Energy Consumption Prediction

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

***Abstract*—A model was created to predict a household energy consumption levels. Data were integrated with weather measurements taken from the nearest weather station. Based on previous related work *K-Nearest Neighbour* and *Random Forest* were the applied algorithms. Random Forest emerged as the best performing: R.M.S.E.: 12.28, RSquared: 0.99, M.A.E.: 3.87, M.S.E.: 150.91.**

***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 recent studies analyzing energy consumption prediction, diverse methodologies have been explored across various settings.

- **Building-Specific Predictions:**

Hangzhou Hotel vs. Shopping Mall [3]: A study compared Support Vector Machine and Neural Networks to predict hourly energy demand. Despite challenges in modeling hotels with varying operation schedules and complex layouts, Support Vector Machine showed superior performance.
Guangzhou Office Building [4]: Another study forecast cooling-electricity demand, with Support Vector Machine outperforming Neural Networks in predicting accurate results.

- **High-rise Buildings and Residential Spaces:**

Hong Kong's Tallest Building [5]: Eight algorithms were compared, with Support Vector Regressor and Random Forest emerging as top performers. An ensemble model combining all eight algorithms yielded improved results. Residential Buildings' Air Conditioning [6]: Behavioral data from infrared signals were utilized to analyze households' consumption. Random Forest outperformed Logistic Regression and Support Vector Machine.

- **Incorporating External Factors:**

Spanish Hotel Energy Prediction [7]: Neural Networks showed superior performance initially, but authors favored Random Forest due to its tunability and handling of categorical variables. New York Multi-family Residential Buildings [8]: Random Forest was utilized to determine feature importance, highlighting the impact of education-related metrics on energy consumption levels.

- **Seasonal and Environmental Factors:**

Chinese Residential District [9]: During the winter season, Random Forest outperformed other algorithms in predicting heating-related consumption levels, emphasizing the significance of historical weather data.

Each study delved into different contexts and algorithms for energy consumption prediction, emphasizing the challenges and advancements in modeling and forecasting across varied settings.

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

Accurate energy consumption levels forecasts are useful for mainly two reasons. On a smaller, consumer-wise scale they can aid households and organizations in developing increased awareness regarding the behaviours that lead to such consumption levels. Enhanced budgeting and expenditure control abilities therefore can be enacted. On a larger, production-wise scale they can facilitate demand forecasting hence 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 above, given recent geopolitical events the ability to accurately forecast energy consumption might constitute a considerable advantage.

B. Data understanding

The data base included 19,735 records and 32 numerical features. Predictors were represented by the temperatures 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

K-Nearest Neighbor and *Random Forest* were the algorithms of choice. The latter 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

Features *'Year'* and *'Sec'* were excluded from the analysis as they both held constant values with no variability. The data set was standardized and outliers with an absolute value greater than three were removed. In addition, since several of the predictors were correlated (Fig.1), a principal component analysis was performed to reduce the data set's dimensionality. Eigenvalues showed that out of the initial 31 independent features fourteen principal components accounted for 95.85% of label's variability (Fig.2). 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*' and *Random Forest*'s combination of hyper parameters yielding the best in-sample performance. The number of neighbours delivering the best result was 40 (Fig.3). The optimal values for *Random Forest*'s number of trees to grow, and of random features to consider at each split were 100 and 14 respectively.

IV. Evaluation

The model based on *Random Forest* was the best performing even though the differences between the two algorithms were negligible as shown in Table I. In addition, both models performed better than more sophisticated architectures adopted in some of the reviewed related works; and than a baseline naîve predictor always predicting target variable's mean value in the test set (i.e. the average energy consumption).

TABLE I
Model Evaluation

Model	*RMSE*	*R Squared*	*MAE*	*MSE*
KNN	13.18	0.99	3.75	173.77
RF	12.28	0.99	3.87	150.91

V. Conclusions and Future Work

Predictive models based on *K-Nearest Neighbour* and *Random Forest* were built to predict a household energy consumption levels. *Random Forest* recorded the most accurate predictions as, on average, predicted consumption levels were a mere 12.28 Watt-per-hour higher than the observed values. Both models performed better than more sophisticated architectures described some of the reviewed related works. Future research should expand the data sets including both more records and predictors. 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 replicating some of the more sophisticated architectures described in the related works, to assess whether better performances could be achieved.

