Estimation of Hourly Utility Usage Using Machine Learning

Albert Wong
Mathematics and Statistics
Langara College
Vancouver BC, Canada
ORCID: 0000-0002-0669-4352

Chunyin Chiu
Mathematics and Statistics
Langara College
Vancouver BC, Canada
ORCID: 0000-0002-5932-539

Abigail Abdulgapul
Mathematics and Statistics
Langara College
Vancouver BC, Canada
ORCID: 0000-0002-7285-1096

Mirza Nomaan Beg
Mathematics and Statistics
Langara College
Vancouver BC, Canada
ORCID: 0000-0002-6769-7151

Youry Khmelevsky
Computer Science
Okanagan College
Kelowna BC, Canada
ORCID: 0000-0002-6837-3490

Joe Mahony
Harris SmartWorks
Ottawa Ontario, Canada
JMahony@harriscomputer.com

2022 IEEE International Systems Conference (SysCon) | 978-1-6654-3992-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/SysCon53536.2022.9773816

***Abstract*—The COVID-19 pandemic has had a major impact on the usage of various utilities. To assess the impact, this research explores the (baseline) estimation of hourly utility usage if the pandemic did not happen. Using usage data from Harris SmartWorks, various machine learning algorithms are implemented to show that they are effective in modelling hourly usage patterns, calendar effects, as well as "lingering" effects of the exogenous factors and produce accurate results.**

***Index Terms*—utility usage, time series, machine learning, deep learning applications, big data**

I. Introduction

The COVID-19 pandemic is impacting personal, family, and business environments in general and utility usage in particular. For some municipalities and cities across North America, residential utility usage increased during the lockdown [1], [2]. The higher energy consumption reflects the increased use of computing such as videos streaming and conferencing due to work-from-home and learning-from-home activities, as well as other stay-at-home activities such as food preparation. A study last year has indicated that residential refrigerators are working overtime due to the increased storage of warm leftovers being placed in the appliance [3].

As providers of critical infrastructure, the utility industry plans for many foreseeable hazards, but it is less likely that health emergencies, such as the COVID-19 crisis, are planned for. There is a need to support utility companies in having the appropriate data available for continuity plans that are adaptable to fully address the fast-moving and unknown variables of an outbreak such as COVID-19. Therefore, it is important to quantify the impact of COVID-19 on utility usage to fulfill this business need.

In this paper, we present the approach to measuring the impact of COVID-19 by estimating, using historical data up to March 2021, utility usage during the pandemic period if COVID-19 did not happen. These baseline estimates could then be compared with the corresponding actual usages during the pandemic periods to arrive at estimates of the pandemic impact. Simply put, the impact of COVID-19 on usage could be estimated as the difference between the actual usage and the estimated (baseline) usage derived from the historical data before the pandemic.

The approach of estimating the baseline usage using historical data was implemented through a number of machine learning models and on several utility data sets in electricity, water, gas, and steam. Given that results from these data sets were similar, we will, in this paper, focus on the work with one particular electricity data set in the United States.

Note that the work presented here was conducted as part of the applied research and capstone projects at Okanagan and Langara Colleges by faculty and students with support from industry [4]–[6], [6]–[9], [9]–[21], [21]–[24].

II. Literature Review

In the past, many studies use the traditional modelling tools, such as Autoregressive Integrated Moving Average (ARIMA) models, to produce predictions for a time series. This approach has been applied in many fields such as medicine, climate research, and energy consumption research [25]–[29]. These studies used trends and patterns of the time series of interest to produce predictions for the future. Some articles also used the combination of a time series model with other statistical techniques, for example, integrating the time series models with smoothing techniques, for the development of forecasts [30], [31].

Apart from the traditional time series analysis, the employment of a regression model is another option [28], [29]. A regression model allowed researchers to take relevant predictors of the forecast values into account. However, it is challenging, from a modelling standpoint, to incorporate historical trends

Supported by Harris Utilities (Harris) and NSERC Grant "A novel approach to COVID-19 Impact Analysis and Reporting for Utilities, 2020-2021.

and patterns of the predicted variable as independent variables in a regression setting.

Recently, researchers have started to use various machine learning algorithms for time series analysis. For example, the Support Vector Regression (SVR) algorithm was used to forecast individual electricity consumption [32]. A number of researchers also used ensemble methods, especially the boosting algorithms, to create predictive models with a high level of accuracy using time series data [27], [30].

In addition, the multiple layer perceptron (MLP) algorithm is commonly used in this area [27], [30], [33], [34]. MLP can explore non-linear associations between numerical or categorical predictors and the variable of interest, but it cannot "learn" directly the autocorrelation pattern of the dependent variable over a period of time. In this regard, in a recent study a transformation technique was used on the calendar data so that a MLP model can take patterns over time into consideration and therefore incorporate the calendar effect [34].

