Buildings Power Consumption: Two-years-Ahead Hourly Forecasting Model

Soufiane Chami, Houda Bouderraoui
University of North Dakota, USA
E-mail: {soufiane.chami, houda.bouderraoui}@und.edu

***Abstract*—The considerable growth of energy demand throughout the years has hastened the migration towards the modern power systems named Smart Grids. To effectively meet the energy demand and well manage its energy flow, forecasting has become a key element for operators and buildings' owners to monitor their energy flows. Predicting the energy demand patterns an hour, a week or a year ahead helps in improving the reliability and efficiency of the power system along with reducing the load consumption. This paper presents three forecasting models based on long-term load prediction for various buildings: Educational, Offices, Residential, etc. Based on one-year training data, we were able to predict the next two-year energy demand of 1500 buildings using three different forecasting models: Light-GBM, Artificial Neural Network, and Linear Regression. We compared the accuracy of each model using Root Mean Square Logarithmic Error. The experimental results show that Light-GBM method performs well since it provided a better RMSLE of 1.09 compared to the other methods.**

***Index Terms*—Forecasting, Energy consumption, Light-GBM, Neural Network, Linear Regression, Load prediction.**

I. Introduction

A. Motivation

THE traditional power systems have known several changes over the last decades. The high penetration of Information and Communication Technologies (ICT) and Distributed Energy Resources (DER), as well as the high energy demand records over time, have fastened the immigration toward a more intelligent power system called Smart Grid. Smart Grid becomes a key enabling for new technologies integration, it is envisioned as the next-generation power grid, promoting the integration of new concepts, and new power resources.

B. State of the Art

One of the main objectives of Smart Grid is monitoring, controlling and reducing energy consumption in different sectors: residential, industrial and commercial which enhanced the use of forecasting techniques in the energy sector especially with the large implementation of smart meters. On the other hand, forecasting techniques as a promising tool have demonstrated a great and significant plus to the scientific community. Predicting the energy consumption rates, for instance, of the following days, weeks or even years will not only help decreasing and shaving the peak demand but ensuring the reliability and efficiency of the power grid. It is in this context that even though it is difficult, energy forecasting becomes a hot topic that grabs increasing attention from academia, industry, and government.
Typically, there are three kinds of forecasting: Short term which stands for one hour to one-week predictions, Medium-term forecasts that range from one week up to one year and last but not least, Long term predictions which intervals are longer than one year [4]. Load forecasting helps in reshaping the energy use over time, as a result, opt for suitable and favorable energy purchase plans [4]. According to [1], [2], [3], [15] residential and commercial buildings consume 20-40% of total energy consumption worldwide and it keeps increasing over time. Building's size, age, location, architecture and the thermal properties of the used materials, number of occupants and their behavior and finally weather conditions [10][6][7][11][8] are all considered as important factors in establishing its energy consumption. As stated by [5], there are two major methods for performing load forecasting: 1) Physics principles-based models and 2) Statistical and machine learning-based models. In this paper, the authors' focus was on machine learning-based models.

In the literature, several papers have discussed buildings' energy demand using different machine learning approaches. In [9], Daut and al. have provided an overview of conventional, artificial intelligence (AI) and hybrid methods and analyzed the performance of these methods for load forecasting. In [5], Marino and al. investigate the effectiveness of two Long Short-Term Memory (LSTM) variations LSTM based Sequence to Sequence (S2S) architecture and standard LSTM to forecast a residential building's energy consumption based on one-hour and one-minute resolution dataset. Authors claim that the standard LSTM does not perform well as the LSTM base S2S which provides better results on both datasets similarly to Factored Conditional Restricted Boltzmann Machines (FCRBM) and the conditional restricted Boltzmann machine (CRBM). As for [12], it provides a performance comparison between the Convolutional Neural Network (CNN) and Artificial Neural Network (ANN), Support Vector Machine (SVM), LSTM, LSTM-S2S and FCRBM algorithms on a One-hour time resolution dataset of a single residential customer. In a similar study, Amber et al. [14] have compared five forecasting techniques: ANN, Deep Neural Network (DNN) Multiple Regression (MR), SVM and Genetic Programming (GP), on five years electricity consumption dataset of an administrative building in London. This comparison presented good MAPE score of 6% for

ANN method. Yu et al.[13] model individual household electricity demand using sparse coding approach to predict next-day and next-week load. Authors affirm the effectiveness of sparse code features in reducing the forecasting error of next-day and next-week total load. Mocanu and al. [4], compared CRBM, FCRBM, ANN, SVM and the Recurrent Neural Network (RNN) performance in predicting residential energy consumption over different time resolutions. Results show that FCRBM outperformed the four other methods. In [6], Weicong and al. demonstrate the performance of LSTM method on residential load forecasting through comparison with Feed-Forward Neural Network (FFNN) and K-Nearest Neighbor (KNN). Aydinalp and al. [10] have used neural networks to model and predict space heating (SH) energy consumption as well as Domestic hot-water (DHW) heating energy consumption.

