FCN | Future Energy Consumer Needs and Behavior

E.ON Energy Research Center

RWTH AACHEN UNIVERSITY

Residential Energy Expenditure in Germany: Machine Learning-Based Modeling and Investigation of the Determinants (Bachelor thesis)

Autor: Behnam Alizadeh Sahzabi
Erstprüfer: Univ.-Prof. Dr.-Ing Aaron Praktiknjo
Betreuung: Jan Priesmann, M. Sc.

Kurzfassung

Die vorliegende Studie konzentriert sich auf die auf maschinellem Lernen basierende Modellierung der Energieausgaben von Haushalten in Deutschland und die Untersuchung der entsprechenden Determinanten. In diesem Zusammenhang wurden die Mikrodaten von 42.226 Haushalten in Deutschland als Datensatz verwendet, der vier Kategorien von Parametern umfasst, darunter solche, die sich auf Geräte, Standort, Gebäudeeigenschaften und sozioökonomische Merkmale beziehen. Die Ausgaben der Haushalte, die dem Strom- und Wärmebedarf entsprechen, sowie der Gesamtenergieverbrauch wurden als Schätzungsziele herangezogen. Für jede Pipeline wurde zunächst die Leistung verschiedener maschineller Lernmodelle wie Multi-Lineare Regression (MLR), Random Forest (RF), Support Vector Machine (SVM), Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Long Short-term Memory (LSTM) sowie die Kombination aus CNN und LSTM ermittelt und verglichen. Um die wichtigsten Determinanten des Energieverbrauchs von Haushalten zu untersuchen, wird ein hybrides Vorwärtsmerkmalauswahlverfahren durchgeführt. Im nächsten Schritt wird durch eine Regressionskoeffizientenanalyse die Auswirkung jedes ausgewählten Parameters auf jedes der betrachteten Schätzungsziele analysiert. Um den Einfluss der einzelnen Parameterkategorien zu vergleichen, wird schließlich die erzielte Leistung für jede Pipeline (unter Verwendung des LSTM-Algorithmus) untersucht, wobei jeweils nur eine Parameterkategorie zur Verfügung gestellt wird.

Die Ergebnisse zeigen, dass der LSTM-Algorithmus (mit Bestimmtheitsmaßen (R2-Werte) von 51,5 % bzw. 45,1 %) bei der Schätzung des elektrischen und des Gesamtenergieverbrauchs die höchste Leistung aufweist. Für die Schätzung des thermischen Energieaufwands hingegen wird das CNN (mit einem R2-Wert von 42,0 %) als das vielversprechendste Modell ermittelt. Es zeigt sich auch, dass der Leistungsbereich dieser Modelle mit demjenigen der meisten ähnlichen Studien übereinstimmt. Darüber hinaus wird festgestellt, dass die Durchführung des Merkmalsauswahlverfahrens die Anzahl der verwendeten Merkmale um 50 % reduziert. Schließlich wird gezeigt, dass die Parameterkategorie, die die Variablen im Zusammenhang mit den Gebäudeeigenschaften umfasst, erwartungsgemäß die einflussreichste ist, und nur durch die Verwendung dieser Variablen kann der LSTM-Algorithmus die elektrischen, thermischen und gesamten Energieausgaben mit R2-Werten von 0,440; 0,425 bzw. 0,361 schätzen.

Abstract

The present study is focused on machine learning-based modeling of the residential energy expenditure in Germany and investigating the corresponding determinants. In this context, the micro-data of 42,226 households in Germany that includes four categories of parameters including those related to appliances, location, building properties, and socio-economic characteristics, has been utilized as the dataset. The households' expenditure corresponding to the electrical and thermal demand along with the total energy consumption has been considered as the estimation targets. For each pipeline, the performance of different machine learning models including Multi-Linear Regression (MLR), Random Forest (RF), Support Vector Machine (SVM), Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Long Short-term Memory (LSTM), along with the combination of CNN and LSTM, has first been determined and compared. Next, in order to investigate the key determinants of residential energy expenditure, a hybrid forward feature selection procedure is performed. In the next step, by conducting a regression coefficient analysis, the effect of each selected parameter on each of the considered estimation targets is analyzed. Finally, in order to compare the influence of each parameter category, the achieved performance for each pipeline (utilizing the LSTM algorithm), while providing only one parameter category at a time, is investigated.

The obtained results demonstrate that, for electrical and total energy expenditure estimation, the LSTM (with coefficients of determination (R2 scores) of 51.5% and 45.1% respectively) is the algorithm with the highest performance. For the estimation of thermal energy expenditure instead, the CNN (with an R2 score of 42.0%) is determined to be the most promising model. The range of achieved performance of these models is also shown to be in line with those reported by the majority of similar studies. Furthermore, it is observed that performing the feature selection procedure reduces the number of utilized features by 50%. Finally, it is demonstrated that the parameter category that involves the variables related to building properties is expectedly the most influential one, and only by utilizing these variables, the LSTM algorithm can estimate the electrical, thermal, and total energy expenditure, with R2 scores of 0.440, 0.425, and 0.361 respectively.