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

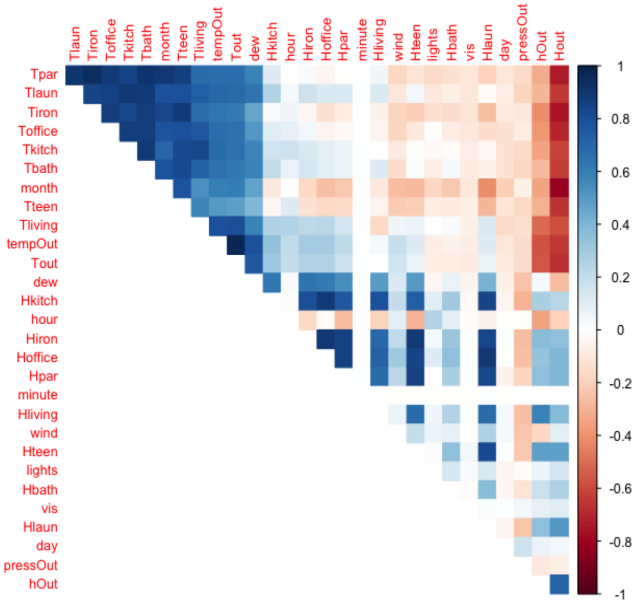
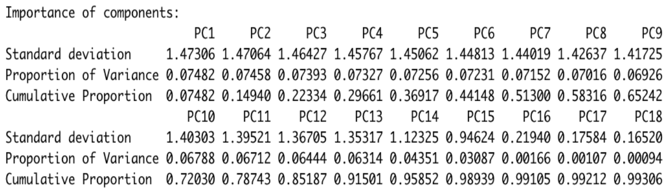


Fig. 1. Correlations Matrix



```
Importance of components:
                           PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     1.47306 1.47064 1.46427 1.45767 1.45062 1.44813 1.44019 1.42637 1.41725
Proportion of Variance 0.07482 0.07458 0.07393 0.07327 0.07256 0.07231 0.07152 0.07016 0.06926
Cumulative Proportion  0.07482 0.14940 0.22334 0.29661 0.36917 0.44148 0.51300 0.58316 0.65242
                          PC10    PC11    PC12    PC13    PC14    PC15    PC16    PC17    PC18
Standard deviation     1.40303 1.39521 1.36705 1.35317 1.12325 0.94624 0.21940 0.17584 0.16520
Proportion of Variance 0.06788 0.06712 0.06444 0.06314 0.04351 0.03087 0.00166 0.00107 0.00094
Cumulative Proportion  0.72030 0.78743 0.85187 0.91501 0.95852 0.98939 0.99105 0.99212 0.99306
```

Fig. 2. Principal Components - Eigenvalues

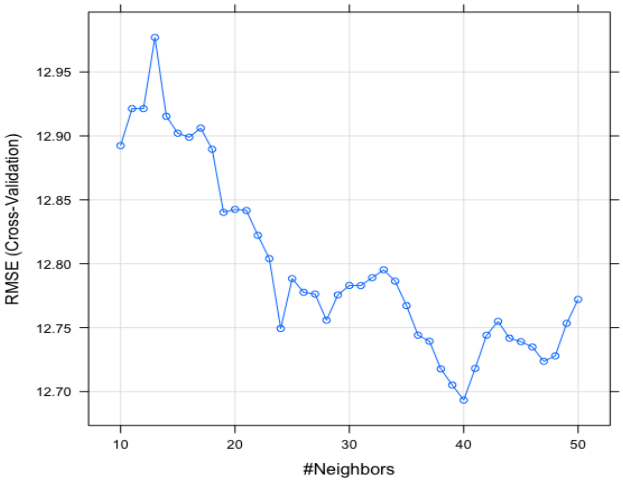


Fig. 3. K-Nearest Neighbour - Number of *K*