# Customized Neural Network training to predict the highly imbalanced data of domestic hot water usage

*Caroline Risoud*, Amirreza Heidari, Dolaana Khovalyg*

Ecole Polytechnique Fédérale de Lausanne (EPFL), Integrated Comfort Engineering (ICE), CH-1700 Fribourg, Switzerland, *caroline.risoud@epfl.ch

**Abstract.** Despite space heating and cooling, the energy use for hot water production has not changed significantly over time and accounts for a big share in modern, well-insulated buildings. The main challenge of hot water generation lies in the highly stochastic nature of the domestic hot water (DHW) demand. Prediction of DHW demand can significantly help to a more efficient operational strategy in water heating systems. However, the time-series data of hot water demand is very sparse and imbalanced, including many zero demands, which makes it challenging to be predicted properly by Machine Learning methods. This study uses data recorded from a single-family building in South Africa and aims to understand how the customizations of a neural network for learning imbalanced datasets can affect the prediction of hot water demand. Four different customizations (Random over-sampling, Random under-sampling, Weight Relevance-based Combination Strategy, Synthetic Minority Over-sampling Technique for Regression) are compared with the baseline model to predict the hot water demand data. The performance of 9 different simulations is compared and the challenges are outlined. The over-sampling technique shows promising results for practical implementation by over-predicting high peaks by up to 20%, which will guarantee enough hot water production at peak usage.

**Keywords.** Hot water demand, Occupant behaviour, Machine Learning, Imbalanced data, Neural network, Time series
**DOI**: https://doi.org/10.34641/clima.2022.141

## 1. Introduction

Hot water production accounts for a big share of energy use in modern buildings [1]. While the energy use of heating and cooling systems has significantly decreased through the generations of buildings [2], the hot water energy use has not significantly decreased. Actual water heating systems do not account for stochastic behavior of occupants and follow a conservative operational strategy to ensure that hot water is available whenever it is demanded. A control strategy that can learn and predict the hot water use behavior of occupants and adapt the hot water production to the demand can significantly reduce the energy consumption [3].

The aim of this project is to learn and understand how the LSTM (Long Short-Term Memory Network) neural network can predict stochastic hot water use behavior. The work copes with the challenge of predicting a continuous target variable within a highly imbalanced dataset, by implementing modifications that resample the training dataset in order to counterbalance the biases towards zero values. The practical implementation of such models could lead to smart, dynamic water heating systems that would adapt to the predicted demand. This could lead to significant energy savings on the large scale since buildings' energy consumption in 2016 represented almost 40% of total EU energy consumption. Finally, this could contribute towards reaching Europe's 20% energy efficiency target for 2020 and beyond [4].

Only limited research has focused on predicting the highly stochastic DHW demand. Heidari et al. [3] proposed a set of Machine Learning (ML) models including Single models, Sequential Multi-task models, and Parallel Multi-task models, to predict the DHW in residential buildings. Then an adaptive water heating system that follows the predicted demand was proposed and compared to the conventional systems. In another study, Heidari et al. [5] proposed the implementation of attention mechanism and time series decomposition to improve the prediction performance of the LSTM neural networks for predicting hot water demand energy use. Ferrantelli et al. [6] developed prediction formulas based on analytical modelling to predict hourly hot water demand in Finland. Gelazanskas and Gamage [7] proposed stochastic models to forecast hot water demand at the individual building

Copyright ©2022 by the authors. This conference paper is published under a CC-BY-4.0 license.

level. Ritchie et al. [8] also evaluated the energy-saving potential by an optimal control strategy based on demand predictions. Previous studies have not addressed the imbalanced nature of hot water demand data which is the main challenge for achieving higher accuracies in predictions. The main objective of this work is to evaluate and compare few potential customizations that are developed to cope with imbalanced datasets used for ML.

# 2. Methodology

## *2.1 Case study description*

This study used the data recorded from a single-family building in South Africa [9]. The data was recorded over approximately 8 months from 18.01.2018 to 15.08.2018. It captures the outflow from the water heater in [ml/m] (millilitres per minute) every 20 minutes. 52% of the timesteps have zero hot water consumption and therefore the dataset can be considered as being imbalanced and biased towards zero values of hot water consumption.

## *2.2 Performance Metrics and Model Evaluation*

Three metrics are used in this work: MAE (Mean Absolute Error), which is an average of the sum of absolute differences between actual and predicted values. RMSE (Root Mean Squared Error), which is an average of the sum of the squares of the error. R2 (R-squared score), which measures the quality of prediction by expressing the percentage of variance explained by the model.

