3.2 the top 20 most important features are selected for each feature engineering method, but 20 is indeed an arbitrarily selected number, why not 10 or 30? Furthermore, how many features would enable the best performance of the ML models? This leads to a curse of dimensionality problem in feature engineering. To better explain this problem, the relationship between ML model (use random forest as an example) performance and feature numbers is investigated in this paper. More in particular, 70% of the given dataset is used to train the random forest model under different feature numbers (e.g. the top K most important features shown in Fig. 8), the remaining 30% dataset is used to label the model performance. Again the Pearson's r (refer to Section 2.1) is used as a performance index to measure the model prediction accuracy: if the predicted values perfectly align with the real values, r equals to 1; in the worst case, r equals to 0. Such curse of dimensionality analysis for random forest model is shown in Fig. 12. It can be seen from Fig. 12 that there is an optimal feature number that would enable the best performance of the proposed random forest model. In our case, when around 12 features are used, the model produced best results. If less features are used, there is an under-fitting problem, which means that the random forest model is not able to capture the inherent dynamics in the training dataset, sequentially the prediction performance is poor; if more features are used, there is over-fitting problem, which means the random forest model uses too many parameters to follow the patterns in the training dataset; as a result, its performance on the test dataset is relatively poor due to its high complexity. Again, the open question here is: it seems that 12 is the threshold of feature numbers in this random forest building energy data model, yet what about the other models for other datasets? How can the optimal feature numbers for these models be found? Is there a general method to conduct such curse of dimensionality analysis for different ML models in the context of building energy data mining? This is also an interesting question that deserves being investigated.

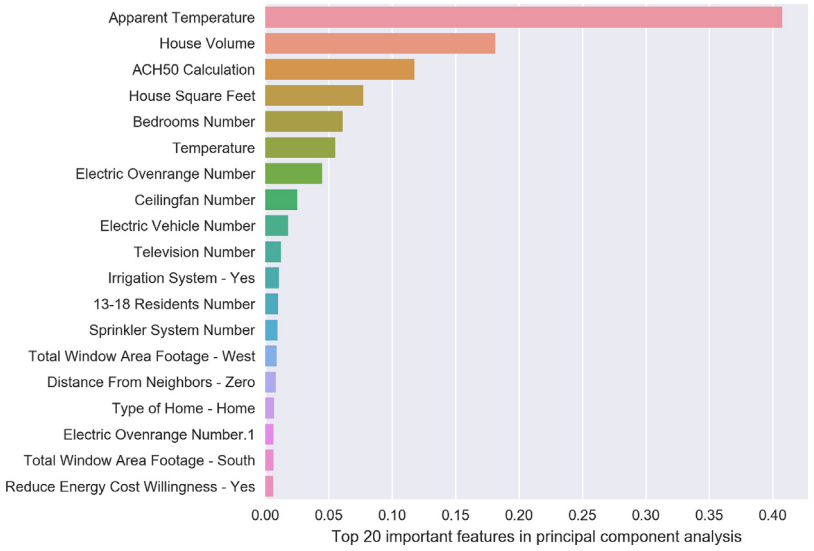


Fig. 9. Top 20 important features in PCA.

