Universality of Forecasting Models on Water Consumption Prediction Tasks

Krzysztof Pałczyński, Tomasz Andrysiak, Magda Czyżewska,
Michał Kierul, Tomasz Kierul

Bydgoszcz University of Science and Technology
Al. prof. S. Kaliskiego 7, 85-796 Bydgoszcz, Poland
krzysztof@palczyński.com.pl

Abstract. Predicting water consumption is beneficial for water management and plays an important role in the water supply of cities. Thanks to it, the sustainable development of water resources is possible. However, in practice, water consumption is influenced by many factors and the different mechanisms of influence are complex and uncertain. Various methods for forecasting water consumption have been developed. Many methods are based on time series models that focus on past water consumption behavior and may be complemented by some exogenous variables such as a statistical regression model. The article focuses on the estimation of how much data is required in order to establish an accurate prediction of the flats water consumption without using data from these particular objects. This work examines how much data from different flats is require in order to train model in generalized fashion enabling accurate prediction of the flats unseen during training phase of the model.

Keywords: Time Series Forecasting, Neural Networks, Water Consumption.

1 Introduction

Predicting water consumption is beneficial for water management and plays an important role in the water supply of cities. Thanks to it, the sustainable development of water resources is possible. However, in practice, water consumption is influenced by many factors and the different mechanisms of influence are complex and uncertain. Various methods for forecasting water consumption have been developed. Many methods are based on time series models that focus on past water consumption behavior and may be complemented by some exogenous variables such as a statistical regression model.

The article also focuses on the estimation of how much data is required in order to establish an accurate prediction of the flats water consumption without using data from these particular objects. This work examines how much data from different flats is require in order to train model in generalized fashion enabling accurate prediction of the flats unseen during training phase of the model.

2 Datasets and Methods

In this article the data used for training and evaluation of the models comes in form of measurements of water consumption from 13 different flats from multiple estates. The measurements were conducted in the span of six months, from January 2020 to June 2020 on flats in Warsaw by SoftBlue S.A company. In order to determine both impact of selected network architecture and amount of data from different flats on results during testing, evaluation procedure has been established:

- Select model algorithm and number of flats involved during training
- Randomly select flats and put them into training set
- Take rest of flats and put them into test set
- Train the model on the test set
- Evaluate model performance on the test set

The experiments were conducted 10 times for each unique pair of model algorithm type and number of flats selected for training. The results were averaged and presented in tab. 1 and on the fig 1. ANOVA analysis and Tukey's post-hoc tests were performed in order to determine significance of differences between averaged results for each pair.

The measurements of water consumption were taken by water meters designed to send electric impulse each time 10 liters of water has flown through the device. The data was pre-processed in order to express flat's hourly water usage. The forecasts were conducted using 24-hour window. As a result, each model performed a forecast of the water consumption in the next hour based on its usage in the last 24 hours.

During nighttime it is expected that water consumption subsides. For some flats there are certain hours when water is not taken from the network at all. Because of that there are observed values of water consumption equal to 0. This is a problem for using relative metrics like Mean Absolute Percentage Error (MAPE) due to its numeric properties. The MAPE metric is calculated by

$$MAPE(Y,\hat{Y}) = \frac{1}{|Y|}\sum_{i=0}^{|Y|}\frac{|\hat{Y}_i - Y_i|}{Y_i}. \quad (1)$$

If observed value at time $t = 0$,

$$\lim_{y\to 0^+} MAPE(y,\hat{y}) = \lim_{y\to 0^+}\frac{\hat{y}-y}{y} = \hat{y}\cdot\lim_{y\to 0^+}\frac{1}{y} = \hat{y}\cdot\infty. \quad (2)$$

If predicted value is greater than zero, then MAPE error is equal to the infinity. However, if predicted value is equal to zero, then equation (2) takes form of zero times the infinity, which is an indeterminant form. Due to this limitation the metric chosen for model evaluation was Mean Squared Error (MSE) over Mean (MSEoM) given by the equation (3).

$$MSEoM(Y,\hat{Y}) = \frac{\frac{1}{|Y|}\sum_{i=0}^{|Y|}(\hat{Y}_i - Y_i)^2}{\frac{1}{|Y|}\sum_{j=0}^{|Y|}Y_j} = \frac{\sum_{i=0}^{|Y|}(\hat{Y}_i - Y_i)^2}{\sum_{j=0}^{|Y|}Y_j} \quad (3)$$

The Mean Squared Error over Mean metric combines both relative character, due to its comparison to the mean, is sensitive to significant differences between observed and predicted values and does not suffer from the numerical issues associated with division by numbers close to zero.

The models tested in this article are linear neural networks, recurrent neural networks and machine learning algorithms based on ensemble of decision trees. The linear networks are made of fully connected layers with ReLU [1] activation function. The evaluated models are one-layer network (FC-1) and two-layer network (FC-2). The recurrent neural networks are made of 2 recurrent layers and one fully connected layer. Every layer uses ReLU activation function. The evaluated recurrent models are Recurrent Neural Network (RNN) [2], Long-Short Term Memory Network (LSTM) [2] and Gated Recurrent Network [3] (GRU). The ensemble algorithms are Random Forest and XGBoost.

