Energy Consumption Prediction

@justgrossi
https://github.com/justgrossi/Portfolio.git

***Abstract*—A predictive model was created to predict households energy consumption. The data base was integrated with weather measurements taken from the nearest weather station. Based on previous related work *K-Nearest Neighbour* and *Support Vector Machine* were the considered algorithms with the former performing better and recording the lowest root mean square error (i.e. *R.M.S.E.*) of 0.0744.**

***Index Terms*—Regression, K-Nearest Neighbor, Support Vector Machine, Energy consumption.**

I. Introduction

The energy industry has become the focal point of widespread interest. Due to recent geopolitical events [1] several European countries foresee challenges in securing the required levels of commodities, and are rethinking long-term strategies. Citizens are experiencing service level reductions, disruptions and increased costs because of new supplying difficulties. Machine learning can assist in developing a deep understanding of the dynamics behind energy consumption, leading to more efficient behaviours and budgeting. Energy consumption analyses allow better production planning and could help rationalize the consumption as well a result. As electricity cannot be stored for later use, one of the greatest challenges is ensuring volumes fed into networks are sufficient to cope with peak hours, while not excessively exceeding demand's levels. Predictive models accurately forecasting demand are crucial to reduce energy production's environmental impact too, making available what is needed when is really needed. The data sets was retrieved from the *UCI Machine Learning Repository* [2].

II. Related Work

In [3], energy consumption was analysed in the Chinese city of Hangzhou comparing a hotel vs a shopping mall. Hourly levels of demand were predicted availing of both *Support Vector Machine* and *Neural Networks*. The data were collected during 29 summer-business-days using smart meters and sensors. *Support Vector Machine* recorded the best performance. Researchers, though, argue that modelling a predictor for a building characterized by varying operation timetables, a complicated planimetry, and individual rooms' random consumption-patterns like a hotel make it challenging to establish a reliable framework that could be generalized and broadly applied. *Support Vector Machine* with a *Gaussian* kernel is the algorithm of choice in [4] too, where hourly cooling-electricity demand of an office building in Guangzhou, China was forecast. Analysing May, June, July, August and October's data the algorithm predicted more accurate results than *Neural Networks*. Eight algorithms were compared in [5] where Hong Kong's tallest building's energy consumption and demand peaks were analysed. *Multiple Linear Regression, Auto Regressive Integrated Moving Average (i.e.* A.R.I.M.A.*), Support Vector Regression, Random Forests, Multi-layer Perceptron, Boosting Tree, Multivariate Adaptive Regression Splines*, and *K-Nearest Neighbors* were the adopted algorithms. *Support Vector Regressor* and *Random Forest* emerged as the better performing. Authors, eventually, created an ensemble model combining all eight algorithms, optimized it using the genetic optimization algorithm, and obtained an increased performance. The weights associated with each single model, determined the objective function the purpose of which was to minimize model's mean absolute percentage error (i.e. *M.A.P.E.*). Households' consumption levels in residential buildings, due to air conditioning, were analysed in [6]. Behavioural data were collected through infra-red signals coming from air conditioning units' remote controllers. Switching on and off the unit, adjusting the desired temperature and the length of usage intervals were all tracked to determine usage rates and, ultimately, quantify consumption. Environmental variables such as location of the apartment within the building, living room size, households' disposal income, number of family members, air conditioning unit's type and model were considered, too. *Random Forest* performed better than *Logistic Regression* and *Support Vector Machine*. *Neural Networks* and *Random Forest* were compared in [7]. The hourly energy consumption of a Spanish hotel was predicted analysing its historical data acquired through building's energy management system. The data base was enriched adding the daily number of reservations tracked by the hotel's reservation system. Finally, the data were also integrated adding weather measurements, such as outdoor air temperature, dew point temperature, wind speed and relative humidity; all collected through a nearby weather station. *Neural Network*'s performance resulted to be superior. Because of the possibility to tune its parameters and work with categorical variable easily, though, authors argue in favor of *Random Forest*. Additionally, it emerged that including the number of guests into the predictive features increased predictions' accuracy only marginally. The average intensity of energy consumption in New York was analysed by [8], to asses which features determined consumption levels of multi-family residential buildings the most. A sample of 3608 buildings was investigated and the data for a total of 171 metrics was collected. The measured features covered 7 macro categories: the building itself, economy, education, environment, households, surroundings, and transportation. *Random Forest, Multiple Linear Regression*, and *LASSO Re-*

gression were compared and *Random Forest* emerged as the best performing and made it possible to rank features in order of importance. The independent variables determining energy consumption levels the most were all related to subjects' education. In particular, the percentage of people holding an academic degree was the most influential. Heating-related consumption levels were investigated, in [9], with regard to a Chinese residential district. The data collected during the winter season were analysed comparing the performances of five algorithms on accuracy, robustness and interpretability. *Random Forest* outperformed *Extreme Gradient Boosting, Artificial Neural Network, Gradient Boosting Decision Tree*, and *Support Vector Regressor*. Historical data related to both previous winter's weather conditions, consumption levels and dry bulbs' temperatures were the most important features in making those predictions. The number of occupants within an households, on the other hand, had little effect on models' predictive capabilities.