In order to evaluate the accuracy of forecasting models, different error metrics can be used such as Root Mean Square Error (RMSE), Mean Square Error (MAS), Mean Absolute Percentage Error MAPE, Relative Error (RE), Mean percentage error (MPE) and Root Mean Squared Logarithmic Error (RMSLE) [14]. In this paper, the authors will use three forecasting models to predict the energy demand of 1500 buildings. These forecasting models are Light-GBM, Neural Network (NN) and Linear Regression. Models' results will be compared to each other using RMSLE.

The remainder of the paper is organized as follows. Section II presents an overview of the problem, dataset description and processing are provided in Section III and Section IV respectively. In Section V, a background on the Light-GBM method has been provided. Section VI summarizes the results analysis and discussion Different forecasting models used in this paper are described in Section IV. Finally, Section V concludes the paper and provides the authors' future works.

II. Problem Statement

As stated in Kaggle platform [16], the purpose of this challenge is to predict energy consumption per building per hour for the two years ahead given the historic data of only one year. The historic data contains approximately 15K buildings existing in anonymous locations around the globe, starting from the 1st January 2016 to the 31st December 2016. The goal is to predict meter reading in the next two years from the 1st January 2017 until the 31st December 2018. The Model predictions will be evaluated with Root Mean Squared Logarithmic Error (RMSLE). In this paper, we will examine different approaches and multiple algorithms as well as comment on the different features and strategies that provide most utilities to predict hourly energy consumption per building.

III. Data Description

The layout of the challenge is to give a year range of historic data (2016) to predict the following two years (2017-2018). The provided dataset includes hourly based historic data of more than 1500 buildings all over the globe and across different meter types. The data is not based only on historic energy consumption, but also the correspondent weather details for each hour. A detailed outline of the data is listed as follow:

A. Building metadata

This table contains the available details about each building such as the year the building was built, the primary use of the building, and the flour count:

1) **site_id**: unique side identifier;
2) **building_id**: unique building identifier;
3) **primary_use**: the type of activities meant for the building: Education, Office, ...etc.
4) **square_feet** : Gross floor area of the building;
5) **year_built** : Year building was opened;
6) **floor_count** : Number of floors of the building.

B. Meteorological data

The other source of data provides historic weather conditions of each site. The features of the weather conditions are listed as follows:

1) **site_id**: unique side identifier;
2) **air_temperature** air temperature in Degrees Celsius;
3) **cloud_coverage**: portion of the sky covered in clouds, in oktas;
4) **dew_temperature**: dwelling temperature in Degrees Celsius;
5) **precip_depth_1_hr** : precipitation depth in millimeters;
6) **sea_level_pressure** : sea level pressure in millimeters;
7) **wind_direction** : compass direction (0-360);
8) **wind_speed** : wind speed in meters per second.

Authors assume that all the buildings of each site have the same weather conditions.

C. Power Consumption data

The final data source provided in the challenge is the time series record of the energy consumption of each building. The dataset has been divided into training and testing sets. The training data is considered from the $1^{st} January, 2016$ to $31^{st} December, 2016$, while the testing data is the historic data from $1^{st} January, 2017$ to $31^{st} December, 2018$. Authors would like to mention that the training and the testing datasets were both depicted from the same dataset of the same building. The features included in these tables are listed as follow:

1) **Timestamp**: the moment the measurement was taken;
2) **Building_id** unique building identifier
3) **Meter**: describes the $metertype$ and labeled as follows $\{0 : electricity, 1 : chilledwater, 2 : steam, 3 : hotwater\}$.
4) **Meter_reading**: the target variable: Energy consumption in kWh (or equivalent). Note that this is real data with measurement error, which we expect will impose a baseline level of modeling error. For the test data, the target variable is filled with zero (hidden).