3 Experimental Results

The averaged results of experiments are presented in the tab. 1 and on the fig. 1. The most significant change in error rate in regard to amount of flats used during training was observed between using singular flat and two flats.

The ANOVA test has been performed for each type of network on the results from experiments clustered by the number of flats used during training. The F metric achieved for each network varies from 10.0 for FC-1 network to 25.23 for XGBoost model. The Tukey's post-hoc tests were performed in order to determine what number of flat produce error distribution separable from others.

Table 1. MSEoM for different models using varying number of flats during training phase

Flats	Models						
	FC-1	FC-2	GRU	LSTM	RNN	Random Forest	XGBoost
1	3.30%	3.60%	3.86%	3.50%	3.91%	3.35%	3.66%
2	2.95%	2.89%	3.05%	2.94%	3.06%	3.02%	3.28%
3	2.96%	2.97%	2.88%	2.80%	2.95%	2.75%	3.18%
4	2.90%	2.75%	2.88%	2.80%	2.95%	2.75%	3.07%
5	2.87%	2.73%	2.78%	2.68%	3.07%	2.69%	2.85%
6	2.90%	2.73%	2.77%	2.80%	2.87%	2.62%	2.87%
7	2.92%	2.65%	2.74%	2.66%	2.85%	2.67%	2.84%
8	2.94%	2.88%	2.71%	2.65%	2.86%	2.61%	2.78%
9	2.95%	2.63%	2.78%	2.72%	2.64%	2.59%	2.72%
10	2.94%	2.65%	2.77%	2.78%	2.79%	2.56%	2.65%
11	2.95%	2.60%	2.78%	2.68%	2.65%	2.54%	2.70%

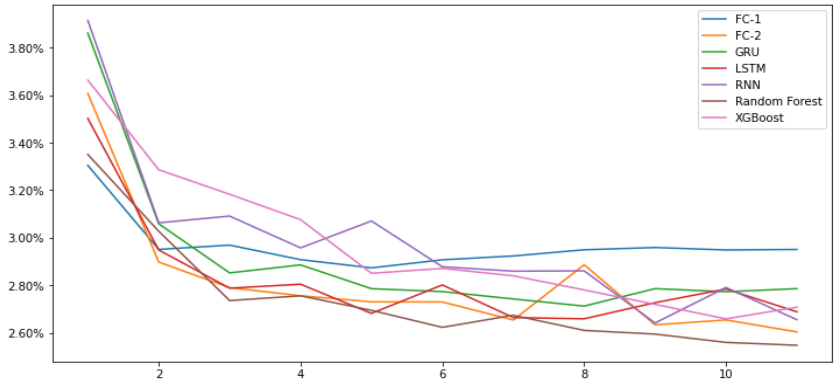

Fig. 1. Series presenting values of Mean Squared Error over Mean for different models using different number of flats for training set creation.

It turns out that error distribution of each model trained on only one flat was significantly different from error distribution of models trained on two flats. However, difference between training on two flats comparing to training on three flats was deemed not significant for every model. The only models, which results while trained on two flats differed from being trained on eleven was Random Forest and XGBoost. For every model and every pair of number of flats selected from 3 to 11 resulted in not significant difference.

The conclusion from these tests is that the smallest number of flats used during training that has significant impact on the model performance is two.

The ANOVA test has been performed on error distributions of each model trained on two flats in order to determine significance of choosing machine learning algorithm. The F metric was equal to 3.6. The Tukey's post-hoc tests were performed. It turns out that the only significant difference in error distributions were determined for XGBoost compared to FC-1, FC-2 and LSTM networks. No other model was deemed significantly different from another.

4 Conclusions

Tukey's post-hoc tests results indicates that using data acquired from two flats is enough to properly represent the task of forecasting water consumption in flats of the similar population. Although averaged error rate was smaller with each added flat for every type of model except of FC-1 and GRU, the mean differences were deemed insignificant. There was no measurable benefit of adding data from more flats.

The analysis of significance of model type selection proved that the only meaningful choice for training the model on data gathered from two flats is to not choose XGBoost algorithm. The performance of the rest of the algorithms did not prove significantly different from themselves. This notion implies selection of the least compu-

tationally expensive algorithm is preferable. However, due to suspicious, rising trend of averaged evaluation results obtained for FC-1 network, the selection of this network is not recommended despite Tukey's post-hoc test negative results. All things considered the FC-2 network is deemed the optimal model for this task.

References

1. Agarap A. F.: Deep learning using rectified linear units (relu). ArXiv Preprint ArXiv:1803.08375 (2018).
2. Sherstinsky A.: Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D: Nonlinear Phenomena 404 (132306) (2020).
3. Chung J., Gulcehre C., Cho K., Bengio Y.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling (2014).