Recommendation system infrastructure for the energy efficiency of buildings

Loup-Noé Lévy[1,2], Guillaume Guerard[3], and Soufian Ben Amor[1]

[1] LI-PARAD Laboratory EA 7432, Versailles University, 55 Avenue de Paris, 78035 Versailles, France; f_author.s_author}@uvsq.fr
[2] Energisme, 88 Avenue du Général Leclerc, 92100 Boulogne-Billancourt France; {f_author. s_author}@energisme.com
[3] De Vinci Research Center, Pole Universitaire Léonard de Vinci, 12 Avenue Léonard de Vinci, 92400 Courbevoie, France ; {f_author. s_author}@devinci.fr

Abstract. Reducing building energy consumption is an important challenge of the XXI century. Such a challenge can be addressed by an energy actor, the Trusted Third Party for Energy Measurement and Performance, by identifying consumption profiles, making diagnostics, proposing energy performance enhancement and evaluating the energy savings resulting from these changes. A thesis was funded by such an actor to study and solve these issues. To accomplish this, the first step is recognising the energy actors as complex sociotechnical systems and identifying the appropriate modeling approach. In this article a detailed data infrastructure adapted to the big data challenges raised by the transformation of today's energy sectors is then proposed. It is designed to address governance guidelines through DevOps methodology. Finally the results of a hierarchical clustering algorithm based on pretopology are briefly presented along with future works.

Keywords: Big Data · Machine Learning Factory · Energy Efficiency.

1 Introduction

1.1 Context

Buildings account for 44% of the energy consumed, ahead of transport (32%) and industry (21%) and by 2050, global energy demand is expected to double. Current networks and production will not be able to handle this load without a profound change in the way we consume, which includes increased energy efficiency.

In a given process, energy efficiency is the ratio of the energy value produced to the energy value consumed. Energy efficiency is the total energy output of a system. It is derived from the performance in terms of what the energy will be used for. The energy efficiency of a house will thus concern the heating, by taking into account the energy output of the appliances, but also insulation, ventilation, etc. Improving energy efficiency requires knowledge of consumption data, change management, building materials and the implementation of an intelligent system.

To know the energy efficiency of a building, the energy performance diagnosis requires to evaluate the energy consumption of the building and its management. The mission of energy saving certificates is to improve the energy efficiency of the residential and tertiary building sectors, transport, industry and networks.

1.2 Objectives

Facing these challenges, the need for a Trusted Third Party for Energy Measurement and Performance (TTPEMP) is felt. These independent actors can make energy efficiency diagnostics and they can propose relevant action to enhance the performance of the building. Theses recommendations can be in term of building renovation or in term of changes in behaviors (humans or non-humans). Its purpose is also to evaluate the energy saving allowed by the energy efficiency recommendation by comparing consumption after modifications to previous consumption [3].

The TTPEMP must be able to process data coming from multiple sources and of heterogeneous format. Aftermath, the TTPEMP propose some strategy to enhance the energy efficiency of a building based on those which give good performance on similar buildings. Thus, the process rise its own knowledge by providing results on each strategy used on similar buildings. Those strategies include building renovation, management of device (position and consumption) and change management concerning the occupants. In the literature, a recommendation system is based on filtering algorithms or nearest neighbors. But the problem is more complex and the data includes numerical, categorical and time series and sometimes outcomes of specific strategies. The goal of the thesis is to propose a recommender system for the energy efficiency of buildings and to perform forecast about the building future consumption after applying any strategy.

2 Recommender system's architecture

2.1 Big Data paragidm

The data treated by the TCMPE are of great **Volume** as it must treat several years of consumption history for hundreds of thousands of buildings. Naturally energy consumption is **constantly being updated** as new consumption is added to the time series causing **Velocity** challenge. Because it is coming from all kinds of sources the quality and format of the data is **Heterogeneous**, causing **Veracity** and **Variety** challenges. The infrastructure must respond to two more constraints: the lineage and governance. Data lineage allows a visualization of the life cycle of the data in order to answer the following questions: from which source does this data come, and what transformations has it undergone. The objective of data governance is to improve data quality, relevance and integrity. The information is thus enriched and enhanced. Because of these challenges, the TCMPE has to have an adequate data infrastructure, presented below.

