Data Mining for Smart Cities: Predicting Electricity Consumption by Classification

Konstantinos Christantonis
The Data Mining and Analytics Research Group
School of Science & Technology
International Hellenic University
Thessaloniki, Greece
k.christantonis@ihu.edu.gr

Christos Tjortjis
The Data Mining and Analytics Research Group
School of Science & Technology
International Hellenic University
Thessaloniki, Greece
c.tjortjis@ihu.edu.gr

***Abstract*—**Data analysis can be applied to power consumption data for predictions that allow for the efficient scheduling and operation of electricity generation. This work focuses on the parameterization and evaluation of predictive algorithms utilizing metered data on predefined time intervals. More specifically, electricity consumption as a total, but also as main usages/spaces breakdown and weather data are used to develop, train and test predictive models. A technical comparison between different classification algorithms and methodologies are provided. Several weather metrics, such as temperature and humidity are exploited, along with explanatory past consuming variables. The target variable is binary and expresses the volume of consumption regarding each individual residence. The analysis is conducted for two different time intervals during a day, and the outcomes showcase the necessity of weather data for predicting residential electrical consumption. The results also indicate that the size of dwellings affects the accuracy of model.

Keywords*—Smart Homes, Smart Cities, Data Mining, Prediction, Classification*

I. Introduction

The exchange of data and information between systems and users has impelled society to assign the term *smart* in almost everything. From smart phones and watches to smart devices all around the dwelling, it is clear that there is a predisposition of people for intelligent systems, which could positively influence their everyday life. Such a system could provide residents with a forecast of the next day's electricity consumption as accurately as possible, so they can plan their activities wisely, i.e. shifting the load of their needs to lower demand periods, to save money and act in a more environmentally friendly way. A smart home is not desired only by residents, but also by businesses that link their products and services with the intelligence factor. Therefore, businesses show willingness to invest in such structures, as they can be proven to bring about a higher quality and therefore a larger profit margin.

However, when the data fail to be captured and stored in databases on a prerequisite way, the whole process is jeopardized. In data mining, especially in real-world scenarios, it is almost impossible to capture all the desired data that could bring up value, due to the data interoperability, which is still a major problem. In addition, data are usually generated through installed sensors, which quite often fail or deviate from the actual values. When the access and administration of data is no longer a problem, then data mining or knowledge discovery techniques turn these "nonsense" values into valuable information. In general, data mining is used to obtain new perspectives and capture hidden factors from unexploited information, which is available in the collected data; however, it is also a scientific field that can validate hypotheses and experience-based knowledge. Data mining is considered the key technology behind smart cities and homes. Accurate predictions of daily unusual behaviors or early warnings of dangerous moments can be crucial and are being systematically studied.

Based on that, different projects of smart cities have begun to develop around the globe in order to capture previously unknown knowledge or test intuitively assumptions. Electricity consumption prediction is a vast field that increasingly attracts researchers' attention. A major reason why this problem is being constantly tracked is that weather is an unstable variable and does not affect equally all the climate zones.

As part of this work, we examine the problem of electricity consumption forecasting. The first part focuses on exploiting the gathered weather data to produce building energy consumption forecasting models while the second aims to simulate the behaviour of grid through the aggregation of consumption of all available homes.

For this purpose, open data from the UMass Trace Repository were used. The selected dataset, named *Home*, includes both weather and consumption data collected for seven different households for three consecutive years (2014-16). However, since some homes were not compatible with each other, it was decided that they are excluded. Indeed, several weather metrics were not exactly matching, neither the recording time of these metrics was identical. Inevitably, only three homes were used to assess the repeatability of the models, which differ a lot, in terms of size and consuming behaviour. The labels of the Homes are identical to those introduced by the repository. Therefore, the homes will be referred as Home B, Home C and Home F. The houses under examination are in Massachusetts, USA. Obviously, weather conditions are similar as the houses are relatively close to each other.

The remaining of this paper is organised as follows: section 2 provides context by reviewing the literature. Section 3 defines the problem we address and details our approach, while section 4 presents experimental results, which are discussed in section 5. The paper concludes by discussing threats to validity and presenting future work.

II. Background

So far, many research teams have relied on data provided by the Smart* project with different goals and directions. In

[1] weather and energy data were processed to predict the latitude and longitude of a smart meter that collects data, as all these energy data are collected anonymously. In addition, in a more market-oriented approach of equal significance, [2] introduced an intelligent charging system called Smart Charge, whose goal is to decrease the electricity bills by shifting consumption to lower price periods. Similarly, SmartCap is a system for monitoring and controlling electric loads, which tries to flatten electricity demand through scheduling algorithms [3].