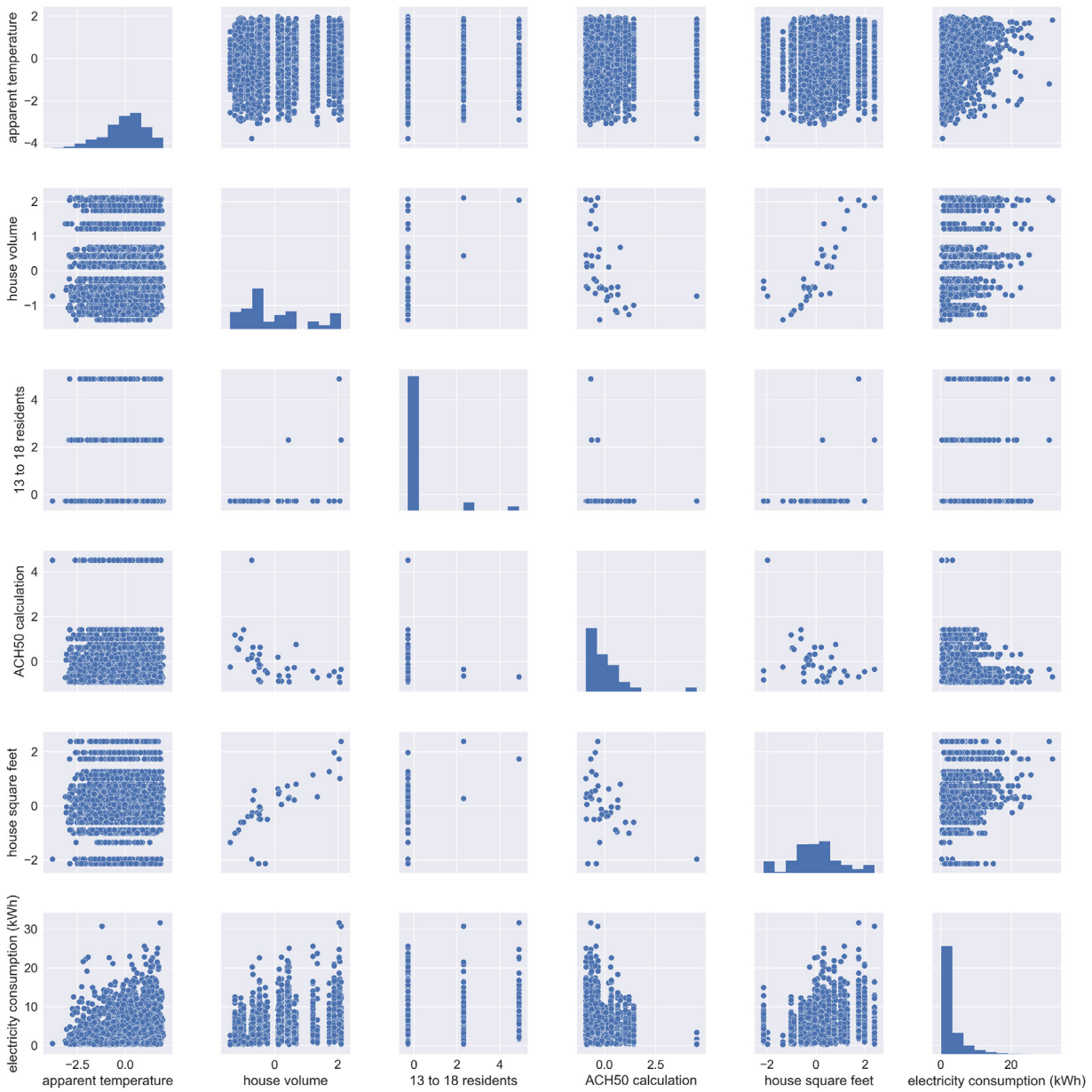


Fig. 10. Scatter plot between building energy and the 5 most important features in original feature space.

4.3. Computational cost

The final topic explored in this section is computational cost. It is mentioned in Section 2.2 that there is high overlap among the top 20 important features discovered by the three different feature engineering methods covered in this study. However, the computational costs to get such insights from these methods are entirely different. On a PC with Intel Core i7-47900 CPU 3.6 GHz × 2, 12 GB RAM, the implementation of EDA in a Python 3 environment only takes 10 min, whereas PCA and random forest take 30 min and 50 min respectively. So if the most important features obtained from all feature engineering methods are almost identical, why not just use the most computationally cheap ones? Starting from here, the performance of random forest model when different feature sets are used as its input are compared. It turns out that the model performance is almost identical when the top 20 most important features discovered from the three different methods are used as inputs. So this study recommends the usage of simple and computationally cheap feature engineering methods, such as EDA, to conduct feature selection task no matter what ML models would be used sequentially. At least, more advanced feature selection and extraction methods do not show advantages in this paper.