2.2 A socio-technical complex systems

To understand the consumption of buildings and to propose an architecture for buildings consumption clustering and forecasting, we must study buildings, and moreover the organisation and systems dealing with the consumption of those buildings, such as the TCMPE but also the local authorities as well as the smart grid and other intelligent energy systems. All those systems are complex systems and moreover socio-technical complex systems. Their modeling must take into account that aspect while also considering the existing infrastructure [3]. The methodology for such modeling was the subject of articles describing the early result of the systemic analysis of such system [1, 2, 12].

2.3 Architecture of the recommender system

Figure 1 presents the recommender system architecture developed during the thesis. In the following text, we will attempt to present it succinctly.

The energy data is massive, multiple and heterogeneous, responding to the Big Data paradigm. It is therefore stored in its raw form in a Data Lake (B). The preprocessing of the data will be done in different data marts. As the preprocessing is specific to the machine learning methods that we want to apply, several different datamarts exist and will be specific to the different issues related to energy performance. In our system, we foresee 3 final datamarts. A datamart for unsupervised profiling (C), a datamart for semi-supervised profiling (D) and a datamart for consumption prediction (E). Other datamarts concerning different problems may be designed during the thesis (F).

An important aspect of preprocessing for machine learning is the extraction of features, they are also specific to the methods to be applied. For profiling, the features used will be the elements that allow to distinguish categories of consumption relevant from the point of view of energy performance. For example, night-time consumption can give indications of unnecessary energy consumption. In addition, high consumption at certain times of the day or week can provide information on the type of use of the building (housing, office, sports hall, business). In return, the type of use of the building can be a new characteristics added to the datamart in order to make finer and more relevant diagnostics. The building profiles that will be identified are not necessarily defined in advance, so it refers to unsupervised learning (C). The profiles identified by our algorithms will then be analyzed in order to extract specific diagnostics and recommendations (H).

One of the objectives of our system is to monitor the evolution of the performance of the buildings and thus to check if they change their category of consumption profile. These categories must be fixed since we cannot follow the passage from one category to another if the categories themselves are moving. Placing an item in a set of predefined categories is semi-supervised learning. This refers to the second type of methods (D). This will allow us to identify changes in building consumption (change in type of activities, deterioration of the building) and to evaluate whether the recommended energy performance actions have been effective (I).

Finally, another branch of the system will consist in comparing the future consumption of a building with its actual consumption. This will allow to evaluate the energy savings resulting from the proposed energy performance actions (J). It will also allow to identify abnormal consumption of buildings.

Results from step (I) and (J) are used to reinforce the machine learning algorithms. Moreover, the strategies to enhance the energy efficiency of building will provide real results that can be evaluated to provide useful feedback for the decision making process done during these two steps.

In order to stay on top of advances in machine learning it is necessary to be able to use new methods and train new models easily as well as select and hybridize the best performing model(s) over time. This will be done with the help of Energisme's machine learning factory and the development of complex inference graphs (G).

Finally, data lineage will be used to allow the adaptation of the system to the evolution of the data and the context, as well as to ensure the quality and the veracity of the exploited data.

3 First results and new algorithms

A building clustering algorithm must be performed on a mixed dataset composed of both numerical and categorical data (with time series). It must be able to treat a large amount of element with a relatively high number of characteristics. Such an algorithm, based on the mathematical structure called pretopology, has been proposed. Because this algorithm is innovative, its details and its early results have been published [11]. Other algorithms could be relevant for buildings clustering, such has AdaBoost/Random Forest [6], K-prototype [9], Self Organizing Maps [8], Adaptive Resonance Theory-based Topological Clustering [14] and comparing their results will be the subject of another article.