Besides prior published work based on the particular *Smart** project, it is very interesting to trace back into significant approaches around the general topic of electrical management systems, such as [4], and the prediction of consumption. More than ten years earlier, [5] compared different structures of ANN's on forecasting, concluding on a 2-hidden layer instruction of form 12-16-16-1 that achieved the highest performance. ANN's are also examined and compared in [6] with SVM algorithm in an extensive review of building electrical energy consumption bibliography that results in superiority of Least squares Support Vector Machine (LsSVM) algorithm.

In [7] customer profiles are pre-defined, based on the total consumption, thus a significant decrease on the daily expenses in electricity is achieved. In contrast, [8-10] propose to first build customer profiles or assign some already known and then try to predict consumption based on them. Another analysis under the umbrella of flattening in consumption is the outlier analysis and the focus on anomaly detection, as it considered a major factor that affects consumers. Since the market's linearization, the kilowatt-hour price changes regularly, and it is expected to minimize unreasonable extra costs by understanding and preventing such phenomena. Such a research was conducted in [11], where the ARIMA methodology was applied.

Load forecasting is a well-studied subject, which has influenced several researchers to try different techniques and approaches. However, there is still a lot of room for improvement as well as several unexplored aspects in such a multidimensional problem. Authors in [12] present three time-based approaches around the forecast, as well as their characteristics and value for the provider; Short-Term (ST), Medium Term (MT) and Long Term (LT). The ST load forecasting could be used for reducing costs and secure operation of power systems. In MT load forecasting the interest focuses on normal operation, while LT load forecasting is studied to ensure safer investments and long-term planning in general. A very interesting approach of load forecasting in distribution system is presented in [13], where Principal Components Analysis (PCA) is applied to multi-linear regression on MT load forecasting.

Regarding weather factors, [14] represents the responses of energy demand due to climate change in Massachusetts. The parameters that are being used describe the Heating Degree-Day (HDD), the hours of daylight and the electricity price in a monthly scale for both residential and commercial sector. The study concludes that 'energy demand in Massachusetts is sensitive to temperature' while the average number of days exceeding 90°F will rise to double by 2030.

The results of prediction cannot always be that accurate, even if the weather indeed affects largely the electricity consuming behaviour. As explained in [13], electrical energy demand has a high non-linear behaviour, thus accurate predictions cannot be guaranteed. Another factor that affects negatively prediction accuracy is the continuous pressure for better living standards [15] on a disproportionate rate. Indeed, according to [16] where the residential electrical consumption in Brazil was analysed, the increase in electricity demand was faster than in income. In a similar study for a very different climate, [17] resulted in similar results regarding temperature, however, as it was highlighted, 'relative humidity is not having significant impact on the energy consumption'.

The base period of forecasting, briefly, affects the complexity of the problem that needs to be modelled. Most studies focus on the ST load forecasting, as it is a more difficult task, due to the noisy effect of environmental factors. As stated in [18], electric power consumption is growing rapidly and introduces a higher level of randomness, due to the increasing effect of environmental and human behaviour. In addition, it is important to understand that like many time-series problems, load forecasting also reflects a seasonality and cyclic component, which leads many researchers to the vertical decomposition throughout the year. The load pattern is a non-stationary time-series problem and thus needs to be carefully fragmentized.

III. Approach and problem definition

Energy consumption prediction in a particular building is usually influenced by many factors, such as the electrical appliances or devices in it, its geographic location, as well as the time range it is operational [6]. Occupancy is also such a factor but is not easy to record it in highly operational buildings. In general, citizens have a decent understanding about the appliances that consume higher rates of electricity; however, the constant change in price prevents the creation of a clear consumption plan. For example, air-conditioning represents the biggest part of electric energy consumption in residential buildings [20], while according to [21] electrical energy consumption increases on summer months over 2.5 % because of the rise in temperature.

Our approach aims at constructing a prediction model which includes usage patterns besides weather data. As long as the target is to predict the fluctuation of consumption individually for each *Home*, the consuming behaviour should also be considered. Authors in [12] state that "Electric demand is often considered as a function of weather variables and human social activities". More specifically, typical families have cycles of consumption on a daily and weekly basis. In general, families use the laundry or any other appliance *X* times per week. If the consumption stays low for a consecutive number of days, it is more likely that next days will show a rise in total consumption. Similarly, the previous day consumption should also be a factor to predict the next day. It is expected for a house that consumes a higher than average amount of electricity, that the next day would result in lower consumption.