5. Conclusions

Feature engineering in building energy data mining is discussed in this paper. Three different feature engineering methods, namely feature visualization, feature selection, and feature extraction, are discussed in the paper. Exploratory data analysis (EDA), random forest (RF), and principal component analysis (PCA) are used to implement feature visualization, feature selection, and feature extraction respectively. By applying such methods to the Pecan Street Project database, some insights about the feature space of building energy data mining can be obtained.

The feature space for building energy data mining problem is usually quite high dimensional. Unfortunately, there is no single dominant dimension in such feature space: the most important features in EDA are only 0.3 positively and 0.2 negatively linearly correlated with building energy consumption, which implies a very weak

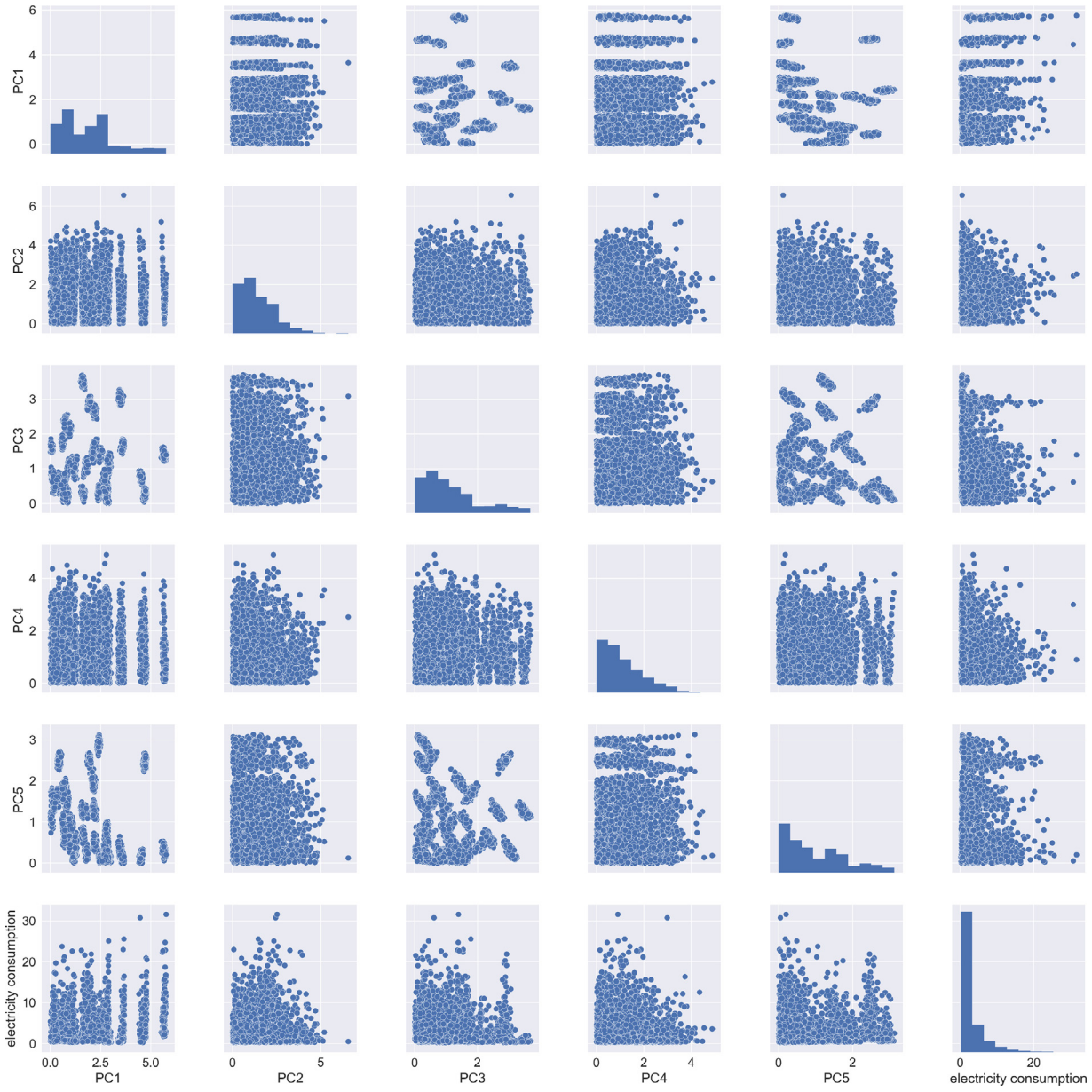


Fig. 11. Scatter plot between building energy and the 5 most important features in PCA-transformed space.

relationship between them. In this sense, building energy consumption is jointly, to some extent evenly, influenced by multiple features from various domains (*e.g.* building physics, weather condition, and occupant behavior). Feature extraction methods such as PCA can transform the original feature space into a new principal component space where machine learning (ML) models are easier to develop.