It is also quite common to use a Long Short-Term Memory (LSTM) algorithm [27], [35], [36]. As a type of Recurrent Neural Network (RNN), its characteristic of feedback connection allowed it to process data in sequence, a key feature of a time series data set.

III. Methodology

As mentioned, the impact of COVID-19 on energy consumption can be evaluated as the difference in utility usage under the COVID-19 environment (actual and observed) and the usage that we would see if the COVID-19 pandemic did not happen (not observable and to be estimated.) This difference can be estimated by comparing the actual utility usage under the pandemic and the usage estimates if the pandemic did not exist, the so-called baseline usage. The baseline usage is not observable but can be estimated using data collected before the pandemic. It is the main objective in this research.

A statistical time series method such as ARIMA is a reasonable approach in generating estimates for the baseline usage. However, other exogenous factors, such as temperature and relative humidity, should be considered as these factors also affect energy usage [37], [38]. As well, machine learning models are considered better candidates in modelling usage for their ability to allow for explicit specification of the usage patterns and other calendar effects such as holidays and weekends. Other features, such as maximum or minimum temperature for the day, could also be generated and specified from the exogenous factors so as to specify the possible "lingering" effect of the exogenous factors.

A. Machine Learning Models for Utility Usage Forecasts

This research considered the following machine learning algorithms for predicting hourly electricity usage: random forest regression (RFR), artificial neural nets (ANN), and support vector regression (SVR). As usual, we experimented with different combinations of features and hyper-parameters for each algorithm so as to maximize accuracy. Long-Short Term Memory (LSTM), along with the traditional multiple layer perceptron (MLP), were the two implemented as the competing ANN algorithms.

With the exception of the SVR algorithm, results from these algorithms are stochastic in nature due to the probabilistic routines (the re-sampling process in RFR and the gradient descent process in ANN) used within them. To ascertain the stability of the predictive power of the models developed using these algorithms, each model developed was run ten times. The

E. Performance Metrics

Two traditional performance metrics for estimation accuracy, Root Mean Square Error and Mean Average Percent Error, were used in this research. In addition, a third metric was developed by the project team during the research to better reflect the needs in measuring estimation accuracy for utility usage.

1) Root Mean Square Error and Mean Absolute Percent Error: The root means square error (RMSE) is commonly used in evaluating the predictive or estimation performance of models. Equation 3 shows the calculation of this metric.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)} \quad (3)$$

The mean absolute percent error (MAPE), which measures the average percentage of absolute error to the actual value, is also used in this project. Equation 4 shows the calculation of MAPE.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{|Y_i|} \quad (4)$$

2) Total Absolute Error Percentage: Besides the two traditional performance metrics, a new metric, called the Total Absolute Error Percentage (TAEP) was developed by the project team to measure accuracy of the estimates comparing to the actuals, taking into account the average magnitude of the usage. This metric is therefore applicable as a performance metric for different time series data sets with different measurements and magnitudes. The TAEP metric is calculated as follows.

$$TAEP = \frac{\sum_{i=1}^{n}|Y_i - \hat{Y}_i|}{\sum_{i=1}^{n}Y_i} \quad (5)$$

We used the TAEP metric as a primary metric for model comparison.

IV. Results

Together with the application of the usual feature engineering and hyper-parameter tuning techniques, various models were developed using the algorithms described above. Table I shows the best models developed for each type of algo-