To construct such a model, it is essential to guarantee that there is a constant tracking of the total consumption in each home. Unfortunately, smart metering is a newly deployed technology that still must surpass numerous challenges and failures and is heavily analysed in several works, such as [29]. It is also key, to clarify that since the available data are limited and the level of information low, this study focuses on forecasting as a result of a binary classification. Reforming the problem from regression to classification was a crucial step in our work. The main reason for that, is our assumptions regarding the level of information that residents would be interested in. Ideally the outcome should be the exact consumption value, but this research is conducted with a broader motivation to suggest and highlight the factors that could affect a prediction either positively or negatively on individual home characteristics. Through this work we aim to investigate how effective would the model be for each of the available homes and which reasons lead to differentiation in accuracy. Since that was decided, the model was transformed into binary around different means. The two labels are 'High' and 'Low' describing the volume of consumption. These means reflect the mean values of consumption throughout the year, the current season or in a monthly basis. This happens mostly because on smoother weather conditions (Spring or Autumn) it is expected for model to return lower accuracy. The binarisation of consumption is performed around standard mean values in order to avoid the 'expensive' handcrafted data engineering. Thus, the whole process can be easily automated

There are many factors that affect consumption. In our case, consumption forecasting is done in the much smaller scale of individual residence for the day ahead. Initially, what should be clear is that such a model includes only variables that are set to be known in advance. The model includes mostly weather data. From the available weather data that match the Homes, only the following were selected:

- Temperature (°F)
- Apparent Temperature (°F)
- Wind Speed (Mph)
- Wind Direction (Bearing)
- Humidity (%)
- Dew Point (°F)
- Weather Icon (Categorical)

There were also some weather metrics that we decided to exclude from the model since the strategy formatting. *Precipitation probability* and *precipitation intensity* were two of them, as it was expected for the level of the information that they could provide to be captured from the categorical metric, *weather icon*. Indeed, the most dominant category was clear, so obviously that leads to zero intensity. Moreover, *visibility* might seem a reasonable variable, but it was not expected to bring higher value to the model than complexity. Beyond that, *cloud cover* had numerous missing values, while *pressure* was not considered as a metric easily understandable by the residents, so as to change their living habits.

In addition, a simplistic dummy-like variable is created that indicates if a day fell on a weekend. Another variable we also introduced, to indicate weekends, if a day fell on a national holiday, as listed here [22]. The last variable records the time of sunset for each day, in a categorical form of five different time intervals (16:00-20:00).

Based on the previous steps it was decided that the model should focus on two different time intervals, on and off-peak hours. Obviously, as in most residences, on Home B the peak consumption hours were between 15:00 and 21:00 so it was decided for the *Usage* (target variable) to be averaged under that time-interval. The off-peak consumption hours were a tricky part as the intention was to be of equal duration as the first, and not overlapping with each other. The remaining three 6-hour intervals are expected to have similar behavior; however, it was desired for only one to be chosen.

Given that providers tend to introduce night tariffs and rates, it was decided that since the two variables of our model refer to Weekends and Holidays, an appropriate approach would be to choose the 09:00-15:00 interval. The on and off-peak hours, is a concept that should be separately adjusted for each Home, based on the occupants' habits. If the purpose is Grid load forecasting, then of course this separation is unified. There are studies that choose different time intervals. For example, the peak period in [23] was defined as 7am to 7pm, Monday to Friday, while all the remaining times and public holidays were considered as off-peak. Consumer behavior is estimated to be more accurate described in the mornings on holidays and weekends than with nightly habits that are pretty much the same. Therefore, the scale of our data is daily, and the same process is followed for weather data. Despite that the available dataset contains hourly information for 2014-16, the targets turn into daily stamps. For example, the average temperature of 01/01/2014 equals to twenty, which is calculated through the half-hourly values between 15:00 and 21:00. Table I shows the number of instances that each class includes for each of the described splits.

TABLE I: BALANCE OF THE TWO CLASSES AROUND THE EXAMINED MEANS

Homes -Time Interval	Consumption Per Total		Consumption Per Month		Consumption Per Season	
	Classes		Classes		Classes	
	High	Low	High	Low	High	Low
Home B ON-peak	361	735	431	665	378	718
Home B OFF-peak	372	724	400	696	393	703
Home C ON-peak	395	685	382	698	380	700
Home C OFF-peak	408	672	390	690	398	382
Home F ON-peak	553	543	553	543	547	549
Home F OFF-peak	438	658	447	649	452	644

The splits reveal that the two classes are rarely balanced and almost in every case the *Low* class outnumbers the *High* one. Also, for classification problems is important to clarify which class is of higher interest. In our case both classes are considered of equal interest.

Based on uncertainty, as it was described above, scientists have already started looking for effective ways to forecast the electricity consumption and therefore the electricity price. Probably the most challenging part is to obtain reasonable data regarding the area of interest. This data gathering leads to valuable results, which mainly affect the life of residents in positive manners, but also allows providers to reschedule their generating and distributing plans. On these days, providers are turning to smart grids that focus on real-time pricing or critical peak rebate. Indeed, according to [24] those kinds of smart pricing are already increasing in the USA and more specifically in States like California and Massachusetts, “this is actually being mandated by the State legislatures”.

