Machine Learning applied to energy demand forecast in IST Alameda campus

Francisco Delca Gouveia Pereira
franciscodelca@tecnico.ulisboa.com

Instituto Superior Técnico, Universidade de Lisboa, Portugal

November 2019

ABSTRACT

Energy consumption forecasting of buildings plays a crucial role in making planning decisions by facility managers and energy providers. These decisions are used to reduce the intrinsic environmental impact of the building sector. Nowadays, with the imminent application of Building Energy Management Systems (BEMS) and the consequent increase of generated data, the use of machine learning algorithms to provide such predictions becomes a natural solution. In this study, four machine learning algorithms (MLP, SVM, RF, and XGB) were compared in three different forecasting horizons (an hour, a day, and a week) for four buildings (Civil, Central, North tower, and South tower) located at Instituto Superior Técnico, Lisbon, (4 algorithms x 3 forecasting horizon x 4 buildings = 48 models). In the development of such models, three years of hourly gathered data of each building consumption and outdoor weather conditions were used. Firstly, due to the missing values presented in the data, an imputation study was carried out in order to guarantee data temporal continuity. Afterwards, based on the energy consumption analysis of each building, different features were created in attempt to describe buildings' behaviour. From the created features, different data sets were developed per building and forecast horizon, where a feature selection analysis supported with the use of a wrapper method, known as RFE, took place. With that selection, it was concluded that the most important features were the type of day, the lagged features, and the average cluster consumption of a typical working day. At last, an hyperparameter search using Bayesian optimization was conducted and the models were then used to forecast the last year of data. Among all the models used, SVM models outstood, showing higher accuracies in most of the forecasting horizons and buildings. Overall, in 93% of the forecasted days, it was achieved a MAPE error of 10.95%, 9.17%, 10.48%, and 12.66% for Civil, Central, North tower, and South tower buildings in a week horizon forecasting, respectively. In addition, it was also noticeable an increasing annual error tendency when the models attempt to predict in greater horizons.

Keywords: energy, building consumption, forecast, machine learning

1. Introduction

Nowadays, the buildings sector accounts with almost 40% of energy consumption and 36% of CO_2 emissions [1]. The forms of energy that are predominant in this sector are electricity and natural gas, accounting with 10 to 15% of overall energy consumption [2] and growing by an average of 1.8% per year from 2010 to 2040 [3]. It is imperative to state the need to reduce building excessive energy consumption, since they represent a significant fraction of the overall energy expenditure, which consequently results in high environmental impacts. Since 2006 [4], the European Commission has been implementing energy efficiency measures for sustainable development, being one of the main objectives to reduce the annual energy consumption by 27% till 2030.

One way to achieve that goal in the building sector, is to effectively predict its consumption, enabling the endorsement of diverse operating strategies to increase energy efficiency and to detect faults related to systems malfunction. To address this need over the past 50 years [5], a large number of investigations have been carried out to ascertain the complexity related with buildings energy consumption and to find out an accurate representation of its energy performance. Currently, building energy simulation can be branched into three different approaches: **white box**, **grey box** and **black box**.

White box approaches, also known as physical models, are widely used in engineering and are grounded by thermodynamic laws, requiring many building details and surrounding environmental conditions as input data. Computationally they are very expensive, and data input requirements may, in some circumstances, not be entirely fulfilled. **Grey box** approaches merge the models mentioned above with statistical modelling, allowing the use of simplified building information and historical data to perform the energy simulation. Nevertheless, they provide reasonable accuracy predictions with high computational cost depending on building information. In order to circumvent the shortcomings referred by the first two approaches, **black box** approach was employed. This purely data-driven approach, when compared with the others, is able to develop a faster and higher accurate consumption forecasting, based only on historical data, avoiding thus the need of physical building details [6]. For those reasons the models that characterize this approach, mainly in machine learning field, have been receiving particular attention in the past years.

2. State of the Art

As it is known, data-driven models, instead of using detailed building information to develop an energy analysis, use only historical and available data to learn the dynamic energy behaviour of the buildings and are often referred to as empirical models. Nowadays, due to their ability to extract useful information at low cost, they have been applied in diverse fields such as commerce [7], political campaigns [8], and medical diagnosis [9].

The most common data-driven models used for energy consumption forecasting may be ramified into two fields: the statistical field and the machine learning field. From the **statistical field**, the models often applied were the autoregressive, integrated and moving average (ARIMA) and the multiple linear regression (MLR). On the other hand, from the **machine learning field**, two models were substantially applied, specifically, artificial neural networks (ANN) and support vector machines (SVM), and another one least used named as ensemble model.

The **statistical models** that have been frequently used to predict building energy consumption are generally regression models [10]. Statistical regression techniques find relationships between the different variables through mathematical formulations to predict a specific target. Several investigations took advantage of this approach to address diverse challenges in the analysis of building energy behaviour, for example, to predict energy used through simplified variables, foresee building energy index, and estimate significant energy parameters for analysis [5]. From regression models, there are at least two models that are mandatory to emphasize, the MLR and the ARIMA. The latter was specifically created to handle and correlate time series data for prediction. Examples of its applicability in short term building energy prediction may be found in [11]-[14]. Although these models are easy to develop and use, they lack on flexibility in coping with the nonlinearity often found in building energy consumption. In consequence of that, the statistical approach presents poorer prediction accuracy, which limits its applicability, when compared with machine learning models, examples of that may be seen in [15]-[17].

Machine learning is an interdisciplinary field based on statistics and optimized mathematics techniques which gives computer systems the ability to learn and improve performance on a given task, being only fed with data without the need to be explicitly programmed [18]. Within machine learning models, ANNs have been particularly popular and applied to forecast buildings energy consumption [10]. In 2005, *Gonzales* and *Zamarreno* [19], used a feedback ANN model developed in [20] to predict next hour consumption of an institutional building. In 2009, *Yokoyama* [21] used an optimization method known as modal trimming [22], instead of the typical gradient descendent in ANN backpropagation stage, to predict the cooling load of a service building. In 2016, *Chae et al.* [23] when forecasting a day a-head energy consumption of a service building, tested the applicability of an ANN with Bayesian optimization to improve model generalization. To input relevant data to the model it was used a feature extraction technique by means of an ensemble machine learning algorithm, known as random forest (RF). The study revealed a decreased in the forecast error as the number of weeks of data available for training increased. Right after ANN, the SVM model was the second used algorithm to predict energy consumption, and although the widely used of ANN, in some of the cases SVM model outperformed it. An example of that was done in 2009 by *Li et al*. [24], which conducted a study comparing the traditional ANN with an SVM model, to predict one hour a-head cooling load of an office building, in China. The SVM model used radial basis function kernel and revealed higher accuracy than the traditional ANN. In the same year, the same type of conclusion was achieved when performed in another office building by *Xuemei et al*. [25].