The three different feature engineering methods discussed in this paper result in three different feature importance rankings. These rankings sort the features according to their relative impact on the corresponding ML model outputs. There are certain features that appear in all three rankings; they are house volume, temperature, construction year, house square feet, air conditioned area, ACH50 calculation (index of house air tightness), bedroom number, attic insulation, irrigation system, and 13–18 years old residents number. These features should be further detailed and investigated to develop better ML models. In other building energy data mining problems with different objectives and/or variables, the feature importance might be different, yet the proposed method provides a good starting point where efficient machine learning models could be further developed.

Use of domain knowledge during feature engineering should be very cautious because the analysis in this study shows that sometimes it is very difficult to understand the machine intelligence based on human being's domain knowledge. As a result, certain features, which are important for ML models, might be wrongly crossed out according to human domain knowledge. In addition to this, curse of dimensionality is another problem that can be observed from this study. Proper feature selection methods should be used to identify the optimal number of features that should be used as input for different ML models. This is also an important research question that needs to be further explored.

Acknowledgements

This research project is funded by the National Research Foundation Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) program.

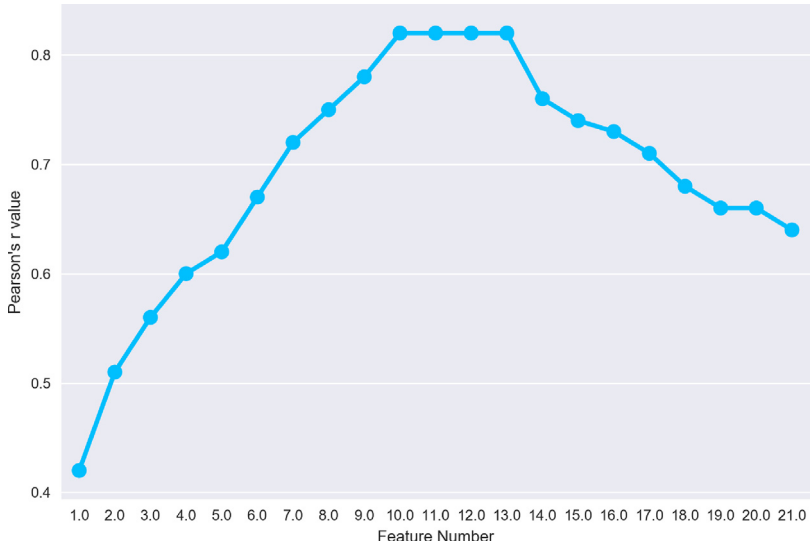


Fig. 12. Curse of dimensionality in random forest model.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.scs.2018.02.016.