IV. EXPERIMENTAL RESULTS

Before reaching the results that each home brings upon, the following figures [1-6] give an indication of the consuming behaviors for each of them. Besides the average hourly consumption, it is important to also examine the monthly averages.

In general, Home B has a tremendous differentiation throughout the three years. Both in Figure 1 and 4 the volume of consumption, despite that follows a similar pattern also increases. On the other hand, Figure 2 and 3 show an insignificant increment.

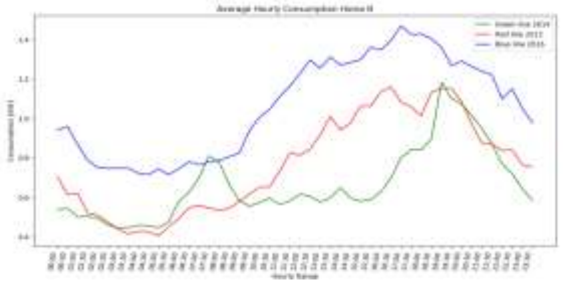

Figure 1: Home’s B avg. hourly consumption

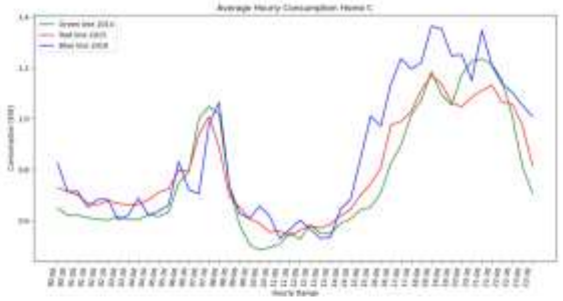

Figure 2: Home’s C avg. hourly consumption

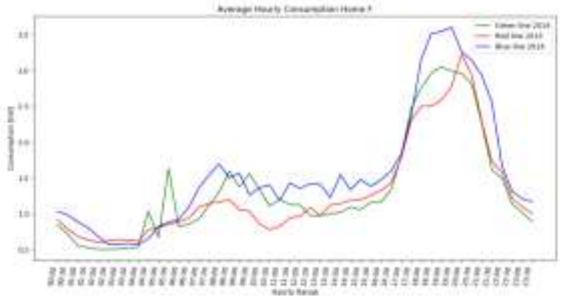

Figure 3: Home’s F avg. hourly consumption

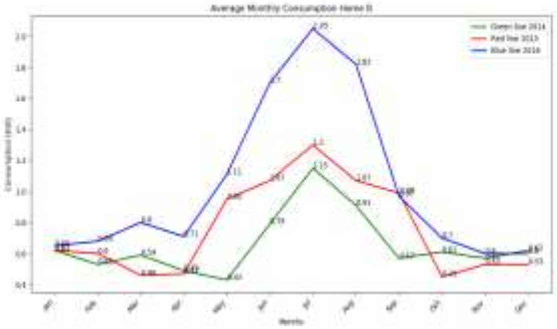

Figure 4: Home’s B avg. monthly consumption

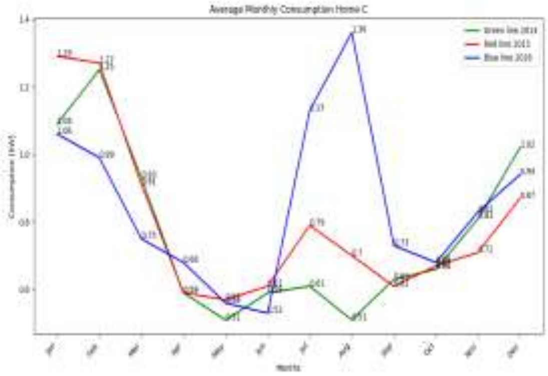

Figure 5: Home’s C avg. monthly consumption

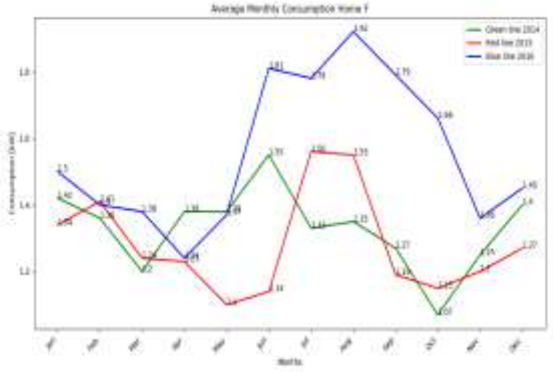

Figure 6: Home’s F avg. monthly consumption

Some more characteristics about the three houses are the following: Home B, as described in [25], is a huge house across two stories with eight rooms and four full-time occupants. It is roughly 1700 square feet and it contains a central A/C as well as a gas-powered heating system. Home C is almost double the size of Home B, around 3500 square feet, again across two stories. Unfortunately, the real number of occupants is unknown. Home C also generates power which not only covers some of the electricity demands but is also possible to ‘reverse direction when the home’s generation exceeds its consumption’. Home F does not come with a description as it is included on the dataset as an update. Information about Home F is expected to be published later in 2019.