In addition, new and advanced techniques were also used in this field. Known as ensemble model they were introduced in 1990 [26]. This kind of models uses multiple learning algorithms to obtain a better accuracy performance than that could be obtained from any of the constituent learning algorithms [27]. When the models used are the same it is named as **Homogeneous** ensemble modelling, on contrary when they use different models to predict it is named as **Heterogeneous** ensemble modelling. Both groups have been used to forecast the building energy consumption, some examples of the ones applied for short term prediction are mentioned below. *Fan et al.* [28], used a homogeneous model to predict half-hourly a-head energy consumption of an institutional building, in Singapore. The forecasting ensemble model used was a weighted SVM model with nu-SVM and epsilon-SVM. On the other way, *Xiao et al.* [29], used a heterogenous ensemble model by combining eight different predictive models, to forecast a day a-head building energy consumption. Furthermore, in the **homogeneous universe**, the ensemble models may use two types of learning procedure which characterizes the order that each model is trained, namely, bagging and boosting. In bagging each model learns with a random subset of training data in a parallel way, e.g. random forest (RF). On the other hand, in boosting each model learns from mistakes made by the previous model in a sequential way, e.g. extreme gradient boosting (XGB). Both random forest and extreme gradient boosting algorithms have been recently used in building energy consumption prediction [30]-[32]. One of the reasons that encourage the application of them is the previously used of their based model, decision trees (DT) algorithm, in the field [10]. A comparison between each type of most used model may be seen in Table 2.1.

In this study the two single models (SVM and MLP) and two ensemble models (XGB and RF) were compared on their forecasting potentialities.

Table 2.1 - Brief comparison of machine learning models used in building energy consumption [33]

Models	ANN	SVM	Ensemble
Advantages	Solve complex nonlinear problems; In general, better performance prediction than SVR	Good balance between prediction accuracy and calculation speed; Few parameters need to be determined	Best prediction accuracy and stability
Disadvantages	Many parameters need to be determined	Kernel function is crucial and difficult to be determined	Difficult to implement
Computational Speed	Medium speed	Medium Speed	Low speed
Accuracy	Good	Average	Best

3. Study case

3.1. Building introduction and energy analysis

In the scope of this study, four different buildings were analyzed, specifically, Civil, Central, North tower, and South tower buildings, located at Alameda campus of Instituto Superior Técnico (IST), Lisbon.

To distinguish the different consumptions patterns, each building was evaluated in monthly, weekly and daily temporal granularities with the data acquired from the previous year of the forecast, 2017. To conduct this analysis, two different tools were used: boxplots and an unsupervised machine learning algorithm, known as k-means.

Monthly Analysis

It was shown from the last energy audits [34]**,** that the main forms of energy use to fulfil their needs are natural gas and electricity. Electricity covers all the demands as regards to lighting, computers, plug-in devices, catering, common systems and laboratories facilities. In addition, natural gas is mainly used for spaces under concession and certain laboratories facilities. Besides that, HVAC system works differently in two groups of buildings: **Central** and **Civil** buildings, operate only with electricity in cooling and heating season; On the other hand, **North** and **South** tower buildings, use electricity and natural gas for heating and cooling seasons, respectively.

The latter statement may be seen in the monthly analysis of Figure 3.1. In both **towers** the consumption fluctuations along the year are characteristic of the different types of energy form used. As regards to **Civil** and **Central** buildings, the stable consumption curve along the year is substantiated by the main use of electricity.

Weekly Analysis

In each building there is a 7 days cycle pattern that repeats almost through the whole year, as it is represented in Figure 3.2. As expected, most of the buildings' energy use occurs during weekdays since it is when the majority of buildings' activities take place. During the weekend, there is an abrupt fall in energy consumption, although Saturday energy use is slightly higher than Sunday.

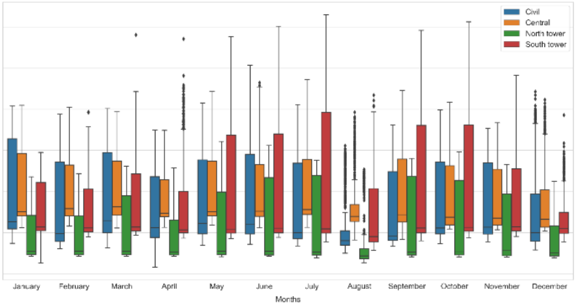

Figure 3.1 - Monthly boxplot energy consumption of each building.

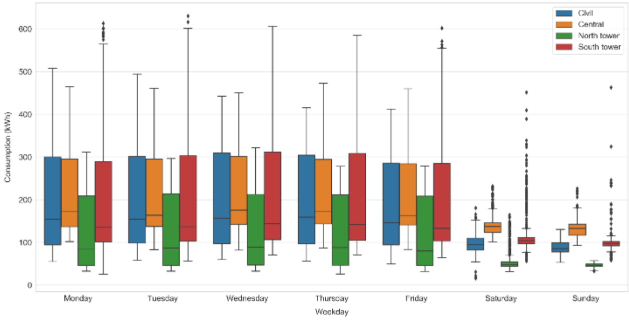

Figure 3.2 - Weekly boxplot energy consumption of each building

Daily Analysis

In hourly analysis, each building was evaluated individually to simplify the visualization and gathered each different daily consumption curve characteristics. In this paper, Civil building was used as role model, further information of the other buildings analysis is given in [35]. The application of k-means algorithm for the daily analysis is shown in Figure 3.3. To complement the analysis,

Table *3.1* shows each cluster type of day percentage.