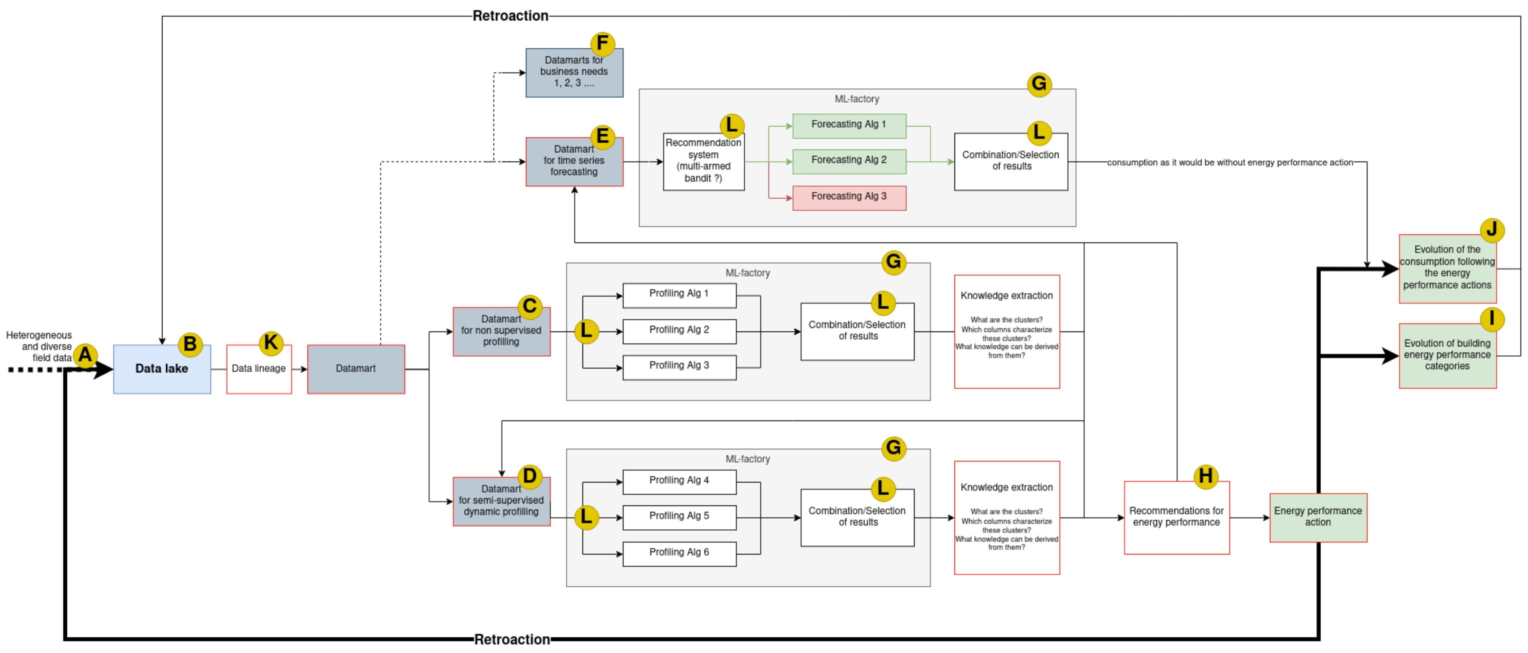
We also implement clustering algorithms specialized to time series. The algorithms include KNN with dynamic time warping [16], Time Series Forest Classifier [13], cBOSS [15] and cRISE [5].

For the prediction of consumption time series, the first algorithms considered are XGBoost [4], LightGBM [10] and a hybrid deep-learning method we developed [7]. The machine learning factory training and comparing these methods on a building dataset will be the main topic of a future article.

Aknowledgement

This paper is the result of research conducted at the energy data management company *Energisme*. We thank *Energisme* for the resources that have been made available to us.