## *2.3 Baseline model*

The baseline model is a LSTM, which is a neural network that is trained using backpropagation through time and overcomes the vanishing gradient problem. LSTM is selected in this study due to its better memory for long-term dependencies. The effects of different architectures (number of LSTM layers, units per layer, presence of dropout layer, presence of dense layer) and hyperparameters (number of epochs, batch size) and input data are analysed to define the best performing baseline model.

## *2.4 Modifications*

The goal of the following sampling methods is to re-balance the distribution of the zero values and the non-zero values.

### *2.4.1 Random under-sampling*

This method performs under-sampling (value removal) on the "*uninteresting values*" of the training set. If the training sample is $D = \{\langle \bar{x}, y \rangle\}_{i=1}^{N}$, the 2 following parameters are defined by the user: t, the threshold on the dataset, and r, the ratio of under-sampling. The values above the threshold are considered as "important" data (which are desired to be predicted with good accuracy). In other words, the domain of the target values can be split into two sub-domains: a domain of rare values, which is of greater importance to the user, and the domain of uninteresting values. The ratio r represents the ratio between the potential candidates for under-sampling in the new (customized) dataset and the potential candidates for under-sampling in the actual (initial) dataset. If the dataset of the rare values is $D_{rare} = \{\langle \bar{x}, y \rangle \in D : y > t\}$ and $D_n$ is the dataset with the remaining observations $D_n = \{\langle \bar{x}, y \rangle \in D : y \leq t\}$ then $r = \frac{|D_{n,New}|}{|D_{n,Actual}|}$. So, if r=0.8 and t=0, then 20% of the zeros were removed from the actual dataset. Noting here that, on one hand, too large values of r will result in a new training dataset that is still too imbalanced and, on the other hand, too small values of r will result in a training data set that is too small or too different [10]. The cases to be under-sampled are then randomly selected from the potential candidates in the dataset. The following three simulations are presented: 1) t=0 and r=0.8 (under-sampling of 20% of the common cases); 2) t=0 and r=0.7; 3) t=2 and r=0.8.

***2.4.2 Random over-sampling.*** Similar to the random under-sampling, the definition of threshold t on the dataset and ratio n that represents the ratio of rare cases to be oversampled have to be specified by the user. Here, $\mathbf{n} = \frac{|\mathbf{D_{rare,New}}|}{|\mathbf{D_{rare,Actual}}|}$. The cases to be oversampled are then randomly selected from the potential candidates in the dataset. Then, they are copied and randomly reintroduced in the new dataset. The training sample resulting from this approach will be larger than the original dataset. The following two simulations are presented: 1) t=0 and n=1.2 (over-sampling of 20% of the rare cases); 2) t=4 and n=1.8.

***2.4.3 Weight Relevance-based Combination Strategy (WECS)***. In this approach, over- and under-sampling are combined. In contrary to the previously discussed methods, the cases are no longer over- or under-sampled randomly. This strategy uses the relevance values as weights to select data points for over- and under-sampling [11]. Previous work [12] proposed the use of a relevance function to map the domain of continuous variables into a [0,1] scale of relevance, i.e., $\phi(Y):Y \rightarrow [0,1]$. This function attributes levels of importance to ranges of the target variable. In this work, the choice of a shifted sigmoid function has been made in order to capture the effect of increased importance for larger DHW values and a relevance score of almost zero for zero DHW demands. This function is expressed by the following equation 1.

$$\mathbf{relevance}(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x}) = \frac{\mathbf{1}}{\mathbf{1+e^{(x-5)}}} \quad (1)$$

With x being the target variable (value of hot water consumption) in [ml/m] (millilitres per minute).