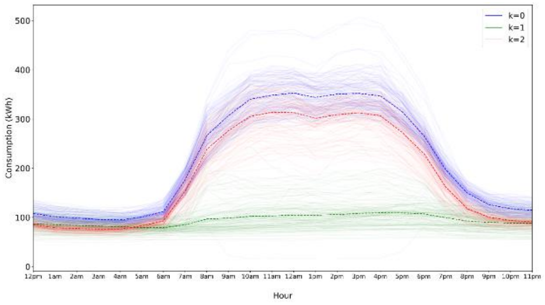

Figure 3.3 - Civil building daily consumption patterns defined by k-means algorithm (k=3)

Table 3.1 - Civil building day type percentage of each daily consumption patterns defined by k-means algorithm (k=3)

Day type (%)	k = 0	k = 1	k = 2
workday	100	3.1	98.1
holiday	0	11	1.9
weekend and summer break	0	85.9	0
heating season	55.8	47.7	43.9
cooling season	44.2	52.3	56.1

Based on Figure 3.3 and

Table 3.1, the following may be concluded:

- In general, the daily energy patterns are not influenced by heating and cooling seasons as referred earlier in monthly analysis, since every cluster has around 50% of the days in each season.

- k=0 and k=2 - consisting mainly of working days represent two typical mean workday consumption patterns. Between both clusters, k=0 consists only of working days, achieving a higher mean energy consumption pattern than k=1.

- k=1 - includes holidays, summer break days and weekends, being the latter predominant. The mean consumption is nearly stationary with a slightly higher expenditure during daytime mostly due to Saturday opening hours and the 24 hours studying area.

4. Methodology

After knowing the main characteristics of Civil building, the present chapter will demonstrate the methodology behind the development of this study main goal: the energy consumption prediction of each building in three different forecasting horizons (1 hour, 1 day, and 1 week). Which may be divided in the following steps:

1. Data Treatment
2. Feature Generation
3. Models Selection
4. Feature Selection
5. Hyperparameter Optimization
6. Forecasting

To conduct each step, a programming language, known as python was used [38].

4.1. Data Treatment

The data set available in this study may be grouped into two different categories. The first category, named as the energy consumption data (ECD), contains values of each building energy consumption, collected at 1-hour intervals for 3 years (2014, 2017, and 2018). The second category, denoted as weather conditions data (WCD), includes the outdoor weather conditions, such as temperature, relative humidity, and solar radiation. This category was also hourly gathered during the same years by an existent weather station in Alameda campus.

In real-world data sets where data is collected through sensors and electric meters, is often found incomplete, imprecise and noisy values. To proceed to the model development, it is mandatory to acknowledge and handle that uncertainty. For that, the following sequential steps were employed:

1. Frequency Preservation;
2. Outliers;
3. Data Imputation Study;
4. Creation of different data sets

Frequency Preservation

In the first step, since the data is time dependent, an hourly frequency preservation was done to keep the data set continuity. Therefore, every missing or repeated hour, were added or deleted, respectively, creating a missing value row if needed.

Outliers

In the second step, an outlier detection was performed in ECD. An outlier is an abnormal data value that considerably diverges from the rest of the data points in the same feature, their inclusion may affect negatively the predictive model accuracy. For that, a statistical metric, named as z-score was employed to detect the outliers, using $(|\sigma| \geq 4)$.

As a consequence of the first two steps, the number of missing values increased. To proceed to the following step, it is mandatory to quantify those values, Table 4.1.

Table 4.1- Data set of the available features and their total missing values and percentage

Data sets	Available Features	Missing Values		Consecutive missing values
		(%)	Total	
BCD	Civil	0.6	157	148
	Central	0.02	4	1
	North tower	0.32	83	76
	South tower	0.36	95	88
WCD	All features	8.97	2357	1165

Data Imputation Study

The third step focuses on the type of imputation applied to fill the missing values encounter in the data set, Table 4.1. In this work data set the main percentage of missing values was found in WCD. For this reason, merely adopting a strategy of dropping those values will lead to a loss of the correspondent ECD. Since the latter, contains the most valuable feature of each building, its own consumption, it is of extreme importance not losing it.

To overcome this situation, a data imputation study, Figure 4.1, has been conducted to measure the accuracy of different imputation strategies. Two models have been used in this study, specifically Multiple Imputation by Chained Equations (MICE) [36] and Miss Forest (MF) [37].

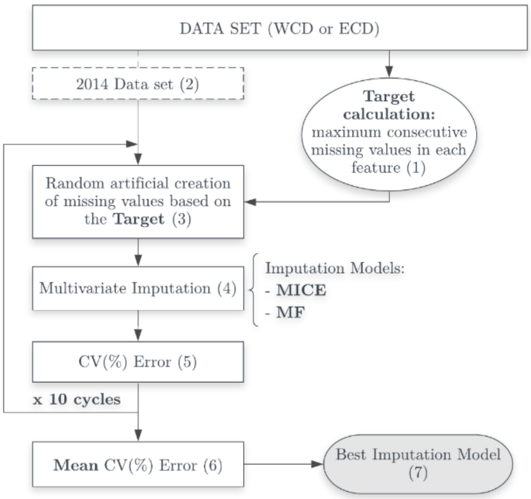
DATA SET (WCD or ECD)
2014 Data set (2)
Target calculation: maximum consecutive missing values in each feature (1)
Random artificial creation of missing values based on the Target (3)
Multivariate Imputation (4)
Imputation Models:
- MICE
- MF
CV(%) Error (5)
x 10 cycles
Mean CV(%) Error (6)
Best Imputation Model (7)

Figure 4.1 - Data imputation study step-by-step diagram

The **target** in Figure 4.1, is referred to the consecutive missing values column in Table 4.1.

After this study completion, although the MF model showed, in general, a great accuracy in each of the data set types when compared with MICE, the gap in WCD was still with undesirable imputation values in terms of the expected seasonality and trend. To address this problem, a method based in an univariate approach, referred in this paper, as hour monthly mean (HMM), was developed. This method basically fills each feature independently with the known values from the other years, using the correspondent mean value of the same month and hour. A comparison of the two models and the new method imputation was done for each of WCD feature. An example of it, for outdoor temperature (wt_temp), can be seen in Figure 4.2. It is possible to conclude that although none of the imputation strategies represents well the actual temperature, the new method gives the daily seasonality and the monthly tendency needed in comparison to the nearly constant values imputed by the MF and MICE.