References

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research, 2*(3), 87–93.

Crawley, D. B., Lawrie, L. K., Winkelmann, F. C., Buhl, W. F., Huang, Y. J., Pedersen, C. O., et al. (2001). Energyplus: Creating a new-generation building energy simulation program. *Energy and Buildings, 33*(4), 319–331.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2*, 224–227.

De Wilde, P. (2014). The gap between predicted and measured energy performance of buildings: A framework for investigation. *Automation in Construction, 41*, 40–49.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78–87.

Ekici, B. B., & Aksoy, U. T. (2009). Prediction of building energy consumption by using artificial neural networks. *Advances in Engineering Software, 40*(5), 356–362.

Fan, C., Xiao, F., & Yan, C. (2015). A framework for knowledge discovery in massive building automation data and its application in building diagnostics. *Automation in Construction, 50*, 81–90.

Foucquier, A., Robert, S., Suard, F., Stéphan, L., & Jay, A. (2013). State of the art in building modelling and energy performances prediction: A review. *Renewable and Sustainable Energy Reviews, 23*, 272–288.

Fumo, N. (2014). A review on the basics of building energy estimation. *Renewable and Sustainable Energy Reviews, 31*, 53–60.

Glasgo, B., Azevedo, I. L., & Hendrickson, C. (2016). How much electricity can we save by using direct current circuits in homes? Understanding the potential for electricity savings and assessing feasibility of a transition towards dc powered buildings. *Applied Energy, 180*, 66–75.

Li, Q., Meng, Q., Cai, J., Yoshino, H., & Mochida, A. (2009). Applying support vector machine to predict hourly cooling load in the building. *Applied Energy, 86*(10), 2249–2256.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News, 2*(3), 18–22.

Noh, H. Y., & Rajagopal, R. (2013). Data-driven forecasting algorithms for building energy consumption. *SPIE Smart Structures and Materials + Nondestructive Evaluation and Health Monitoring, International Society for Optics and Photonics* pp. 86920T–86920T.

Raschka, S. (2015). *Python Machine Learning.* Packt Publishing Ltd.

Rodger, J. A. (2014). A fuzzy nearest neighbor neural network statistical model for predicting demand for natural gas and energy cost savings in public buildings. *Expert Systems with Applications, 41*(4), 1813–1829.

Sahakian, M. D. (2011). Understanding household energy consumption patterns: When "west is best" in metro manila. *Energy Policy, 39*(2), 596–602.

Shiraki, H., Nakamura, S., Ashina, S., & Honjo, K. (2016). Estimating the hourly electricity profile of Japanese households-coupling of engineering and statistical methods. *Energy, 114*, 478–491.

Street, P. (2010). *The Pecan Street Project*Austin, TX: Working Group Report.

Tian, W. (2013). A review of sensitivity analysis methods in building energy analysis. *Renewable and Sustainable Energy Reviews, 20*, 411–419.

Trnsys, A. (2000). *Transient System Simulation Program.* University of Wisconsin.

Yu, Z., Fung, B. C., & Haghighat, F. (2013). Extracting knowledge from building-related data-a data mining framework. *Building Simulation, vol. 6* (pp. 207–222).

Yu, Z. (2012). *Mining hidden knowledge from measured data for improving building energy performance.* Ph.D. thesis), Concordia University.

Zhao, H.-x., & Magoulès, F. (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews, 16*(6), 3586–3592.

Zhao, J., Yun, R., Lasternas, B., Wang, H., Lam, K. P., Aziz, A., et al. (2013). Occupant behavior and schedule prediction based on office appliance energy consumption data mining. *CISBAT 2013 Conference-Clean Technology for Smart Cities and Buildings,* 549–554.

Zhou, K., & Yang, S. (2016). Understanding household energy consumption behavior: The contribution of energy big data analytics. *Renewable and Sustainable Energy Reviews, 56*, 810–819.