The algorithms that are tested are Support Vector Machines (SVM), Random Forest, Stochastic Gradient Descent (SGD) and Logistic Regression. Following, there are three different technical stages of classification while there are three different classification implementations. The first one splits the total consumption instances around the general mean value. The second one splits the total consumption instances regarding each month's mean value, while the last one regarding each season's mean value. Each algorithm will be examined for each of the following stages:

Stage 1 represents the results that are given by calling the default algorithms as they are set on Scikit-Learn package.

Stage 2 reflects what happens when the training data are scaled, due to the different units and ranges. For some classifiers such as SGD this is a crucial step, while others are not affected.

Stage 3 reflects a hyper-parameter tuning as it can boost algorithmic performance. For each algorithm, the tuned parameters are chosen empirically, so not all of them are set to be tuned.

It is important to clarify that for SVM different kernels delay the process. This happens mostly because the linear kernel is an almost identical implementation with SGD's hinge kernel. Moreover, the polynomial kernel requires the data to be scaled. For the assessing of generalisation of our model the most common approach of a 10-fold cross-validation was used. Cross validation is a resampling procedure, used to evaluate classification models on a limited data sample [26]. Shortly this concept splits the available data into k random groups of equal size if possible and uses each time one of these groups (folds) as the test set; the rest k-1 folds are used for training. Each fold is used after the completion once, and the final score is the average of all these tests. The results clearly indicate that there is not an algorithm superior to others. Logistic regression required less hyper-parameter tuning and clearly was not highly affected by that.

The most stable algorithms are random forest and SVM, however the latter is slower, while it shows unpredictable behavior during the scaling stage. Regarding the two different time intervals, it was expected for the off-peak period to achieve higher results, but the results concluded on the exact opposite. This was initially assumed since on off-peak period the fluctuations of consumption are smaller, but as the experiments resulted; higher fluctuations lead to more information.

Table II illustrates that the most promising results were achieved for Home B and especially the ON-peak time interval. On the other hand, Home C and F are not showing significant differences and all the algorithms are performing similarly. The last indication from results is that when the binary transformation is being around the general mean value of consumption the accuracy is higher. In addition, as we shorten the "focus" of transformation, the more difficult it is to achieve successful predictions.

V. Grid Load

In a similar way, the grid load could be simulated in order to assist the providers. Since all the houses are in the same region, weather data are very similar, thus an average value for each weather metric is calculated. Regarding the electrical consumptions, the values are summed both for *Yesterday* and *Past week* load. The target values again are transformed into two classes, however, since previous analysis resulted in better performance for the general mean value, it was decided this to be the only one to be examined. The split of classes for On-peak period is 474 High-606 Low, while for Off-peak period it is 433 High-647 Low. The structure of the model remains the same and it is presented in Tables III and IV.

As seen in tables III and IV, the performance of the model does not change drastically. However, at this point a different perspective of generalisation can be examined through merging. The accuracy for both periods remains similar however, for Off-peak period a slightly higher accuracy is achieved.

TABLE II: Final results for each stage of all the algorithms for both time intervals