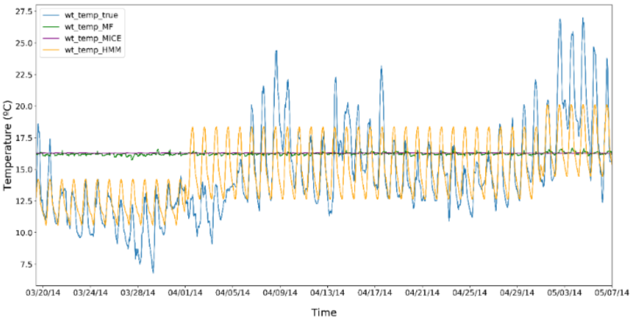

Figure 4.2 - Imputation comparison example of MF, MICE, and HMM imputations with the true value of the outdoor temperature (wt_temp).

Creation of different data sets

The fourth and last step, consists in the creation of three different data sets, that differ in the strategies adopted for the missing values shown in Table 4.1.

In this step, it was introduced another technique of univariate imputation, named as linear interpolation. The limit applied was of three hours, meaning that if any of the features has three consecutive missing values a linear interpolation is applied between the value before and after the gap.

Therefore, the first strategy common to each of the data sets, named as dt_01, dt_02, and dt_03, was the three hours linear interpolation. After that, the three data sets differ from one to another as it is described below:

- In dt_01 the remaining missing values were dropped, making it the least influenced by imputation;

- In dt_02 was applied a MF imputation just in the ECD. This was done to check the relevance of the imputation only in building energy consumption, since their error metrics shown the best mean results among the other features. Afterwards, the rest of the missing values were dropped.

In attempt to not losing any part of the data, the following data set was created:

- In dt_03 it was used the MF imputation model in ECD. After that, it was implemented the HMM method to the WCD, reasoned by the outcome shown in Figure 4.2.

In Table 4.2, it is possible to check the different techniques applied in each data set and the correspondent lost rows.

4.2 Features Generation

In this dissertation, the new features, may be grouped according to what they are based on, therefore, there are three main categories, time dependent, calendar, and energy consumption. The first includes every new feature that was time dependent, the second group of features was based on the national and academic calendar, and lastly, the third group generates features that were based on each building own consumption.

Table 4.2 - Summary of the imputation techniques applied per data set

Data sets			dt_01	dt_02	dt_03
Imputation	Linear Interpolation		✔	✔	✔
	ECD	MF		✔	✔
	WCD	HMM			✔
Drop remaining missing values			✔	✔	
Lost rows			2590	2286	0

Time

This group of new features, was supported by the patterns encounter in each of the temporal partitions performed in the consumption analysis of 3. In attempt to offer to the forecasting models the possibility to distinguish those patterns three features were generated with integers values that differ depending on the temporal partition, e.g. for monthly partition integers from 1 to 12 were set in the new feature. The three new features were named after each temporal partition, specifically t_month, $t\_dayofweek$, and t_hour.

Calendar

In the absence of each building real occupancy data, this category was implemented. It basically attempts to replicate the real occupancy by day, with two different features. For that, each of the features use integer values, levels, that quantify the expected daily occupancy rate, being the lowest level the one that among the others has the smallest occupancy rate. The first one, named as $s\_workday$, specifies the type of day in three different levels (level 0 - weekends and summer break days; level 1 - holidays; level 2 - workdays). The second one, named as s_epochs, uses the academic calendar to distinguish also three levels (level 0 - break period between semesters; level 1 - exams period, no classes; level 2 - classes period).

Energy consumption

This category is used to create a sort of "guidelines" using each building consumption, to enhance the performance of the forecasting models. These "guidelines" can be distinguished into two different groups, according to the technique used to generate them.

The first group used the mean daily pattern of every cluster in consumption daily analysis per building of chapter 3. Each pattern was repetitively replicated along all the data set to create a new feature. With this process, the number of features created correspondent to each cluster identified, e.g. in the case of **Civil** building, 3 features were generated, named as cl_civil_0, cl_civil_1, and cl_civil_2, which represent the same clusters identify in Figure 3.3.

The second and last group of features, was defined considering the common use of lagged features and also the high autocorrelation of each building own consumption. From each building autocorrelation, it was observed that the three most autocorrelated periods take place firstly at the previous hour, then a week before, and as lastly at the previous day. These three periods were used to generate lagged features for each building. The new features were named using the lagged period as suffix, e.g. $civil\_lag1hour$ for civil building previous hour. To complement each of the chosen periods, it was also created three other features, that provide the minimum, the maximum and the average of the three hours prior to each period, through the use of a rolling window technique. For the rolling window features name it was added a suffix to the lagged feature that characterizes the period that was selected from, e.g. for civil building one hour lagged rolling window maximum, $civil\_lag1hour\_rollmax$.

Nonetheless, each data set created in 4.1 was split per building and time horizon, based on the set of features that were available to use for each of the different forecasting horizons (an hour, a day, and a week).

4.3 Models selection

The machine learning models used, in this study, were four models: ANN(MLP), SVM, RF, and XGB. The first three models were given by scikit-learn [39] and the last one by XGBoost [40] python packages. In addition, ANN(MLP) and SVM were chosen taking into account their substance use in the literature reviewed. The other two models, RF and XGB, despite their insignificant usage when compared with the first two models in forecasting, they are emerging in the machine learning field, showing in recent studies, [31] and [32], better performance than the first two models.

Furthermore, to evaluate and compared the chosen models, each created data set, was split into two subsets: the training and the testing set. The training set was responsible for the learning process of every model and the posterior comparison of the diverse strategies employed during this work (using a time series cross-validation technique). It consists of the first two years of data, specifically, 2014 and 2017. The rest of the data, 2018, was used only to test the model's ability to forecast.

4.4 Feature Selection

The objective of this section is to select the most relevant created features, before modelling the data sets of each building and forecasting horizon (an hour, a day, and a week). This procedure, when properly applied, its known to improve the models, in terms of over fitting, accuracy, and by reducing the training time.

In this dissertation it was firstly used a filter method by Pearson correlation, to elect the most relevant WCD features, and afterwards, with the remaining features, a wrapper method, named as recursive feature elimination (RFE), was employed using XGB model.