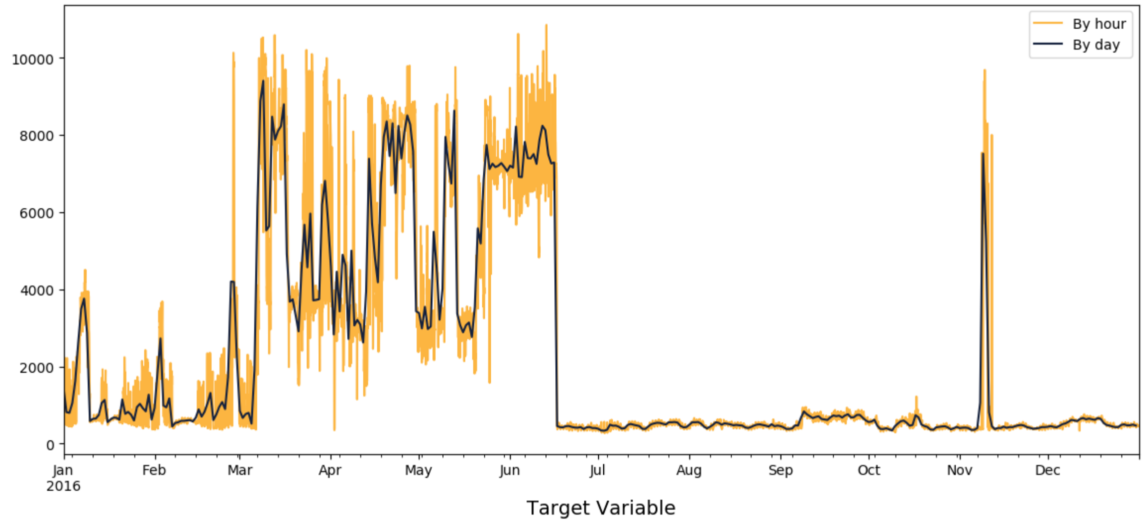

Fig. 1. Target Variable : Meter reading measurements aggregated by day during one year 2016

D. Data Exploration

As illustrated in Fig.1, the average energy consumption during the first three months of 2016 has increased during the next four months. Then it decreases for the rest of the year except for a peak in November. A preliminary hypothesis that can explain this phenomenon is that buildings included in the training may have a dominant pattern that drives the whole consumption profile in the training dataset.

This hypothesis can be validated to a large extent using the figure 2 . We observe that the education buildings have almost the same profile as the global meter reading in the training data. Besides, the magnitude of the education is extremely dominant over all the other buildings with different primary use. In other sites, where the energy consumption of Education buildings takes different shapes, the level of magnitude of consumed energy is very low and can not drive the global energy profile shown in figure 1.

IV. Data Processing

To make it possible to feed the machine model and ensure it can capture the consistent and entire information, we need to perform features engineering on the training data and any other preprocessing methods if necessary. In the following sections, we explain the main features engineering operations applied during the modeling pipeline. Then we will evaluate the impact of each technique on the model performance.

*1) **Outliers Removal**:* As deduced earlier from previous sections, the energy consumption profile is too different from a building type to another. Therefore, outliers should be carefully and separately treated since they usually create discrepancies between the training and test data. Outliers are usually known by their characteristic of ruining the validation schema in machine learning if they are not well treated.

Most of what we have observed in the training data confirmed that outliers come mainly from Education buildings **??**. Practically speaking, in order to detect outliers, we aggregated the train data by day and $meter\_type$. We believe the outliers should be observed exclusively with respect to the type of energy consumed. Thus, we plotted the $meter\_reading$ of the whole year with daily energy consumption. As can be seen in Fig.**??**, there is a quite flagrant difference in meter_reading for each meter type. Table I shows the number of outliers building detected within each meter type.

It is important to check some of these buildings' outliers and find the reason for their high detected values. For instance, and as illustrated in Fig.**??**, building 799 seems it did not have any data before June. This can be explained by the fact that the building is new, but according to the dataset, it was built in 1976. Besides, two periods with a daily average of 75 000 kWh then 175 000 kWh for 500 000 ft^2 are observed. Which is equivalent to 0,15-0,35 $kWh/ft^2/day$ and 55-130 $kWh/ft^2/year$.Th is is 3-7 times more than the typical consumption that is 20 $kWh/ft^2/year$.