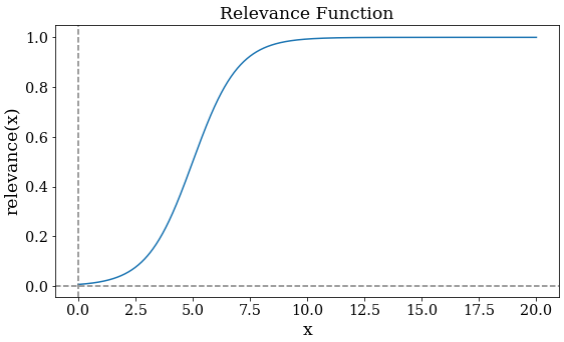



**Fig. 1:** Chosen relevance function

The key idea of this strategy is to use the relevance function scores as probabilities for resampling. A threshold of t=0 is fixed, so that the relevance scores are only used for over-sampling while the under-sampling is strictly performed on zero DHW values. Before performing over- and under-sampling, the relevance scores are translated into probabilities for each DHW value being an over-sampling candidate in the training set as shown in equation (2).

$$\mathbf{p_i^{over}} = \frac{\boldsymbol{\phi}(\mathbf{y_i}>\mathbf{t})}{\sum_{\mathbf{i=1}}^{\mathbf{N}} \boldsymbol{\phi}(\mathbf{y_i}>\mathbf{t})} \qquad (2)$$

With $(\mathbf{y_i} > \mathbf{t})$ representing the candidates for over-sampling (target values in the dataset that are bigger than the user-defined threshold t), **N** is the total number of over-sampling candidates. In that way, the higher the relevance for a DHW value, the more likely it is to be selected for over-sampling. Two simulations are performed using this strategy, namely: 1) 300 data points are oversampled and 300 data points are undersampled; 2) 300 data points are oversampled and 200 data points are undersampled.

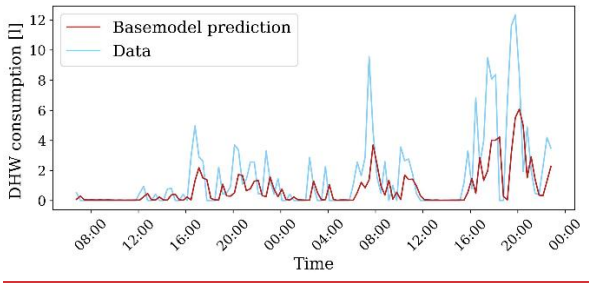
***2.4.4 SMOTE for Regression (SMOTER).*** This method combines under- and over-sampling [13] and is a modification of the classical SMOTE algorithm so that it is suitable for regression tasks [14]. This time, the over-sampling candidates are not just copied and inserted in the training set; "synthetic" cases with rare target values are generated as an interpolation of the values of its k-nearest neighbours [15]. Again, the relevance function from equation (1) is used and the threshold is held at t=0 in this method. The two following simulations are performed: 1) Over-sampling rate [%] = 100, Under-sampling rate [%] = 100, k (number of nearest neighbours to use for the generation) = 2; 2) Over-sampling rate [%] = 300, Under-sampling rate [%] = 200, k=2.

## 3. Presentation and Analysis of the results

The goal is to compare the 9 (2 over-sampling, 3 under-sampling, 2 WECS, 2 SMOTER) sampling approaches against the baseline model and to identify which methods could lead to a better prediction of the true DHW demand. Table 1 summarizes the numerical results . Figure 2 shows the results as a plot for the specific period between 10-12 July 2018.

**Tab. 1:** Comparison of accuracy metrics achieved by the different methods

| | Base-model | Over-sampling1 | Over-sampling2 | Under-sampling1 | Under-sampling2 | Under-sampling3 | WECS 1 | WECS 2 | SmotR 1 | SmotR 2 |
|---|---|---|---|---|---|---|---|---|---|---|
| MAE | 1.692 | 3.044 | 3.130 | 3.017 | 3.035 | 3.033 | 3.086 | 3.061 | 1.739 | 1.774 |
| RMSE | 3.306 | 5.959 | 6.165 | 5.859 | 6.002 | 5.911 | 6.064 | 6.071 | 3.360 | 3.459 |
| R2 score | 0.352 | 0.322 | 0.256 | 0.340 | 0.345 | 0.340 | 0.294 | 0.313 | 0.311 | 0.350 |