Filter method - Pearson Correlation

When Pearson correlation between each of WCD features and the corresponding buildings was employed, it was found that half of the available features had a correlation below 0.15 with every building. That features were automatically removed, since their lack of information about each building dynamic behaviour, Figure 4.3.

As a result, the only useful features were temperature, apparent temperature, relative humidity, and mean solar radiation. In addition, the correlation between each feature was done, and as expected, the correlation between the almost similar features: temperature and apparent temperature was of 0.93. With that in mind, the first was elected (temperature) since it presents higher correlation score with every building.

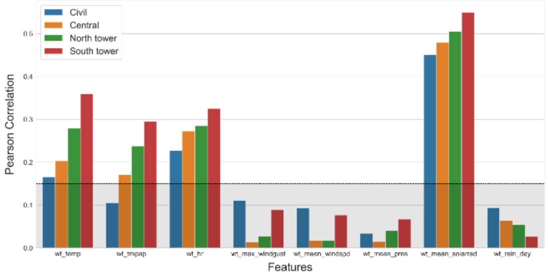

Figure 4.3 - Pearson correlation between WCD features and each building

Wrapper Method - RFE

The RFE uses a given external model, that assigns weights to each of the features. From the machine learning models selected in 4.3, only two models, specifically RF and XGB have an attribute that allows the implementation of RFE method to weight each feature and elect the best set of features. However, due to the computation expenditure of this process it was only performed by XGB model due to its computational speed characteristics.

To ensure the consistency of the selection performed by the RFE method, each set of features selected per time horizon and building was visualized and discussed. After that, some features that were considered as important based on the knowledge acquired in 3.1 energy analysis were added to the RFE method selection, creating a new selection of features.

Lastly, the forecasting models chosen in 4.3 were used to calculate the average time series cross-validation error to compare: the inexistence of selection, the RFE method selection, and the new selection.

4.5 Hyperparameter Optimization

Hyperparameter searching techniques when applied by trial and error, such as grid search and random search, may lead to an inefficient and time-consuming process, since they roam the given space of available hyperparameters values in an isolated way without paying attention to past results. To ease this process, techniques such as evolutionary algorithms and Bayesian optimization may be employed, as it was in [23] and [41], respectively. In the case of this study a Bayesian optimization grounded by Gaussian processes was used. To apply this technique a python package, named as Scikit-Optimize [42] was used.

4.6 Error metrics

Several performances measure may be used in the forecasting of energy consumption, although the ones used in this work were selected based on the most frequently applied in the reviewed literature [26]. Therefore, the two error metrics used were: the coefficient of variation of the root mean square error (CV(RMSE)) (4.1) and the mean absolute percentage error (MAPE) (4.2).

$$CV(RMSE)(\%) = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_{predict,i} - y_{true,i})^2}{n}}}{\bar{y}_{true}} \times 100 \qquad (4.1)$$

$$MAPE\ (\%) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_{predict,i} - y_{true,i}}{y_{true,i}}\right| \times 100 \qquad (4.2)$$

Where $y_{predict,i}$ is the predicted energy consumption at time point i, $y_{true,i}$ is the actual energy consumption at time point i, y_{true} is the average energy consumption, and n is the total number of data points in the dataset.

5. Results

In this chapter, the models chosen in 4.3, were compared and used for decision making of the diverse strategies adopted during the methodology. The error metric applied was the CV(RMSE) error (4.6), calculated by the time series cross-validation in the training set. The diverse strategies, namely, the data imputation study and posterior data sets generation of 4.1 and the features selection performed in 4.4, were divided by sections to ease the visualization and consequent decision. After that, the last section reveals the forecasting results of each developed model for every building and time horizon. To evaluate the forecasting results, the error metrics referred in 4.6 were used, expressed in percentage.

5.1. Data Imputation Study

As it was mentioned in 4.1, two multiple imputation algorithms (MICE and MF) were studied with the intend of filling the gaps of the original data set. Only the results for ECD imputation are represented in Table 5.1, since for WCD as said before, it was used the HMM method, Figure 4.2.

Table 5.1 - Mean Error Evaluation of 10 cycles for MF and MICE algorithms in ECD

Buildings		Civil	Central	North tower	South tower
MICE	CV(RMSE)	14.82	32.88	54.38	38.65
	MAPE	14.82	27.78	42.51	23.37
MF	CV(RMSE)	**2.46**	**12.32**	**20.13**	**16.86**
	MAPE	2.46	8.98	11.97	8.36

For the ECD the results show a clear dominance in terms of accuracy for MF algorithm. Based on that, this algorithm was chosen for the posterior data sets generation.

5.2. Data Sets Analysis

In this analysis, the data sets generated in 4.1 were compared. To procced with it the average error of the four models (ANN(MLP), SVM, RF, and XGB) in an hour, a day, and a week horizon is shown in Table 5.2.

Table 5.2 - Average time series cross-validation of CV(RMSE) of the four models for each data set, displayed by building and time horizon

Building		Civil	Central	North tower	South tower
1 hour	dt_01	13.39	16.36	23.97	24.76
	dt_02	13.80	16.43	**23.69**	**24.69**
	dt_03	**13.38**	**15.74**	23.87	24.90
1 d	dt_01	37.54	39.69	45.52	39.17
	dt_02	**37.49**	40.19	45.74	39.08
	dt_03	38.98	**39.43**	**42.90**	**38.43**
1 week	dt_01	**39.25**	44.44	48.33	41.47
	dt_02	39.61	45.17	48.16	41.69
	dt_03	40.45	**42.38**	**44.81**	**40.11**

It is possible to conclude that despite the slight error variations between the data sets, the one with the most accurate results was dt_03. This might be explained by the absence of data lost, which gives the models the possibility to improve their predictions since there is more data to learn from. In addition, the use of dt_03 allows the prediction of most of the hours of the forecasted year, one of the reasons why the data imputation study was conducted in the first place.

5.3. Feature Selection Analysis

From the RFE method, to each building and time horizon a set of features was selected. From the results obtained, it is concluded that for the same building, the features selected for different time horizons tend to decrease in ratio when the horizon of prediction increases, e.g. for North tower building, the ratios were 0.76, 0.33, and 0.14 for an hour, a day, and a week, respectively. Meaning that, several features that were considered as important in the hour horizon were neglected in the other two. With that in mind, a new set of features supported by the RFE method selection and the energy analysis was performed, as earlier mention in 4.4.