The most probable explanation is something wrong with the software configuration that did not make the measure conversion correctly. We observe a similar problem in other outliers. The solution is $normalization$ that converts the meter reading to a value between 0 and 1.

*2) **Timestamp Parsing**:* This feature has valuable information that would help in mapping the energy consumption and capture seasonality and time decencies. For example, the power consumption in some buildings may be higher or lower at the weekends or holidays. It can also change in the summer and winter. For this reason, we can apply a parsing function to extract all this information from the timestamp. It is also worth noting that we can not use the holiday calendar since we do not have indications about the locations of the building. Finally, we could extract the three main features from the timestamp: day of the week, if it is weekend, week of the month and the hour of the day. Some of these features can be seen more as categorical features such as the day of the week.

*3) **Categorical Features**:* Handling the categorical feature impacts the performance of the model. It is very important to make sure that the model can handle them without losing the

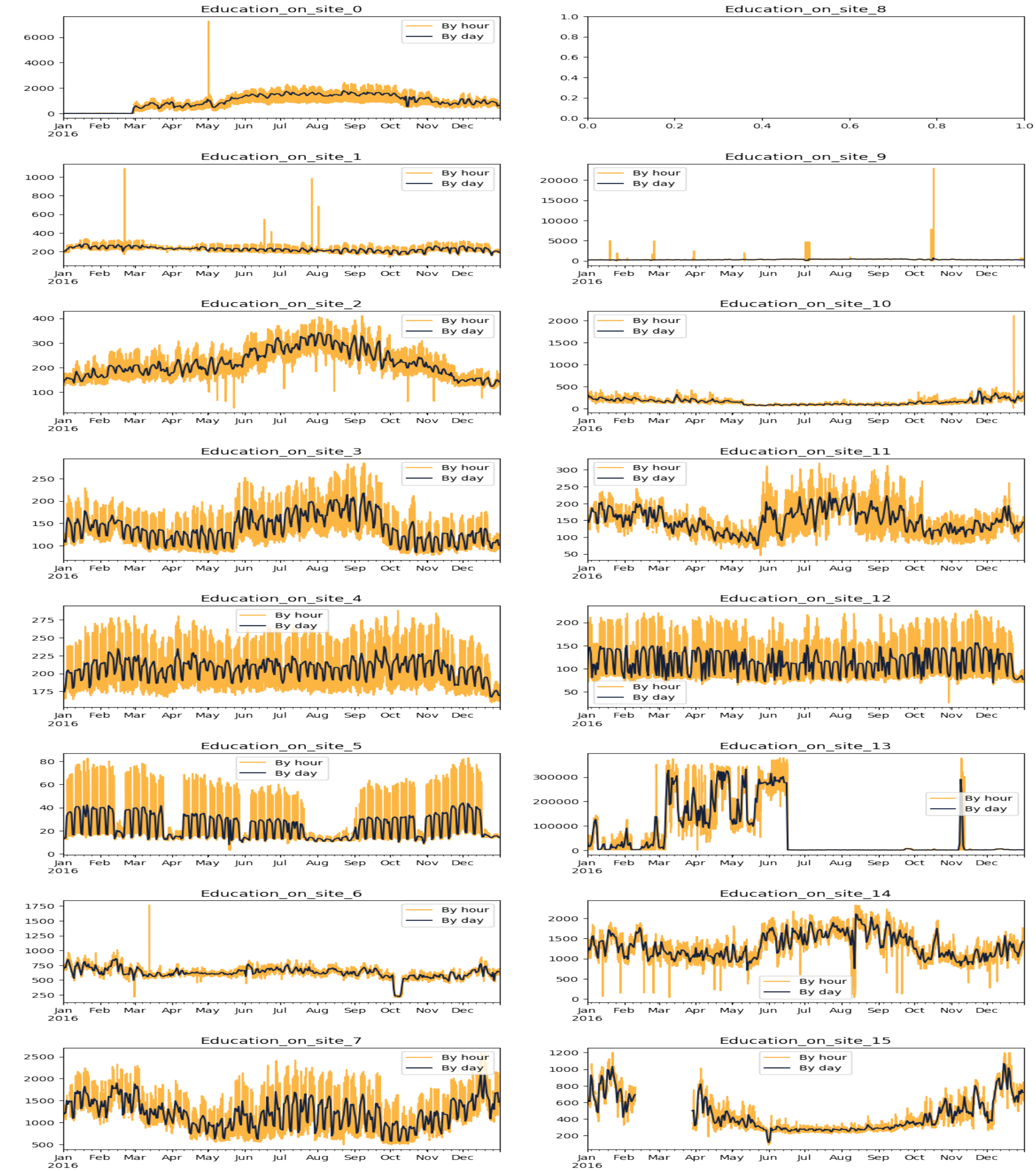

Fig. 2. Education Energy Profile per site

information they contain as much as possible. For this reason, we decide to highly consider Gradient boosting Machine Models such as Light-GBM. It is a very excellent model that can categorize features without the need for encoding and converting them into numerical columns. In table II, we enabled the options of categorical features recognition by Light-GBM, we notice that using this property helps considerably in improving the accuracy of the model by around 10%.

4) ***Missing values***: The weather table contains a lot of missing values as well as other tables' columns, such as

TABLE I
OUTLIERS BUILDING FOR ELECTRICITY METER

	site_id	building_id	primary_use	square_feet	year_build	floor_count
794	7	794	Education	731945	1969	11
799	7	799	Education	527431	1976	26
801	7	801	Education	484376	1952	5
803	7	803	Education	182986	1962	3
993	9	993	Education	428647	NAN	NAN
1088	13	1088	Education	287594	NAN	NAN

building metadata. For the weather data, we will use interpolation as a method to compute the missing values. Whereas, for the categorical features in building metadata, we will use NAN values as a separate category along with other categories within the columns.