TABLE I
ALGORITHMS COMPARISON - RMSE

Reference	Technique	RMSE
[3]	*Support Vector Machine*	*0.85*
[4]	*Support Vector Machine*	*0.06*
[5]	*Support Vector Machine*	*282.8*
[6]	*Random Forest*	*NA*
[7]	*Neural Network*	*4.97*
[8]	*Random Forest*	*0.879*
[9]	*Random Forest*	*0.21*

III. METHODOLOGY

The Cross Industry Standard Process for Data Mining (i.e. *CRISP-DM*) [10] was the methodology of choice. The research question was approached and facilitated by framework' six stages.

A. *Business understanding*

Predictive models accurately forecasting energy consumption levels are useful for mainly two reasons. On a smaller, consumer-wise scale accurate predictions could assist households, companies and organizations in general in developing increased awareness regarding the behaviours that lead to specific consumption levels. Hence, enhanced budgeting and expenditure control abilities could be enacted. On a larger, production-wise scale accurate predictions could facilitate demand forecasting, capacity and production planning. As electricity cannot be stored for later use, one of the greatest challenges is ensuring volumes fed into networks are sufficient to cope with peak hours, while not excessively exceeding demand's levels. Finally, as mentioned already, given the consequences of recent geopolitical events, the ability to accurately forecast energy consumption might constitute a considerable advantage.

B. *Data understanding*

The data set included 19,735 records for a total of 31 numerical features. Predictors were represented by the temperature and humidity of a household's different rooms. There were no missing values. Wireless sensors and Internet of Things (i.e. *I.O.T.*) type of devices were used to record the data, every ten minutes, over a period of 4.5 months. The target label was represented by the *'Consumption'* variable expressing the household's energy consumption in Watt per hour. Values were logged every 10 minutes availing of building's energy meter. Data were also integrated with weather measurements taken from the nearest weather station. Pressure, wind, visibility, dew point, both external temperature and humidity were considered as well.

C. *Algorithms Selection*

As the data set was composed entirely of numerical features, *K-Nearest Neighbor* and *Support Vector Machine* were the algorithms of choice, as the latter also consistently delivered a strong performance in a predominant part of the reviewed literature. In contrast, *K-Nearest Neighbor* was selected for the opposite reason, as it seemed to have been neglected, and very rarely mentioned, in previous related works. For such reasons, it was investigated whether a comparable or better performance could be achieved.

D. *Data preparation*

As both *K-Nearest Neighbor* and *Support Vector Machine* base their predictions on a calculated measure of distance, a custom function was created to normalise all features, and control the different scales of measurement's effect on the computation. Finally, an 80/20 split was applied to get the train and test sets.

E. *Modelling*

A 10-folder cross-validation procedure was used to determine *K-Nearest Neighbor*'s optimal number of '*K*' (i.e. neighbours) to consider when computing the Euclidean distance. The 1-10 range was tested with 3 delivering both the lowest *root mean square error* (i.e. *R.M.S.E.,* R^2) and mean absolute error (i.e. *MAE*) (Fig.1). The most suitable *Support Vector Machine*'s *kernel* function was found comparing the *linear, polynomial* and *radial* kernels on a 5-folder, cross-validation procedure applied to the same folders. Performances, hence, were directly comparable and the *polynomial* kernel emerged as the best performing (Fig.2). The model was then retrained tuning the hyper parameters as follows (Fig.3):
- ***'degree'***: of the polynomial function; 2 yielded he best result,
- ***'scale'***: optimal scaling factor, the best was 0.2,
- ***'C'***: cost function influencing misclassified instances. The higher the value hence the penalty applied, the fewer the errors. The best value was 2.

IV. EVALUATION

K-Nearest Neighbour and *Support Vector Regressor* recorded almost identical performances the former, however, accounted for a higher percentage of dependent variable's variability (40.4%). *K-Nearest Neighbour* performed also better than more sophisticated models described in the reviewed literature.

On average, the consumption levels predicted were a mere 0.0354 Watt-per-hour higher than the real values. *Support Vector Machine*'s performance was in line with those reported in the reviewed literature. When comparing models' mean absolute error the difference was negligible. The fact, however, that *Support Vector Regressor* accounted for, roughly, only 27% of the dependant variable's variability, cast a shadow on its reliability, and on whether the model could be generalized and applied to other use cases. Lastly, models' *M.S.E.* was compared against a naive regressor, used as a baseline and represented by the target variable's mean value in the test set (i.e. 0.0098), and both models scored lower; meaning they were both an improvement over a model simply predicting the average energy consumption.