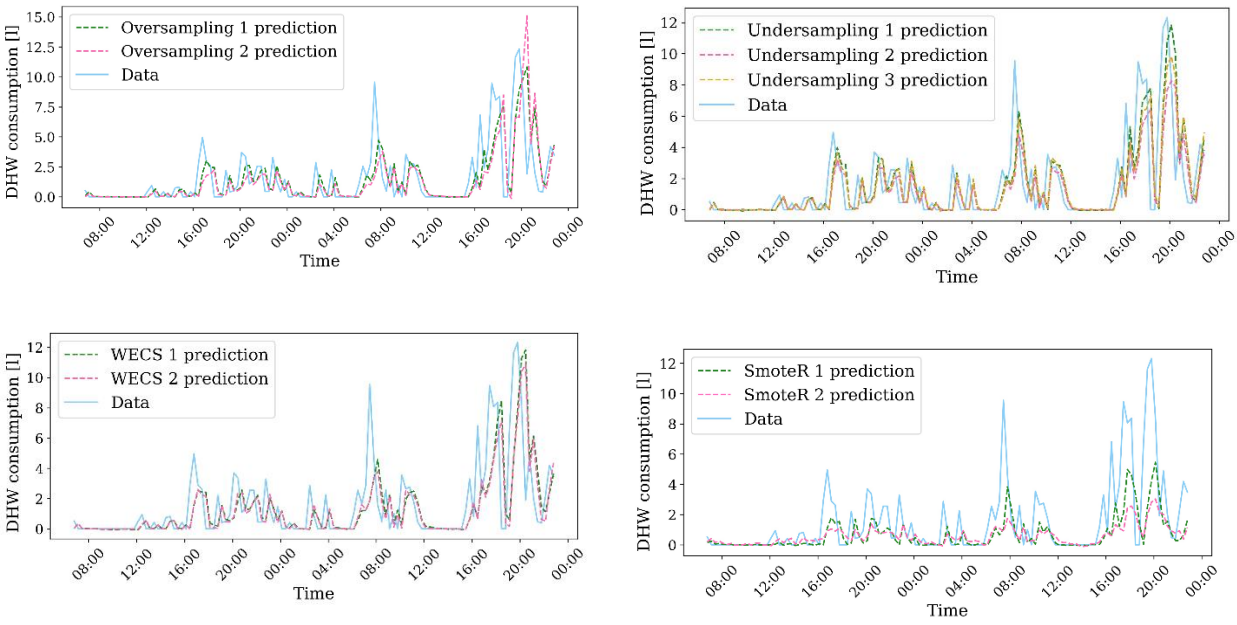



**Fig. 2:** Predicted versus actual demand (Data) by the proposed methods for the period between 10-12 July 2018

It is clear that all applied modifications work, zero values are very well predicted and the general trends are respected for each simulation. The basemodel predictions clearly show the problem of predicting an unbalanced dataset. Almost all non-zero values are underpredicted while zero-values are well predicted. However, the non-zero values are the most important values to predict accurately.

Neither of the implemented modifications overperformed the baseline model, which performed the best. WECS 1 shows the worst results. The over-sampling 2 simulations predicted higher DHW demand than the actual demand during peak hour around 8 PM. This is a real benefit given that the peak demand is the most crucial demand that has to be guaranteed. In order to guarantee the DHW demand at peak usage, predictions have to be slightly higher than the actual demand to account for the potential uncertainty while guaranteeing the peak demand.

The fact that the modifications did not perform better is due to the shift in the forecast, which means that the forecasted and the actual data are shifted by one time-step. Shifts in time-series predictions are a well-known issue [16], to which we propose the following modifications and improvements:

1. Using an alternative performance metric to evaluate the model which would be more adapted to imbalanced regression tasks. An evaluation framework for forecasting rare extreme values of a continuous target variable (precision and recall for regression) is proposed by [10]. Their key idea is weighting the error also by its relevance and not only by its magnitude.

2. Improving the input data by adding categorical features (month of the year, day of the month, hour of the day) and additional features such as the total number of hot water demands until the current moment.

3. Reformulating the problem as a multi-class classification problem. A classification task will make the forecasting much easier for the model and could overcome the shift issue. This implementation could easily be practically applied by creating a storage tank with different 'levels' where each level corresponds to a new class.

## 4. Conclusion

Based on the obtained results, the following conclusions can be drawn:

a. Predicting the exact value and time of hot water demand is a very challenging task for ML due to the stochastic nature of occupants' behavior. Thus, the proposed models usually underestimate the demand and show a shift in predicted time versus actual time of demand.

b. The SMOTER method is capable of reaching almost the same performance as the baseline model.

c. The Random Over-sampling method is very promising by being able to predict higher values at peak hours.

In sum, if the proposed strategies to cope with the

time-series shift can be introduced, the over-sampling technique shows promising results for practical application by over-predicting high peaks, which will guarantee enough hot water production at peak usage. Still, further research is required to better predict the highly stochastic occupant behavior by ML models.