*5) **Target Preprocessing**:* Since the evaluation metric of this challenge relies on the Root Mean Squared Logarithmic Error (RMSLE), we will consider converting the target variable to log to rescale the target variable and applies reverse functions to return to the kWh scale.

V. Methods

A. Light-GBM

Light-GBM [17] is a gradient boosting machine algorithm that is similar to XGBoost with few variations. Generally speaking, gradient boosting machines algorithms are based on a fundamental Idea: Weak learners make strong predictors. This training process of GBM algorithms starts by training a weak learner such as decision trees and observes its error and mispredicted data points. Then, special weights are attributed to each of these data points. The following learner will be trained and tries to avoid making the same mistakes as the previous model. This cycle will be repeated and iterated until a reasonable performance is achieved.

This principle is applied to both XGBoost and Light-GBM, the only difference between these algorithms is that XGBoost follows vertical decision trees growth (depth-wise) and Light-GBM follow a horizontal growth (leaf-wise)

In this model, our purpose is to observe the model behavior toward the training data at hand. We also wanted to set up a benchmark for our project to evaluate the utility and impact of each adjustment made on the model. This point is very important to avoid modeling complexities that do not provide any added values to the objective of this project. At this stage, we chose Light-GBM as our selected algorithm to run the experiment. Later, it will be compared with different algorithms.

VI. Results and Discussion

In this section, we will present the evaluation metrics and the corresponding results obtained for each model. We will also outline the most important factors that contributed to the prediction of the energy consumption per building.

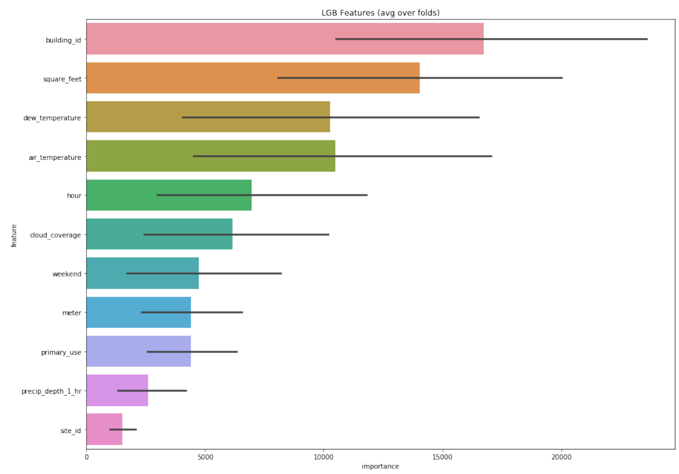

Fig. 3. Features classification according to their importance and contribution in predicting buildings' energy consumption

A. Features Engineering Strategy Evaluation

We first run a baseline model having all the preprocessing strategies enabled. Then we disable one strategy each time to assess the model. Next, we compare the performance variation and report a decreasing percentage. Thereby, we can observe the impact of each feature engineering strategy and quantify its utility. Table II summarizes the obtained results.