Fig. 1. The proposed recommender system


Retroaction
Heterogeneous and diverse field data
A
B
Data lake
K
Data lineage
Datamart
F
Datamarts for business needs 1, 2, 3 ....
E
Datamart for time series forecasting
G
ML-factory
L
Recommendation system (multi-armed bandit ?)
Forecasting Alg 1
Forecasting Alg 2
Forecasting Alg 3
L
Combination/Selection of results
consumption as it would be without energy performance action
C
Datamart for non supervised profilling
L
G
ML-factory
Profiling Alg 1
Profiling Alg 2
Profiling Alg 3
L
Combination/Selection of results
Knowledge extraction
What are the clusters?
Which columns characterize these clusters?
What knowledge can be derived from them?
D
Datamart for semi-supervised dynamic profilling
L
G
ML-factory
Profiling Alg 4
Profiling Alg 5
Profiling Alg 6
L
Combination/Selection of results
Knowledge extraction
What are the clusters?
Which columns characterize these clusters?
What knowledge can be derived from them?
H
Recommendations for energy performance
Energy performance action
J
Evolution of the consumption following the energy performance actions
I
Evolution of building energy performance categories
Retroaction

References

1. Amor, S.B., Guerard, G., Levy, L.N.: Systemic approach for modeling a generic smart grid. In: Proceedings of the Tenth International Symposium on Information and Communication Technology. pp. 15–22 (2019)
2. Amor, S.B., Tran, H.: Modeling and recommendation system for improving the energy performance of buildings. In: Distributed Computing and Artificial Intelligence, Volume 2: Special Sessions 18th International Conference. vol. 2, p. 206. Springer Nature
3. Bosom, J.: Conception de microservices intelligents pour la supervision de systèmes sociotechniques : application aux systèmes énergétiques. Ph.D. thesis (2020), http://www.theses.fr/2020UPSLP051, thèse de doctorat dirigée par Bui, Marc et Ben Amor, Soufian Université Paris sciences et lettres 2020
4. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., et al.: Xgboost: extreme gradient boosting. R package version 0.4-2 **1**(4), 1–4 (2015)
5. Flynn, M., Large, J., Bagnall, T.: The contract random interval spectral ensemble (c-rise): the effect of contracting a classifier on accuracy. In: International Conference on Hybrid Artificial Intelligence Systems. pp. 381–392. Springer (2019)
6. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences **55**(1), 119–139 (1997)
7. Guerard, G., Pousseur, H., Taleb, I.: Isolated areas consumption short-term forecasting method. Energies **14**(23), 7914 (2021)
8. Hsu, C.C., Kung, C.H., Jheng, J.J., Chang, C.Y.: Unsupervised distance learning for extended self-organizing map and visualization of mixed-type data. Intelligent Data Analysis **23**(4), 799–823 (2019)
9. Ji, J., Pang, W., Zhou, C., Han, X., Wang, Z.: A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. Knowledge-Based Systems **30**, 129–135 (2012)
10. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems **30** (2017)
11. Levy, L.N., Bosom, J., Guérard, G., Amor, S.B., Bui, M., Tran, H.: Application of pretopological hierarchical clustering for buildings portfolio. In: SMARTGREENS. pp. 228–235 (2021)
12. Lévy, L.N., Bosom, J., Guerard, G., Amor, S.B., Tran, H.: Modeling and recommendation system for improving the energy performance of buildings. In: International Symposium on Distributed Computing and Artificial Intelligence. pp. 206–209. Springer (2021)
13. Lucas, B., Shifaz, A., Pelletier, C., O'Neill, L., Zaidi, N., Goethals, B., Petitjean, F., Webb, G.I.: Proximity forest: an effective and scalable distance-based classifier for time series. Data Mining and Knowledge Discovery **33**(3), 607–635 (2019)
14. Masuyama, N., Amako, N., Yamada, Y., Nojima, Y., Ishibuchi, H.: Adaptive resonance theory-based topological clustering with a divisive hierarchical structure capable of continual learning. arXiv preprint arXiv:2201.10713 (2022)
15. Middlehurst, M., Vickers, W., Bagnall, A.: Scalable dictionary classifiers for time series classification. In: International Conference on Intelligent Data Engineering and Automated Learning. pp. 11–19. Springer (2019)
16. Oehmcke, S., Zielinski, O., Kramer, O.: knn ensembles with penalized dtw for multivariate time series imputation. In: 2016 International Joint Conference on Neural Networks (IJCNN). pp. 2774–2781. IEEE (2016)