## 5. References


[1] Diana Ürge Vorsatz et al. “Heating and cooling energy trends and drivers in buildings”. In: *Renewable and Sustainable Energy Reviews* 41 (Jan. 1, 2015), pp. 85–98. ISSN: 1364-0321. DOI: 10.1016/j.rser.2014.08.039.

[2] Zhihua Zhou et al. “Heating energy saving potential from building envelope design and operation optimization in residential buildings: A case study in northern China”. In: *Journal of Cleaner Production* 174 (Feb. 10, 2018), pp. 413–423. ISSN: 0959-6526. DOI: 10.1016/j.jclepro.2017.10.237.

[3] Amirreza Heidari et al. “Adaptive hot water production based on Supervised Learning”. In: *Sustainable Cities and Society* (Dec. 1, 2020), p. 102625. ISSN: 2210-6707. DOI: 10.1016/j.scs.2020.102625.

[4] European Commission. *Towards reaching the 20% energy efficiency target for 2020, and beyond.* European Commission. 2017. URL: https : / / ec . europa . eu / commission/presscorner/detail/en/MEMO_17_162 (visited on 01/04/2021).

[5] Amirreza Heidari and Dolaana Khovalyg. “Short-term energy use prediction of solarassisted water heating system: Application case of combined attention-based LSTM and time-series decomposition”. In: *Solar Energy* 207 (Sept. 1, 2020), pp. 626–639. ISSN: 0038- 092X. DOI: 10.1016/j.solener.2020.07.008.

[6] Andrea Ferrantelli et al. “Analytical modelling and prediction formulas for domestic hot water consumption in residential Finnish apartments”. In: *Energy and Buildings* 143 (2017), pp. 53–60. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2017.03.021.

[7] L. Gelažanskas and K. A. A. Gamage. “Forecasting hot water consumption in dwellings using artificial neural networks”. In: *2015 IEEE 5th International Conference on Power Engineering, Energy and Electrical Drives (POWERENG).* 2015 IEEE 5th International Conference on Power Engineering, Energy and Electrical Drives (POWERENG). ISSN: 2155- 5532. May 2015, pp. 410–415. DOI: 10.1109/PowerEng.2015.7266352.

[8] Michael Ritchie, Jacobus Engelbrecht, and M.J. (Thinus) Booysen. “Practically-Achievable Energy Savings with the Optimal Control of Stratified Water Heaters with Predicted Usage”. In: *Energies* (Apr. 1, 2021). DOI: 10.3390/en14071963.

[9] Joshua Siegel. “*Single Family Hot Water Flow Data*”. fr. In: (Sept. 2019). type: dataset. DOI: 10.7910/DVN/BQEHMU. URL: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BQEHMU (visited on 11/02/2020).

[10] Luís Torgo and Rita Ribeiro. “Precision and Recall for Regression”. In: *Lecture Notes in Computer Science*. Vol. 5808. Oct. 3, 2009, pp. 332–346. ISBN: 9783642047466. DOI: 10 . 1007/978-3-642-04747-3_26.

[11] Paula Branco, Luís Torgo, and Rita Ribeiro. “Pre-processing Approaches for Imbalanced Distributions in Regression”. In: *Neurocomputing* 343 (Feb. 1, 2019). DOI: 10 . 1016 / j . neucom.2018.11.100.

[12] Luís Torgo et al. “Resampling strategies for regression”. In: *Expert Systems* 32 (Aug. 2014). DOI: 10.1111/exsy.12081.

[13] Luís Torgo et al. “*SMOTE for Regression*”. In: vol. 8154. Sept. 2013, pp. 378–389. DOI: 10.1007/978-3-642-40669-0_33.

[14] Luís Torgo and Rita Ribeiro. “*Utility-Based Regression*”. In: Sept. 2007, pp. 597–604. ISBN: 9783540749752. DOI: 10.1007/978-3-540-74976-9_63.

[15] Aleksandra Deis. *Regression Addressing Extreme Rare Cases*. en. 2019. URL: https://kaggle.com/aleksandradeis/regression-addressing-extreme-rare-cases (visited on 11/02/2020).

[16] Flovik, V. (2018, June 7).In towardsdatascience: https://towardsdatascience.com/how-not-touse-machine-learning-for-time-series-forecasting-avoiding-the-pitfalls-19f9d7adf424


### *Data Statement*

The dataset analyzed during the current study is available in the Harvard Dataverse, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BQEHMU (visited on 11/02/2020).