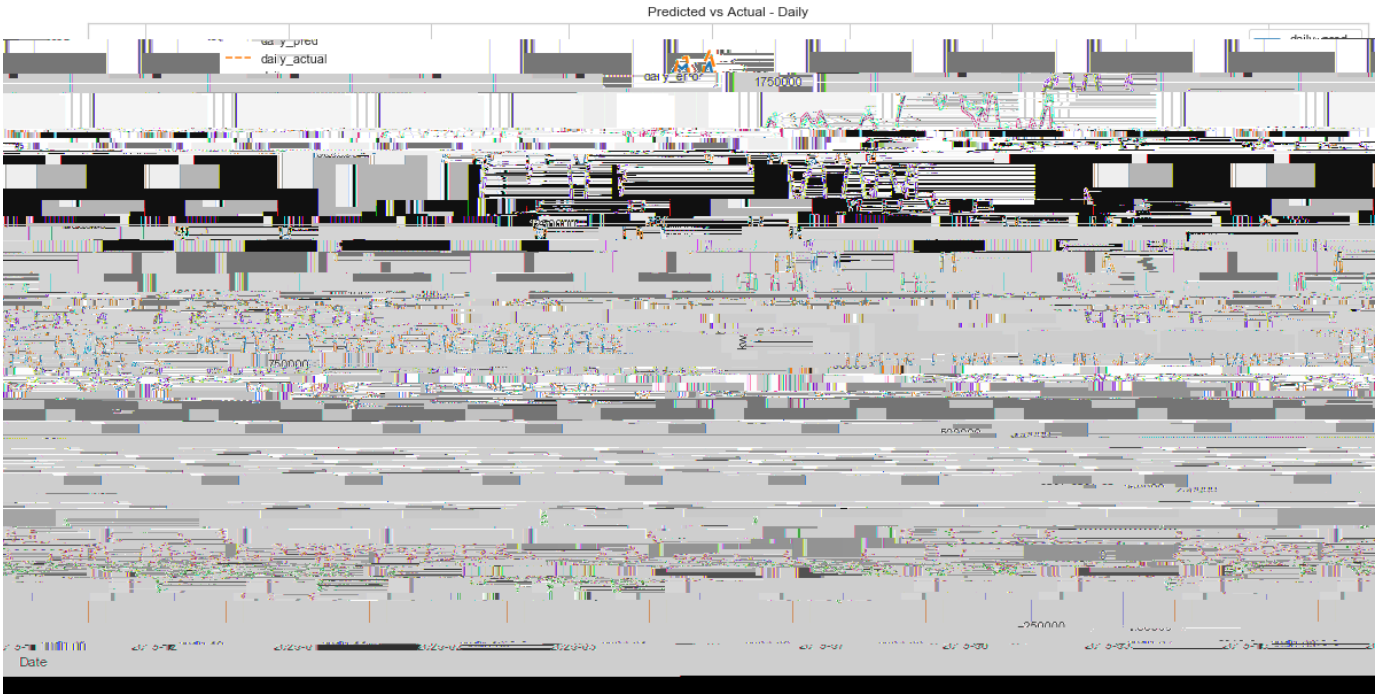

Fig. 1. Actual versus Forecasting Summarized by Day Under the MLP Model

this paper for their thoughtful comments and suggestions for improvement.

REFERENCES

[1] H. Cooley, "How the Coronavirus Pandemic is Affecting Water Demand," 2020. [Online]. Available: https://pacinst.org/how-the-coronavirus-pandemic-is-affecting-water-demand/

[2] B. Marohl and O. Comstock, "U.S. energy consumption in April 2020 fell to its lowest level in more than 30 years." [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=44556

[3] S. Hinson, "COVID-19 is changing residential electricity demand," Renewable Energy World, 4 2020.

[4] Y. Khmelevsky and V. Voytenko, "Cloud computing infrastructure prototype for university education and research," in *Computing*. ACM Press, 2010, pp. 1–5. [Online]. Available: https://doi.org/10.1145/1806512.1806524

[5] Y. Khmelevsky, V. Ustimenko, G. Hains, C. Kluka, E. Ozan, and D. Syrotovsky, "International collaboration in SW engineering research projects," in *Proceedings of the 16th Western Canadian Conference on Computing Education - WCCCE '11*, 2011.

[6] G. Hains, C. Li, Y. Khmelevsky, B. Potter, J. Gaston, A. Jankovic, S. Boateng, and W. Lee, "Generating a Real-Time Algorithmic Trading System Prototype from Customized UML Models (a case study)," no. 1, pp. 1–14, 2012.

[7] Y. Khmelevsky, M. Rinard, and S. Sidiroglou-Douskos, "A Source-to-source Transformation Tool for Error Fixing," in *Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research*, ser. CASCON '13. Riverton, NJ, USA: IBM Corp., 2013, pp. 147–160. [Online]. Available: http://dl.acm.org/citation.cfm?id=2555523.2555540

[8]

[20] M. Cocar, R. Harris, and Y. Khmelevsky, “Utilizing Minecraft bots to optimize game server performance and deployment,” in *Canadian Conference on Electrical and Computer Engineering*, 2017.
[21] G. Hains, C. Mazur, J. Ayers, J. Humphrey, Y. Khmelevsky, and T. Sutherland, “The WTFast’s Gamers Private Network (GPN®) Performance Evaluation Results,” in *2020 IEEE International Systems Conference (SysCon)*. IEEE, 2020, pp. 1–6.
[22] C. Mazur, J. Ayers, J. Humphrey, G. Hains, and Y. Khmelevsky, “Machine Learning Prediction of Gamer’s Private Networks (GPN®S),” in *Proceedings of the Future Technologies Conference*. Springer, 2020, pp. 107–123.
[23] A. Wong, C. Chiu, G. Hains, J. Behnke, Y. Khmelevsky, and C. Mazur, “Network Latency Classification for Computer Games,” in *The IEEE International Conference on Recent Advances in Systems Science and Engineering (submitted)*, 2021.
[24] A. Wong, C. Chiu, G. Hains, J. Humphrey, Y. Khmelevsky, C. Mazur, and H. Fuhrmann, “Gamers Private Network Performance Forecasting - From Raw Data to the Data Warehouse with Machine Learning and Neural Nets,” 2021. [Online]. Available: https://arxiv.org/abs/2107.00998
[25]