From the RFE method selection, it is mandatory to state that three type of features were considered as indispensable for every building and time horizon: , the $s\_workday$, the $cl\_(a)\_0$, and each different lagged feature when available ($(a)\_lag\_1hour$, $(a)\_lag\_1day$, and $(a)\_lag\_1week$).

Furthermore, to the new selection, four type of features supported by the following reasons were always added:

- t_hour - due to the temporal granularity of the data and since it was always selected by the RFE method for the hour horizon data set of each building;

- s_epoch - mainly due to the support that provides to the $s\_workday$ feature to address the stochastic behaviour of the daily inhabitants of each building, in terms of the different types of occupancy occurring during the different periods of the school calendar;

- For the cluster average features, e.g. $cl\_(a)\_0$, the only ones that were added were the ones that exhibit the most different average consumption patterns;

- Lastly, the weather conditions features were added in every data set, firstly because of their good correlation with each building, shown in Figure 4.3, and secondly, since they are the only features that contain information about the exact moment of prediction.

5.4. Forecasting

As a result of all the previous studies completion, this section will show Civil building forecasting by the four machine learning models chosen in 4.3 for the three time horizons.

In Civil building, for every time horizon the forecasting results did not exceed an CV(RMSE) error of 16.75%, which according to Table 5.3 was achieved in a week horizon prediction by XGB model. According to that, it is noticeable an increasing error tendency when the models attempt to predict in greater horizons, e.g. the RF model obtained CV(RMSE) errors of 7.14%, 12.18%, and 14.29% for an hour, a day, and a week horizon predictions, respectively. It may also be concluded that all the models tend to adapt easily when predicting an hour a-head consumption, achieving similar prediction accuracies, on contrary to what occurs in a day and a week a-head predictions, where the best model was easily selected.

Table 5.3 - Annual results for Civil building forecast for an hour, a day, and a week horizon

Models		MLP	SVM	RF	XGB
1 hour	CV(RMSE)	6.93	**6.32**	7.14	6.52
	MAPE	5.04	**4.22**	4.44	4.22
1 day	CV(RMSE)	13.42	**11.09**	12.18	14.3
	MAPE	10.94	**7.69**	8.63	11
1 week	CV(RMSE)	16.65	**14**	14.29	16.75
	MAPE	12.5	**10.1**	10.75	12.98

Furthermore, from all the models used, the SVM model achieved the best annual forecasting results for all the horizon predictions, in the three use error metrics. Consequently, this model was used to check the monthly predictions accuracy for each of the horizons, Table 5.4.

Table 5.4 - Civil building monthly forecast results of the best models selection for each time horizon

Months	1 hour - SVR		1 day - SVR		1 week - SVR	
	CV(RMSE)	MAPE	CV(RMSE)	MAPE	CV(RMSE)	MAPE
January	6.93	4.45	10.85	7.54	11.29	8.15
February	6.38	3.91	12.16	8.18	18.01	9.94
March	6.16	4.05	10.39	7.37	12.91	9.32
April	6.35	4.3	11.47	7.79	14.8	10.52
May	5.93	4.1	11.1	7.25	11.53	8.29
June	5.7	4.29	9.83	7.68	13.76	10.87
July	5.81	4.12	**8.12**	**5.85**	**9.68**	**7.77**
August	8.74	4.52	14.37	9.35	17.41	14.77
September	7.3	4.41	12.77	6.93	18.97	10.79
October	**5.52**	**3.86**	8.77	6.7	10.19	7.85
November	5.68	4.16	10.83	8.47	11.4	9.51
December	5.95	4.49	12.76	9.18	15.69	13.08

Based on Table 5.4, the months that yielded better results were October for an hour horizon, and July for a day and a week horizon. This might be explained by the small daily consumption variations felt in those months, making the correspondent real consumption ease to predict when compared to other months. On the other hand, the worst case scenario was found in August for an hour and a day horizon prediction, and in September for a week horizon prediction. The unpredictability of these type of months derivates from particular events that somehow change abruptly the energy consumption of the building, such as the two weeks of summer break in August and the beginning of the first semester in September.

After all, two weeks of July and August, were chosen to be visualized, as an example of the best and worst case scenarios for the all the forecasting horizons, respectively, Figure 5.1.

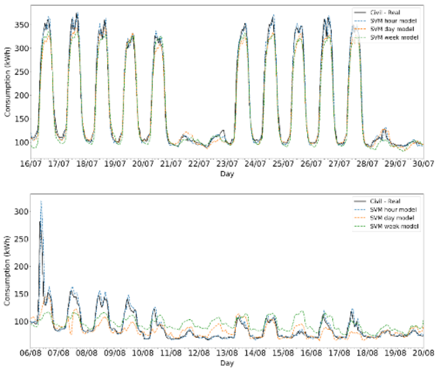

Figure 5.1 - Civil building forecasting of the best two weeks (July) and worst two weeks (August) for each time horizon best model

Table 5.5 - Day type results for Civil building best models forecast of each time horizon

Day type		Workday	Weekends	Holidays	Summer break
1 hour SVM	CV(RMSE)	6.01	**4.89**	9.24	11.52
	MAPE	4.45	**3.35**	6.46	4.47
1 day SVM	CV(RMSE)	**9.89**	12.52	26.24	22.77
	MAPE	**6.93**	8.07	17.59	10.68
1 week SVM	CV(RMSE)	**12.79**	16.00	28.17	26.16
	MAPE	**9.16**	10.95	20.86	18.16
Day	(%)	65	28	4	3

From the day type analysis results, in Table 5.5, it may be seen that the workdays were in most of the time horizons the easiest type of day to predict, which means that in 65% of the days in the three forecasting horizons the MAPE values were below 10%. Right after those type of days, weekends that account with 28% of all the days, obtained the second best predictions, displaying MAPE values below 11%. That being said, it is possible to conclude that 92% of the days (workdays and weekends) were successfully forecasted in the three time horizons when compared to the remaining 8% of days that included holidays and the two weeks of summer break, previously shown in Figure 5.1.