B. Important factors

Using Light-GBM, we got interesting results about the most important features that helped in predicting the future consumption profile of each building. Form Fig.3, it can be concluded that the weather conditions, building metadata, and day-time are the most important aspects in predicting the energy consumption of buildings.

C. Evaluation and Metrics

On the performance evaluation, the challenge adopted The Root Mean Squared Logarithmic Error (RMSLE) metric to judge models' accuracy. The Root Mean squared Error (RMSE) measures the bias and variance. It is defined in (1) as follows:

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(p_i - a_i\Big)^2} \qquad (1)$$

TABLE II
SUMMARY OF OBTAINED RESULTS

Strategy	Contribution	Details
Log-scaled target	45.70%	Rescale the meter readings using log1p
Dropping Outliers	17.80%	Drop bad measurements form site 0 electrical data from the first 141 days
Allow LGBM to handle Categorical Features	9.70%	Define Categorical Features and allow LGBM to handle them itself
Per-meter models	3.90%	Build separate models per meter type
Time-based features	4.00%	Timestamp parsing, and fix difference across different sites
Filling missing data	0.40%	Input missing values in weather data by interpolation
Missing value as Category	0.10%	Use missing as categorical features to mentioned if the columns has Nan value or not

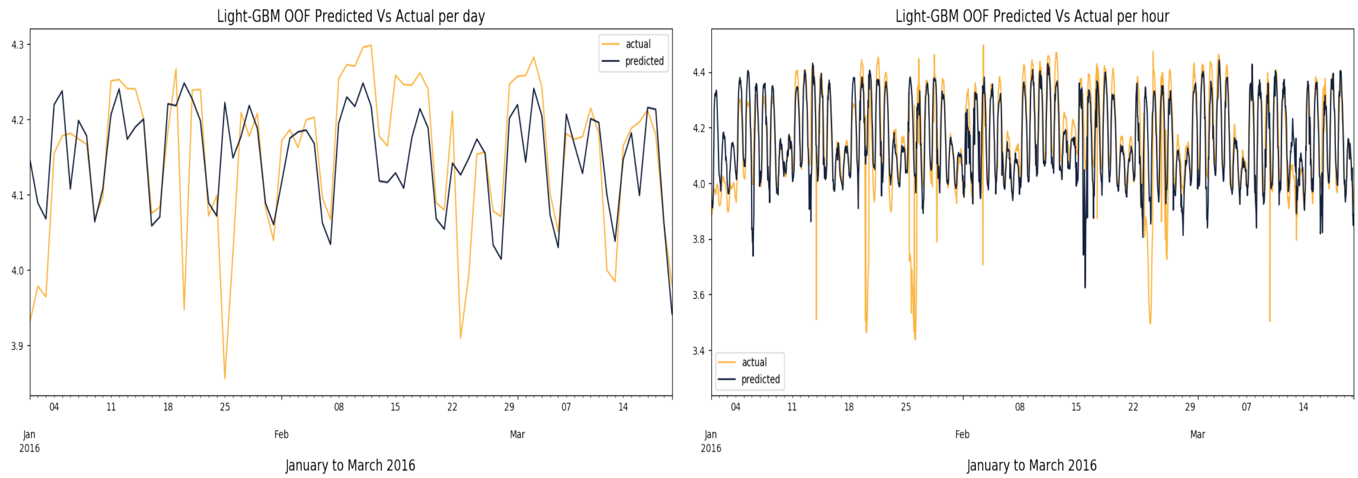

Fig. 4. Light-GBM Out-Of-Folds (OOF) predictions Vs Actual values per day and per hour

While RMSLE is used when the predictions and actual values are replaced by their logarithmic values as described by (2):

$$RMSLE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(\log(p_i) - \log(a_i)\Big)^2} \quad (2)$$

There is a reason behind choosing RMSLE over the RMSE and MAE or any other regression loss function. The RMSLE is used when we do not want to penalize the model for huge differences between predicted and actual values when the response variable takes a big number. More specifically, RMSLE does not penalize the model when it makes an over-estimation, but it punishes the model for making under-estimations. It is more robust against outliers.