Algorithms	Homes Time Inr	STAGE 1	STAGE 2	STAGE 3	STAGE 1	STAGE 2	STAGE 3	STAGE 1	STAGE 2	STAGE 3
		Consumption Per Total			Consumption Per Month			Consumption Per Season		
SVM	HOME B ON PEAK	0.8869	0.8705	0.891	0.6825	0.7043	0.7166	0.7138	0.7411	0.7493
RF		0.8773	0.8773	0.895	0.643	0.643	0.6934	0.7043	0.7043	0.7561
SGD		0.5986	0.7752	0.8801	0.6021	0.6267	0.7029	0.6035	0.6457	0.7397
LR		0.8746	0.876	0.8828	0.6798	0.6771	0.7029	0.7356	0.7288	0.7411
SVM	HOME B OFF PEAK	0.7833	0.7915	0.8024	0.6839	0.7002	0.7084	0.6975	0.7057	0.7125
RF		0.7724	0.7724	0.7973	0.6675	0.6675	0.7152	0.6811	0.6811	0.7179
SGD		0.6811	0.7152	0.7915	0.5572	0.598	0.6975	0.5054	0.6117	0.7057
LR		0.797	0.7847	0.797	0.6961	0.692	0.6989	0.6907	0.6893	0.6989
SVM	HOME C ON PEAK	0.7441	0.7339	0.7759	0.6652	0.6929	0.697	0.6694	0.6984	0.7233
RF		0.7261	0.7261	0.7676	0.65	0.65	0.7123	0.6556	0.6556	0.7109
SGD		0.5532	0.6846	0.7897	0.5311	0.5975	0.6915	0.6334	0.6639	0.7192
LR		0.7842	0.7773	0.7869	0.6929	0.6929	0.6984	0.7095	0.7081	0.715
SVM	HOME C OFF PEAK	0.7233	0.7634	0.7731	0.6307	0.6957	0.7136	0.6237	0.7178	0.7385
RF		0.7219	0.7219	0.7593	0.6666	0.6666	0.7136	0.668	0.668	0.7247
SGD		0.6915	0.6777	0.7717	0.6071	0.6071	0.715	0.6002	0.6559	0.7289
LR		0.7745	0.7662	0.7745	0.7067	0.7109	0.7136	0.7206	0.7192	0.7316
SVM	HOME F ON PEAK	0.564	0.6784	0.6839	0.5217	0.6512	0.6621	0.5299	0.6798	0.6852
RF		0.643	0.643	0.6866	0.5912	0.5912	0.6512	0.6212	0.6212	0.6716
SGD		0.5027	0.5855	0.6757	0.5068	0.5871	0.6607	0.4986	0.5994	0.6662
LR		0.6825	0.6716	0.6852	0.6662	0.6457	0.6716	0.673	0.6634	0.6811
SVM	HOME F OFF PEAK	0.643	0.6757	0.6811	0.6185	0.6294	0.6416	0.6008	0.6253	0.6471
RF		0.6267	0.6267	0.6893	0.6076	0.6076	0.628	0.5953	0.5953	0.6348
SGD		0.5313	0.6335	0.6771	0.5027	0.5871	0.6376	0.5231	0.5912	0.6362
LR		0.6716	0.6689	0.6784	0.6294	0.6294	0.6416	0.6294	0.6376	0.6485

TABLE III. ON-PEAK / GENERAL MEAN VALUE

Grid	On-peak / General mean value			
	SVM	*Random Fst*	*SGD*	*Logistic Regr*
Stage 1	0.6846	0.6680	0.5892	0.7026
Stage 2	0.7358	0.6680	0.6154	0.7150
Stage 3	0.7358	0.7178	0.7247	0.7192

TABLE IV. OFF-PEAK / GENERAL MEAN VALUE

Grid	Off-peak / General mean value			
	SVM	*Random Fst*	*SGD*	*Logistic Regr*
Stage 1	0.6376	0.6860	0.5767	0.7302
Stage 2	0.7219	0.6860	0.6528	0.7275
Stage 3	0.7275	0.7495	0.7247	0.7302

VI. DISCUSSION

Electrical load forecasting is a complex problem, which needs detailed design and a deep understanding of the domain. The purpose of this work was not to build a state-of-the-art model, but to analyse and review different algorithms having applied appropriate pre-processing and selected suitable parameters. The bibliography generally suggests several factors that could positively affect the electrical load predictability; however, due to the limited available data, the effect of weather conditions on the electrical load forecasting was examined.

The overall aim of this study was to examine this general problem within the concept of smart city. Unlike many studies, which attempt to predict electricity consumption of the grid or large blocks of apartments, this work focuses on single households. In contrast with commercial buildings, where electricity consumptions follow a pattern (for example 08:00 to 17:00 with a decline around 13:00 during the lunch break), occupant behaviour has a major effect on single households. Thus, in single household scale, a thoroughly prediction model is hard to be established.

The model accuracy for two out of the three houses was over 75%. Higher accuracy was not expected in this analysis, as major factors (such as occupancy or activity inputs) were not available, thus not incorporated into the model. In general, it would be easier to extract safer conclusions if the initial data met the criteria. Furthermore, the decomposition of this time-series problem into explanatory input variables gives more space for creativity and understanding of the problem itself.

For smart cities administrators and electricity providers, such an approach, would face the problem of cold start. For a consumer to receive information and forecasts about their home, it is necessary to record and process data for a long time. However, this approach, which decomposes the time factor and assumes that same weather conditions induce same consumer behaviour, irrespective of the day or season, can be implemented much faster.