6. Conclusions

In this work, four machine learning models (MLP, SVM, RF, and XGB) were compared in three different forecasting horizons (an hour, a day, and a week) for four main buildings (Civil, Central, North tower, and South tower) located at Instituto Superior Técnico, Lisbon. Giving a total of 48 models developed. To conduct this study, two types of hourly collected data (WCD and ECD) for three years (2014, 2017, and 2018) was used. In order to create the right conditions for each forecasting model development, the available data was first treated and then analysed.

In the first stage, a data imputation study was performed, concluding that from the two multiple imputation algorithms used (MICE and MF), the one that achieved greater results, was the MF algorithm, which was successfully applied in each building consumption data incompleteness (ECD). However, in the presence of simultaneous wide gaps of missing values, which occurred in WCD, this algorithm was not able to supply the tendency and seasonality intrinsic to the type of variable that was being filled. For that reason, a single imputation method created with that propose (HMM) was employed, guaranteeing with that the necessary dynamic of the imputed variable.

In the second stage of this work, each building energy consumption was analysed, and several features were created in attempt to better define the buildings' behaviour. With all the features generated, a feature selection analysis via the RFE method took place to understand the importance of each feature. From that analysis, it was concluded, that three type of features were indispensable in every building and forecasting horizon data set, specifically, the day type, the lagged features (one hour, one day, and one week), and the cluster average consumption of the larger group of days of each building from the year of 2017. In addition, it was also noticeable that the used of this method was too restrict in the feature selection process, which might be explained by the use of the clusters average consumption features that were generated from one of the years where this method was applied (training set), influencing with that the importance of each feature and the consequent selection. To overcome that situation a new selection supported by the RFE method selection was employed.

Furthermore, after the hyperparameter optimization and the forecasted results obtained, it was concluded that, although, all the models developed did not show larger error variations when predicting the same building consumption and forecasting horizon, SVM model outstood, achieving the most accurate results in the majority of the predictions performed. One of the reasons which might explain this situation is supported by the small number of hyperparameters that this model has, leading to reach its full potential faster and easiest than any other model used in this work. Moreover, the second most accurate model was XGB performing the best prediction in three of the cases, followed by the MLP and RF that achieved the highest accuracy value in one case each. Considering the best models' selection, in most of the buildings and forecasting horizons it was found that August was the hardest month to predict, due to the intrinsic unpredictability of two summer break weeks of all Alameda's campus facilities. On the other hand, in 93% of the predicted days, which accounts with working days (65%) and weekends (28%) the buildings achieved an MAPE error of 10.95%, 9.17%, 10.48%, and 12.66% for Civil, Central, North tower, and South tower buildings in a week horizon prediction. In addition to that, it was noticeable an increasing annual error tendency when the models attempt to predict in greater horizons.

Afterall, using the CV(RMSE) metric to compare the different buildings predictions, it was found that in every forecasting horizon studied, Central building was the easiest one to predict and South tower building the hardest one. Which might be related by the small variations of consumption found in the first study case, against the sudden peaks and dips encounter in the latter.

6.1. Future Work

In this work, it was concluded that the stochastic behaviour of each building's inhabitants is one of the major reasons for the variations of the energy consumption patterns. With that in mind, two solutions can be adopted:

- The simplest one, is to perform a different day type feature for each of the buildings, since it was shown that different buildings have different consumption patterns in the same type of day;

- The second solution, it only can be applied in buildings with Wi-Fi systems, with that being said, it basically uses each people network log in to count the online number of occupants in each of the buildings.

Another future work that may be explored is supported by the fact that each model achieved similar error measurements. With that, an ensemble model with an evolutionary algorithm to define each model percentage in the final output may be employed, which, normally, leads to a generalized and enhanced prediction.

Lastly, it has been shown in several recent studies the tendency of using deep learning models to predict sequential data, such as time series. The use of those models such as long-short term memory (LSTM) could be beneficial, due to its intrinsic long-term dependencies in recurrent architectures.

6.2. Contributions

From this study, four contributions may be stood out:

- An imputation study for this study dataset was developed;
- Adaptation for the first time of XGB model to forecast building energy consumption in hourly granularity;
- Development of forecasting models in three different horizons for four buildings located at Alameda campus , IST, Lisbon;
- The code developed along this study is publicly available in [43].

References

[1] European Commission, "Buildings - European Commission," 2018. [Online]. Available: https://ec.europa.eu/energy/en/topics/energy-efficiency/buildings. [Accessed: 30-Nov-2018].

[2] Eurostat, "2014 energy consumption by sector in the EU," 2014. [Online]. Available: https://epthinktank.eu/2016/07/08/energy-efficiency-in-buildings/energy-consuption-by-sector/. [Accessed: 06-Dec-2018].

[3] M. Leahy, J. L. Barden, B. T. Murphy, N. Slater-thompson, and D. Peterson, "International Energy Outlook 2013."

[4] E. Commission, "European Commission, Action Plan for Energy Efficiency: Realizing the Potential, Communication from the Commission," *Eur. Comm. Action Plan Energy Effic. Realiz. Potential, Commun. from Comm.*, 2006.

[5] H. X. Zhao and F. Magoulès, "A review on the prediction of building energy consumption," *Renew. Sustain. Energy Rev.*, vol. 16, no. 6, pp. 3586–3592, 2012.

[6] H. X. Zhao and F. Magoulès, "A review on the prediction of building energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6. pp. 3586–3592, 2012.

[7] F. Rabhi *et al.*, "An overview of the commercial cloud monitoring tools: research dimensions, design issues, and state-of-the-art," *Computing*, vol. 97, no. 4, pp. 357–377, Apr. 2014.

[8] J. Sides, "The Victory Lab: The Secret Science of Winning Campaigns," *Public Opin. Q.*, vol. 78, no. S1, pp. 363–364, Jan. 2014.

[9] W. J. Kuo, R. F. Chang, D. R. Chen, and C. C. Lee, "Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images," *Breast Cancer Res. Treat.*, vol. 66, no. 1, pp. 51–57, Mar. 2001.

[10] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies," *Renewable and Sustainable Energy Reviews*, vol. 81. 2018.

[11] G. R. Newsham and B. J. Birt, “Building-level occupancy data to improve ARIMA-based electricity use forecasts,” in *BuildSys*, 2010, p. 13.