The obtained results from the three tested models are as follows:

TABLE III
TRAINING RESULTS

Approach	Train	Validation
Light-GBM	1.1017	1.099
ANN	1.061	1.15
Linear Regression	1.7	1.909

Generally speaking, Light- GBM model is more performant than the other algorithms. On one hand, despite their great success on complex data structures like text and images, neural networks are very limited on tabular data due to their data hanging and the fact that they require to be fed by a huge amount of data. Besides, in the context of this challenge, it is in-efficient to treat each building as a sequence to make it meaningful for an ANN. On the other hand, Light-GBM is very successful in this type of data and can learn more than different extracted features.

Table III summarizes the obtained results of the three tested algorithms in terms of RMSLE. As can be seen, Light-GBM outperforms all other methods. It achieves 1.09 on tests dataset.

Finally, for the sake of providing a visual description of the behavior of the Light-GBM model, we included Fig.4 that provides an snapshot of the Out-Of-Folds predictions throughout four months from January to March 2016.

VII. CONCLUSION

In this paper authors have compared the performance of three forecasting models: Light-GBM, ANN and Linear regression to predict two years energy demand of 1500 buildings over the world based on one year data. From the research that has been carried out, Light-GBM model outperforms the other proposed models since it provided a low RMSLE 1.09 compared to the other methods.

REFERENCES

[1] Pérez-Lombard, L., Ortiz, J., & Pout, C. (2008). "A review on buildings energy consumption information". Energy and buildings, 40(3), 394-398.

[2] "The U.S. Energy Information Administration (EIA)", online, accessed 11/25/19, https://www.eia.gov/

[3] Costa, A., Keane, M. M., Torrens, J. I., & Corry, E. (2013). "Building operation and energy performance: Monitoring, analysis and optimisation toolkit". Applied Energy, 101, 310-316..

[4] Mocanu, E., Nguyen, P. H., Gibescu, M., & Kling, W. L. (2016)."Deep learning for estimating building energy consumption". Sustainable Energy, Grids and Networks, 6, 91-99.

[5] Marino, D. L., Amarasinghe, K., & Manic, M. (2016, October). "Building energy load forecasting using deep neural networks". In IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society (pp. 7046-7051). IEEE.

[6] Kong, W., Dong, Z. Y., Hill, D. J., Luo, F., & Xu, Y. (2017). "Short-term residential load forecasting based on resident behaviour learning". IEEE Transactions on Power Systems, 33(1), 1087-1088.

[7] Yu, Z., Fung, B. C., Haghighat, F., Yoshino, H., & Morofsky, E. (2011). "A systematic procedure to study the influence of occupant behavior on building energy consumption". Energy and buildings, 43(6), 1409-1417.

[8] Zhao, H. X., & Magoulès, F. (2012). "A review on the prediction of building energy consumption". Renewable and Sustainable Energy Reviews, 16(6), 3586-3592.

[9] Daut, M. A. M., Hassan, M. Y., Abdullah, H., Rahman, H. A., Abdullah, M. P., & Hussin, F. (2017). "Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review". Renewable and Sustainable Energy Reviews, 70, 1108-1118.

[10] Aydinalp, M., Ugursal, V. I., & Fung, A. S. (2004). "Modeling of the space and domestic hot-water heating energy-consumption in the residential sector using neural networks". Applied Energy, 79(2), 159-178.

[11] Aksoezen, M., Daniel, M., Hassler, U., & Kohler, N. (2015). "Building age as an indicator for energy consumption". Energy and Buildings, 87, 74-86.

[12] Amarasinghe, K., Marino, D. L., & Manic, M. (2017, June). "Deep neural networks for energy load forecasting". In 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE) (pp. 1483-1488). IEEE.

[13] Yu, C. N., Mirowski, P.,& Ho, T. K. (2016). "A sparse coding approach to household electricity demand forecasting in smart grids". IEEE Transactions on Smart Grid, 8(2), 738-748.

[14] Amber, K. P., Ahmad, R., Aslam, M. W., Kousar, A., Usman, M., & Khan, M. S. (2018). "Intelligent techniques for forecasting electricity consumption of buildings". Energy, 157, 886-893.

[15] Cao, X., Dai, X., & Liu, J. (2016). "Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade". Energy and buildings, 128, 198-213.

[16] "ASHRAE - Great Energy Predictor III", Kaggle Competitions Platform, 2019, https://www.kaggle.com/c/ashrae-energy-prediction, online, accessed 2019-12-15.

[17] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). "Lightgbm: A highly efficient gradient boosting decision tree". In Advances in Neural Information Processing Systems (pp. 3146-3154).