TABLE II
MODEL EVALUATION

Model	*RMSE*	*R Squared*	*MAE*	*MSE*
KNN	0.0744	0.4041	0.0354	0.0060
SVM	0.0854	0.2719	0.0367	0.0068

V. CONCLUSIONS AND FUTURE WORK

Predictive models based on *K-Nearest Neighbour* and *Support Vector Machine* were built to predict households energy consumption levels. The model based on the former algorithm recorded the highest precision and accounted for dependent variables' highest percentage of variability (40.4%). On average, the predicted energy consumption levels were a mere 0.0354 Watt-per-hour higher than the observed values. Additionally, both models performed better than the models described in the reviewed literature. Future research should expand the data sets including both more records and predictors, considering running a principal component analysis to eliminate possible redundant features. It would also be interesting to analyse and understand if, and how, behaviours and consumption levels changed during the SARS-CoV-2 pandemic. Lastly, improved computing capabilities would be beneficial in assessing the performances of more sophisticated models.

REFERENCES

[1] J.F. Adolfsen, F. Kuik, E. M. Lis, T. Schuler, The impact of the war in Ukraine on euro area energy markets, European Central Bank Economic Bulletin, Issue 4/2022. Available at: www.ecb.europa.eu/pub/economic-bulletin/focus/2022/html/ecb.ebbox202204$_0$1 68ef3c3dc6.en.html

[2] UCI Machine Learning Repository. Available at: https://archive.ics.uci.edu/ml/datasets.php

[3] Y. Chen, H. Tan, Short-term prediction of electric demand in building sector via hybrid support vector regression, Applied Energy, Volume 204, 2017, Pages 1363-1374, ISSN 0306-2619, https://doi.org/10.1016/j.apenergy.2017.03.070.

[4] Q. Li, Q. Meng, J. Cai, H. Yoshino, A. Mochida, Applying support vector machine to predict hourly cooling load in the building, Applied Energy, Volume 86, Issue 10, 2009, Pages 2249-2256, ISSN 0306-2619, https://doi.org/10.1016/j.apenergy.2008.11.035.

[5] C. Fan, F. Xiao, S. Wang, Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, Applied Energy, Volume 127, 2014, Pages 1-10, ISSN 0306-2619, https://doi.org/10.1016/j.apenergy.2014.04.016.

[6] S. H. Mun, Y. Kwak, J. H. Huh, A case-centered behavior analysis and operation prediction of AC use in residential buildings, Energy and Buildings, Volumes 188–189, 2019, Pages 137-148, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2019.02.012.

[7] M. W. Ahmad, M. Mourshed, Y. Rezgui, Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption, Energy and Buildings, Volume 147, 2017, Pages 77-89, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2017.04.038.

[8] J. Ma, J. C. P. Cheng, Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests, Applied Energy, Volume 183, 2016, Pages 193-201, ISSN 0306-2619, https://doi.org/10.1016/j.apenergy.2016.08.096. https://www.sciencedirect.com/science/article/pii/S0306261916311941

[9] R. Wang, S. Lu, Q. Li, Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings, Sustainable Cities and Society, Volume 49, 2019, 101623, ISSN 2210-6707, https://doi.org/10.1016/j.scs.2019.101623.

[10] CRISP-DM, Data Science Project Management. https://www.datascience-pm.com/crisp-dm-2/ (accessed Jun. 09, 2021).

APPENDIX

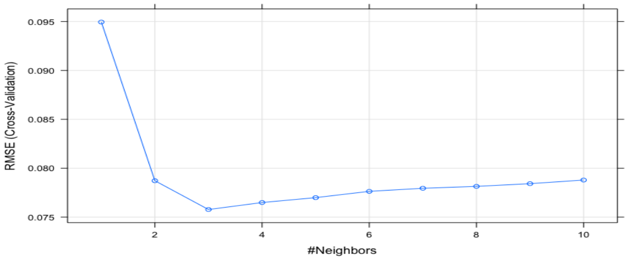

Fig. 1. K-Nearest Neighbour - Number of *K*

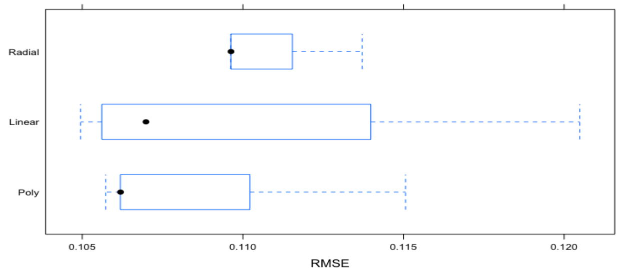

Fig. 2. Support Vector Machine - *Kernel* Comparison

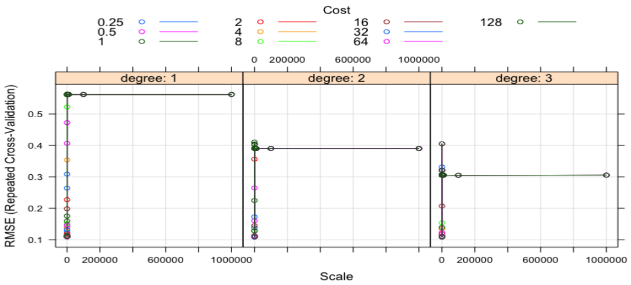

Fig. 3. Support Vector Machine - Hyper parameters tuning