VII. CONCLUSION

This research was conducted mostly with a citizen-centric approach; regarding the algorithms used, the basic conclusion is that the complexity of the model was not as high as to allow one of them to stand out. All the results are very close and sometimes identical. The number of involved houses may be very small, but some safe conclusions can be extracted. The most important results of this work are summarized below:

- The accuracy of the model increases when the consumption follows similar paths throughout the months. For Home F, although it has, a smooth and similar average consuming daily behaviour; it is not the same for average monthly consumption.
- Home C is twice as big as Home B, and this may indicate that the bigger a home is, the more difficult it is to predict its consumer behaviour. Occupancy in such larger houses can also deviate much more in comparison to typical-size homes, thus monitoring of occupancy could bring higher value. Unfortunately, there is no information for Home F.
- Transforming the consumption from a real number into a binary class achieves better results when the point of division is just the mean of all instances. Next in performance is when the division is based on a seasonal mean and finally on a monthly mean value; the first two seems to be the more reasonable.
- Unexpectedly, Home F and Home C have no differentiation between the predictions around On and Off-peak periods. In Home B there is a significant decrease on Off-peak period, possibly due to the fact that for Home B none of the periods is actually of low consumption.

The most important part of the analysis is the proper feature selection, as the selection of algorithms or the evaluation technique (i.e. Cross-validation, Manual split) does not affect the result that much.

A. *Threats to validity*

Every research ‘suffers’ from threats that might question the validity of the approach and its results. In this case, a main threat is that of sufficiently deseasonalising data. The factor of time could possibly be analysed to more explanatory variables. Another significant factor is access to accurate weather data. Since the model examines if weather data can accommodate forecasting, these data should be accurate. However, obviously for forecasting on day-ahead period the weather data themselves are also acquired via prediction.

Since the problem is not treated as a time-series one, it is not clear which method against over-fitting is more appropriate. Research has shown that decision tree classifiers improve generalisation accuracy via pre-pruning [27], [28]. In general, in time-series problems there is a different way to use cross-validation than the classic. In addition, the size of the homes is much bigger than a typical house or a regular apartment. Since the electricity consumption is also affected by the size, may the results be different in smaller residences. Finally, the last threat is the selected metric to compare the models against. Accuracy is not always the more appropriate metric, especially when the classes are imbalanced. Generally, the recall for *Low* class was much higher while for *High* class it was much lower.

B. *Future research directions*

An important extension of this work is the evaluation of the constructed model on different houses. The *UMass* Trace repository is expected to release new data for different homes in 2019, thus it could bring extra knowledge about the model by applying the same techniques for those. Regarding feature selection, it was desired to introduce the concept of the *week of the month*, which is increasingly used by similar efforts. More specifically, it is claimed that people tend reasonably or not to consume higher amounts of electricity a specific week of the month. Such a distinction between days could bring probably better results.

In addition, another feature could also regard the summation of consumption if the provider follows an escalated pricing policy. It is common in such cases, that the consumers reduce their consumption when they pass a threshold, thus reaching a new level of pricing during a month. Occupancy or activity monitoring could also be recorded and examined as an input in the electricity forecasting models. This could have a significant impact on model's accuracy given the scale of the application. Regarding the classification algorithms under examination, it is desired to also test neural networks, which need a wide hyper-parameter tuning, and it was decided to be excluded from this work, as well as other classifiers including logistic regression [19]. The perspective of merging the data of each house in order to simulate a tiny grid could also gather more interest, but as such, it could also eliminate the factor of personalisation, which was the biggest challenge of this work.

References

[1] D. Chen, D. Irwin (2017) Weatherman: Exposing Weather-based Privacy Threats in Big Energy Data, 2017 *IEEE Int'l Conf. Big Data* (Big Data).

[2] A. Mishra, D. Irwin, P. Shenoy, J. Kurose, T. Zhu (2012) SmartCharge: Cutting the Electricity Bill in Smart Homes with Energy Storage, e-Energy '12 *Proc. 3rd Int'l Conf. on Future Energy Systems: Where Energy, Computing and Communication Meet*, Article No. 29.

[3] S. Barker, A. Mishra, D. Irwin, P. Shenoy, J. Albrecht (2012) SmartCap: Flattening Peak Electricity Demand in Smart Homes, 2012 *IEEE Int'l Conf. Pervasive Computing and Communications*.

[4] O.M. Longe, K. Ouahada, S. Rimer, H. Zhu, H.C Ferreira (2015) Effective Energy Consumption Scheduling in Smart Homes, AFRICON 2015.

[5] A. Azadech, S.F Ghaderi, S. Sohrabkhami (2007) Forecasting electrical consumption by integration of Neural Network, time-series and ANOVA, *Applied Mathematics and Computation*, Vol. 186, No. 2, pp. 1753-1761.

[6] A. Ahmad, M. Hassan, M. Abdullah, H. Rahman, F. Hussin, H Abdullah, R. Saidur (2014) A review on applications of ANN and SVM for building electrical energy consumption forecasting, *Renewable and Sustainable Energy Reviews*, Vol. 33, pp. 102-109.

[7] A. Sauhats, R. Varfolomejeva, O. Linkevics, R. Petrecenko, M. Kunickis, M. Balodis (2015) Analysis and prediction of electricity consumption using smart meter data, *IEEE 5th Int'l Conf. Power Engineering, Energy and Electrical Drives* (POWERENG).