[12] S. R. Twanabasu and B. A. Bremdal, “Load forecasting in a smart grid oriented building,” in *22nd International Conference and Exhibition on Electricity Distribution (CIRED 2013)*, 2013, pp. 0907–0907.

[13] N. Mohamed, M. H. Ahmad, Suhartono, and Z. Ismail, “Improving short term load forecasting using double seasonal arima model,” *World Appl. Sci. J.*, vol. 15, no. 2, pp. 223–231, 2011.

[14] J. Zhuang, Y. Chen, X. Shi, and D. Wei, “Building Cooling Load Prediction Based on Time Series Method and Neural Networks,” *Int. J. Grid Distrib. Comput.*, vol. 8, no. 4, pp. 105–114, 2015.

[15] Y. K. Penya, C. E. Borges, D. Agote, and I. Fernández, “Short-term load forecasting in air-conditioned non-residential Buildings,” in *Proceedings - ISIE 2011: 2011 IEEE International Symposium on Industrial Electronics*, 2011, pp. 1359–1364.

[16] J. Colomer, L. Burgas, J. Massana, C. Pous, and J. Melendez, “Short-term load forecasting in a non-residential building contrasting models and attributes,” *Energy Build.*, vol. 92, pp. 322–330, Apr. 2015.

[17] D. Zhao, M. Zhong, X. Zhang, and X. Su, “Energy consumption predicting model of VRV (Variable refrigerant volume) system in office buildings based on data mining,” *Energy*, vol. 102, pp. 660–668, May 2016.

[18] M. Borovcnik, H.-J. Bentz, and R. Kapadia, “A Probabilistic Perspective,” in *Chance Encounters: Probability in Education*, 1991, pp. 27–71.

[19] P. A. González and J. M. Zamarreño, “Prediction of hourly energy consumption in buildings based on a feedback artificial neural network,” *Energy Build.*, vol. 37, no. 6, pp. 595–601, Jun. 2005.

[20] S. Eric, “Predicion and control using feedback neural networks and partial models,” 1996.

[21] R. Yokoyama, T. Wakui, and R. Satake, “Prediction of energy demands using neural network with model identification by global optimization,” *Energy Convers. Manag.*, vol. 50, no. 2, pp. 319–327, 2009.

[22] R. Yokoyama and K. Ito, “Capability of Global Search and Improvement in Modal Trimming Method for Global Optimization,” *JSME Int. J. Ser. C*, vol. 48, no. 4, pp. 730–737, 2006.

[23] Y. T. Chae, R. Horesh, Y. Hwang, and Y. M. Lee, “Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings,” *Energy Build.*, vol. 111, pp. 184–194, Jan. 2016.

[24] Q. Li, Q. Meng, J. Cai, H. Yoshino, and A. Mochida, “Applying support vector machine to predict hourly cooling load in the building,” *Appl. Energy*, vol. 86, no. 10, pp. 2249–2256, Oct. 2009.

[25] X. Li, J. H. Lǔ, L. Ding, G. Xu, and J. Li, “Building cooling load forecasting model based on LS-SVM,” *Proc. - 2009 Asia-Pacific Conf. Inf. Process. APCIP 2009*, vol. 1, pp. 55–58, 2009.

[26] L. K. Hansen and P. Salamon, “Neural Network Ensembles,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 10, pp. 993–1001, 1990.

[27] D. Opitz and R. Maclin, “Popular Ensemble Methods: An Empirical Study,” *J. Artif. Intell. Res.*, vol. 11, pp. 169–198, Jul. 1999.

[28] F. Zhang, C. Deb, S. E. Lee, J. Yang, and K. W. Shah, “Time series forecasting for building energy consumption using weighted Support Vector Regression with differential evolution optimization technique,” *Energy Build.*, vol. 126, pp. 94–103, Aug. 2016.

[29] C. Fan, F. Xiao, and S. Wang, “Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques,” *Appl. Energy*, vol. 127, pp. 1–10, 2014.

[30] M. W. Ahmad, M. Mourshed, and Y. Rezgui, “Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption,” *Energy Build.*, vol. 147, pp. 77–89, 2017.

[31] Z. Wang, Y. Wang, R. Zeng, R. S. Srinivasan, and S. Ahrentzen, “Random Forest based hourly building energy prediction,” *Energy Build.*, vol. 171, pp. 11–25, 2018.

[32] C. Robinson *et al.*, “Machine learning approaches for estimating commercial building energy consumption,” *Appl. Energy*, vol. 208, no. May, pp. 889–904, 2017.

[33] Z. Wang and R. S. Srinivasan, “A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models,” *Renewable and Sustainable Energy Reviews*, vol. 75, no. September 2015. Elsevier Ltd, pp. 796–808, 2017.

[34] M. Ana, "Plataforma de gestão de energia para o Campus do IST: Modelação e representação dos consumos de energia", Instituto Superior Técnico, 2019.

[35] P. Francisco, “Machine Learning applied to energy demand forecast in IST Alameda campus”, Instituto Superior Técnico, 2019.

[36] S. van Buuren and K. Oudshoorn, “Flexible multivariate imputation by MICE,” pp. 1–20, 1999.

[37] D. J. Stekhoven and P. Bühlmann, “Missforest-Non-parametric missing value imputation for mixed-type data,” *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[38] Python Software Foundation, “Welcome to Python.org,” *2001*, 2017. [Online]. Available: https://www.python.org/. [Accessed: 19-Oct-2019].

[39] “scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation.” [Online]. Available: https://scikit-learn.org/stable/index.html. [Accessed: 30-Oct-2019].

[40] Xgb. Developers, “XGBoost Documentation,” *XGBoost*, 2016. [Online]. Available: https://xgboost.readthedocs.io/en/latest/. [Accessed: 30-Oct-2019].

[41] X. Li, L. Ding, and L. Li, “A novel building cooling load prediction based on SVR and SAPSO,” *3CA 2010 - 2010 Int. Symp. Comput. Commun. Control Autom.*, vol. 1, no. 1, pp. 528–532, 2010.

[42] “skopt API documentation,” *scikit-optimize.github.io*. [Online]. Available: https://scikit-optimize.github.io/. [Accessed: 14-Oct-2019].

[43] “FDelca/energy_consumption_forecasting.” [Online]. Available: https://github.com/FDelca/energy_consumption_forecasting. [Accessed: 31-Oct-2019].