[8] X. Li, C.P. Bowers, T. Schnier (2010) Classification of Energy Consumption in Buildings With Outlier Detection, *IEEE Transactions on Industrial Electronics*, Vol. 57, No 11.

[9] M. Espinoza, C. Joye, R. Belmans, B.D Moor (2005) Short-Term Load Forecasting, Profile Identification, and Customer Segmentation: A Methodology Based on Periodic Time Series, *IEEE Transactions on Power Systems*, Vol. 20, No 3.

[10] Q. Wang (2009) Grey Prediction Model and Multivariate Statistical Techniques Forecasting Electrical Energy Consumption in Wenzhou, China, *2nd Int'l Symposium on Intelligent Information Technology and Security Informatics*.

[11] J.Contreras, R. Espinola, F. Nogales, A. I. Conejo (2003) ARIMA Models to Predict Next-Day Electricity Prices, *IEEE Transactions on Power Systems*, Vol. 18, No 3.

[12] C. Xia, J. Wang, K. McMenemy (2010) Short, medium and long term load forecasting model and virtual load forecaster based of radial basis function neural networks, *Int'l Journal of Electrical Power & Energy Systems*, Vol. 32, No. 7, pp. 743-750.

[13] R. Torkzadeh ET. Al. (2014) Medium Term Load Forecasting in Distribution Systems Based on Multi Linear Regression & Principal Component Analysis: A Novel Approach, *19th Conf. Electrical Power Distribution Networks (EPDC)*.

[14] A. D. Amato, M. Ruth, P. Kirshen, J. Horwitz (2005) Regional energy demand responses to climate change: Methodology and Application to the commonwealth of Massachusetts, *Climatic Change*, Vol. 71, No. 175.

[15] V. Bianco, O. Manca, S. Nardini (2009) Electricity consumption forecasting in Italy using linear regression models, *Energy*, Vol. 34, No. 9, pp. 1413-1421.

[16] G. Jannuzzi, L. Schipper (1991) The structure of electricity demand in the Brazilian household sector, *Energy Policy*, Vol. 19, No. 9, pp. 879-891.

[17] D. Jose, M. Mathew, A. Krishman (2016) Weather Dependency of Electricity Demand: A case study in Warm Humid Tropical Climate, *3rd Int'l Conf. Electrical Energy Systems* (ICEES).

[18] E. Almeshaiei, H. Soltan (2011) A methodology for Electric Power Load Forecasting, *Alexandria Engineering Journal*, Vol. 50, No. 2, pp. 137-144.

[19] S. Zhang, C. Tjortjis, X. Zeng, H. Qiao, I. Buchan, J.A. Keane (2009) Comparing data mining methods with logistic regression in childhood obesity prediction, *Information Systems Frontiers*, Vol. 11, No. 4, pp. 449-460.

[20] H. M Fadzili, S. Hazlina, Y. Rubiyah, B. Salinda, F. S. Ismail (2013) Review of HVAC scheduling techniques for buildings towards energy-efficient and cost-effective operations, *Renewable and Sustainable Energy Reviews*, Vol. 27, pp. 94-103.

[21] S. Mirasgedis, Y. Sarafidis, E. Georgopoulou, DP. Lalas, M. Moschovits, F. Karagiannis et al (2006) Models for mid term electricity demand forecasting incorporating weather influences, *Energy*, Vol. 31, No. 2–3, pp. 208-227.

[22] https://www.sec.state.ma.us/cis/cishol/holidx.htm/

[23] R. Lawson, P. Thorsnes, J. Williams (2012) Consumer Response to Timen Varying Prices for Electricity, *Energy Policy*, Vol. 49, pp. 552-561.

[24] http://www.whatissmartgrid.org/featured-article/what-you-need-to-know-about-dynamic-electricity-pricing/

[25] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy (2012) Smart*: An open data set and tools for enabling research in sustainable homes, ACM SustKDD'12.

[26] https://machinelearningmastery.com/k-fold-cross-validation/

[27] C. Tjortjis, J.A Keane (2002) T3: an Improved Classification Algorithm for Data Mining, Intelligent Data Engineering and Automated Learning (IDEAL 2002). *Lecture Notes in Computer Science*, vol. 2412. Springer.

[28] C. Tjortjis, M. Saraee, B. Theodoulidis, J.A. Keane (2007) Using T3, an improved decision tree classifier, for mining stroke-related medical data, *Methods of information in medicine*, Vol. 46, No. 5, pp. 523-529.

[29] P. M. Santos et al. (2018), "PortoLivingLab: An IoT-Based Sensing Platform for Smart Cities," in IEEE Internet of Things Journal, Vol. 5, no. 2, pp. 523-532.